Springer Series in Statistics

Robert E. Kass
Uri T. Eden
Emery N. Brown

Analysis of Neural Data

Springer

Springer Series in Statistics

For further volumes:
http://www.springer.com/series/692

Robert E. Kass · Uri T. Eden
Emery N. Brown

Analysis of Neural Data

Springer

Robert E. Kass
Carnegie Mellon University
Pittsburgh, PA
USA

Emery N. Brown
Massachusetts Institute of Technology
Cambridge, MA
USA

Uri T. Eden
Boston University
Boston, MA
USA

ISSN 0172-7397 ISSN 2197-568X (electronic)
ISBN 978-1-4614-9601-4 ISBN 978-1-4614-9602-1 (eBook)
DOI 10.1007/978-1-4614-9602-1
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013955054

To our families

Preface

This book serves as a guide and reference for anyone who wishes to understand analysis of neural data generated from studies that range from molecules, to circuits, to systems, to behavior.

Its origins may be traced to the decision by two of us (E.N.B. and R.E.K.), in 1998, to write a review article on statistical analysis of spike train data. Shortly after commencing we realized that some of the methods we thought we ought to be reviewing had, in fact, not yet been developed. After we and others rectified this situation, we published a pair of reviews (Brown et al. 2004; Kass et al. 2005). During this time we also broadened our interests to other experimental modalities, such as neuroimaging, and we began teaching workshops and semester-long courses on statistical methods for neuroscience. In addition, we met the third author of this book (U.E.), who came to share our interests in research and pedagogy (and who pursued his Ph.D. thesis under the guidance of E.N.B.).

It became clear that a book on this subject was desperately needed, and we agreed to write one. While this turned into a longer project than we anticipated, numerous research collaborations, conversations with colleagues at meetings, and extensive comments from students gave us many insights into the content and presentation of the principles and techniques that evolved to form this volume. We feel we are much wiser than when we started, and we hope we have succeeded in imparting a good deal of what we have learned in the process.

Some readers may expect a book organized by type of neural data. We decided, instead, to organize by analysis, with each chapter devoted to broadly categorized statistical concepts described succinctly in section headings that are available in the extended version of the table of contents. Each chapter, however, also contains multiple examples of the way these analytical ideas have been used in the brain sciences: there are more than 100 such examples throughout the book, and they are indexed. A reader wishing to see how we have discussed fMRI data, for instance, should start with the example index. More specific organizational guidelines are given in Chapter 1.

The book is intended as either a reference, or a text. R.E.K. has used preliminary versions of the manuscript in classes populated by graduate students of varying backgrounds, ranging from biologists with minimal mathematical knowledge, who were looking for conceptual understanding, to engineers, who needed to see derivations. We opted to try to satisfy both kinds of audiences.

An appendix is provided as a reminder of key mathematical ideas, and derivations are often marked as optional by indenting them. To those who wish to use the book as a text, R.E.K. would suggest the following ordering of topics:

Part I (Elementary Statistics): Chapters 1–7, 10, 12.1–12.4, 13.1.

Part II (Basic Statistical Theory): Chapters 8, 9, 11, 12.5, 13.2–13.4.

Part III (Advanced Topics): Selections from Chapters 14–19.

In his experience, Parts I and II take approximately 12 and 7 classes, respectively.

Many readers will want to see computer code for the methods we have described. We ourselves used both Matlab and R to produce figures. Although we decided not to inject Matlab or R code into the body of the book, we have put code up on our the website http://www.stat.cmu.edu/∼kass/KEB.

In addition to the many colleagues and students who made suggestions along the way, including those who are acknowledged within the text, we are indebted to Spencer Koerner, who helped clean up and create much code and many figures, Patrick Foley, who created the website, Heidi Sestrich, who fixed numerous defects in our LATEX, and Matthew Marler, who read the whole manuscript carefully and provided extremely helpful comments. We are also grateful to Elan Cohen and Ryan Sieberg, who each created several figures.

Robert E. Kass
Uri T. Eden
Emery N. Brown

Short Table of Contents

Preface vii

1 Introduction 1

2 Exploring Data 23

3 Probability and Random Variables 37

4 Random Vectors 71

5 Important Probability Distributions 105

6 Sequences of Random Variables 137

7 Estimation and Uncertainty 149

8 Estimation in Theory and Practice 179

9 Propagation of Uncertainty and the Bootstrap 221

10 Models, Hypotheses, and Statistical Significance 247

11 General Methods for Testing Hypotheses 287

12 Linear Regression 309

13 Analysis of Variance 361

14 Generalized Linear and Nonlinear Regression 391

15 Nonparametric Regression 413

16 Bayesian Methods 439

17 Multivariate Analysis 491

18 Time Series 513

19 Point Processes 563

Appendix: Mathematical Background 605

References 623

Example Index 635

Index 639

Contents

1 Introduction 1
1.1 Data Analysis in the Brain Sciences 1
1.1.1 Appropriate analytical strategies depend crucially on the purpose of the study and the way the data are collected. 3
1.1.2 Many investigations involve a response to a stimulus or behavior.. 6
1.2 The Contribution of Statistics. 8
1.2.1 Statistical models describe regularity and variability of data in terms of probability distributions. 9
1.2.2 Statistical models are used to express knowledge and uncertainty about a signal in the presence of noise, via inductive reasoning. 13
1.2.3 Statistical models may be either parametric or nonparametric. 14
1.2.4 Statistical model building is an iterative process that incorporates assessment of fit and is preceded by exploratory analysis. 17
1.2.5 All models are wrong, but some are useful. 17
1.2.6 Statistical theory is used to understand the behavior of statistical procedures under various probabilistic assumptions. 19
1.2.7 Important data analytic ideas are sometimes implemented in many different ways. 20
1.2.8 Measuring devices often pre-process the data. 20
1.2.9 Data analytic techniques are rarely able to compensate for deficiencies in data collection.. 21
1.2.10 Simple methods are essential. 21
1.2.11 It is convenient to classify data into several broad types. 21

2 Exploring Data 23
2.1 Describing Central Tendency and Variation 23
2.1.1 Alternative displays and summaries provide different views of the data. 23
2.1.2 Exploratory methods can be sophisticated. 26
2.2 Data Transformations 28
2.2.1 Positive values are often transformed by logarithms. 28
2.2.2 Non-logarithmic transformations are sometimes applied. 33

3 Probability and Random Variables 37
3.1 The Calculus of Probability 38
3.1.1 Probabilities are defined on sets of uncertain events. 38
3.1.2 The conditional probability P(A|B) is the probability that A occurs given that B occurs. 40
3.1.3 Probabilities multiply when the associated events are independent. 41
3.1.4 Bayes' theorem for events gives the conditional probability P(A|B) in terms of the conditional probability P(B|A). 42
3.2 Random Variables 46
3.2.1 Random variables take on values determined by events. 47
3.2.2 Distributions of random variables are defined using cumulative distribution functions and probability density functions, from which theoretical means and variances may be computed. 48
3.2.3 Continuous random variables are similar to discrete random variables. 52
3.2.4 The hazard function provides the conditional probability of an event, given that it has not yet occurred. 61
3.2.5 The distribution of a function of a random variable is found by the change of variables formula. 62
3.3 The Empirical Cumulative Distribution Function 64
3.3.1 P–P and Q–Q plots provide graphical checks for gross departures from a distributional form. 65
3.3.2 Q–Q and P–P plots may be used to judge the effectiveness of transformations. 69

4 Random Vectors 71
4.1 Two or More Random Variables 71
4.1.1 The variation of several random variables is described by their joint distribution. 73
4.1.2 Random variables are independent when their joint pdf is the product of their marginal pdfs. 75
4.2 Bivariate Dependence 76
4.2.1 The linear dependence of two random variables may be quantified by their correlation. 77
4.2.2 A bivariate normal distribution is determined by a pair of means, a pair of standard deviations, and a correlation coefficient. 82
4.2.3 Conditional probabilities involving random variables are obtained from conditional densities. 84
4.2.4 The conditional expectation $E(Y|X = x)$ is called the regression of Y on X. 85
4.3 Multivariate Dependence 90
4.3.1 The mean of a random vector is a vector and its variance is a matrix. 90
4.3.2 The dependence of two random vectors may be quantified by mutual information. 92
4.3.3 Bayes' theorem for random vectors is analogous to Bayes' theorem for events. 98
4.3.4 Bayes classifiers are optimal. 99

5 Important Probability Distributions 105
5.1 Bernoulli Random Variables and the Binomial Distribution . . . 105
5.1.1 Bernoulli random variables take values 0 or 1. 105
5.1.2 The binomial distribution results from a sum of independent and homogeneous Bernoulli random variables. 106
5.2 The Poisson Distribution 110
5.2.1 The Poisson distribution is often used to describe counts of binary events. 110
5.2.2 For large n and small p the binomial distribution is approximately the same as Poisson. . . . 113
5.2.3 The Poisson distribution results when the binary events are independent. 115
5.3 The Normal Distribution 116
5.3.1 Normal random variables are within 1 standard deviation of their mean with probability $2/3$; they are within 2 standard deviations of their mean with probability .95. 116

5.3.2 Binomial and Poisson distributions are approximately normal, for large n or large λ. 118
5.4 Some Other Common Distributions 119
5.4.1 The multinomial distribution extends the binomial to multiple categories. 119
5.4.2 The exponential distribution is used to describe waiting times without memory. 120
5.4.3 Gamma distributions are sums of exponentials. 123
5.4.4 Chi-squared distributions are special cases of gamma distributions. 124
5.4.5 The beta distribution may be used to describe variation on a finite interval. 124
5.4.6 The inverse Gaussian distribution describes the waiting time for a threshold crossing by Brownian motion. 125
5.4.7 The t and F distributions are defined from normal and chi-squared distributions. 128
5.5 Multivariate Normal Distributions 129
5.5.1 A random vector is multivariate normal if linear combinations of its components are univariate normal. 129
5.5.2 The multivariate normal pdf has elliptical contours, with probability density declining according to a χ^2 pdf. 130
5.5.3 If X and Y are jointly multivariate normal then the conditional distribution of Y given X is multivariate normal. 132

6 Sequences of Random Variables 137
6.1 Random Sequences and the Sample Mean 137
6.1.1 The standard deviation of the sample mean decreases as $1/\sqrt{n}$. 139
6.1.2 Random sequences may converge according to several distinct criteria. 142
6.2 The Law of Large Numbers. 143
6.2.1 As the sample size n increases, the sample mean converges to the theoretical mean. 143
6.2.2 The empirical cdf converges to the theoretical cdf. 144
6.3 The Central Limit Theorem 145
6.3.1 For large n, the sample mean is approximately normally distributed. 145
6.3.2 For large n, the multivariate sample mean is approximately multivariate normal. 147

7 Estimation and Uncertainty 149
7.1 Fitting Statistical Models 149
7.2 The Problem of Estimation 151
7.2.1 The method of moments uses the sample mean and variance to estimate the theoretical mean and variance. 153
7.2.2 The method of maximum likelihood maximizes the likelihood function, which is defined up to a multiplicative constant. 154
7.3 Confidence Intervals 158
7.3.1 For scientific inference, estimates are useless without some notion of precision. 158
7.3.2 Estimation of a normal mean is a paradigm case. 160
7.3.3 For non-normal observations the central limit theorem may be invoked. 162
7.3.4 A large-sample confidence interval for μ is obtained using the standard error $s/\sqrt{n}$. 162
7.3.5 Standard errors lead immediately to confidence intervals. 164
7.3.6 Estimates and standard errors should be reported to two digits in the standard error. 169
7.3.7 Appropriate sample sizes may be determined from desired size of standard error. 169
7.3.8 Confidence assigns probability indirectly, making its interpretation subtle. 170
7.3.9 Bayes' theorem may be used to assess uncertainty. 173
7.3.10 For small samples it is customary to use the t distribution instead of the normal. 176

8 Estimation in Theory and Practice 179
8.1 Mean Squared Error 181
8.1.1 Mean squared error is bias squared plus variance. 181
8.1.2 Mean squared error may be evaluated by computer simulation of pseudo-data. 186
8.1.3 In estimating a theoretical mean from observations having differing variances a weighted mean should be used, with weights inversely proportional to the variances. 190
8.1.4 Decision theory often uses mean squared error to represent risk. 195

8.2 Estimation in Large Samples 196
8.2.1 In large samples, an estimator should be very likely to be close to its estimand. 196
8.2.2 In large samples, the precision with which a parameter may be estimated is bounded by Fisher information. 196
8.2.3 Estimators that minimize large-sample variance are called efficient. 200
8.3 Properties of ML Estimators 202
8.3.1 In large samples, ML estimation is optimal. 202
8.3.2 The standard error of the MLE is obtained from the second derivative of the loglikelihood function. 203
8.3.3 In large samples, ML estimation is approximately Bayesian. 207
8.3.4 MLEs transform along with parameters. 208
8.3.5 Under normality, ML produces the weighted mean. 209
8.4 Multiparameter Maximum Likelihood 209
8.4.1 The MLE solves a set of partial differential equations. 210
8.4.2 Least squares may be viewed as a special case of ML estimation. 212
8.4.3 The observed information is the negative of the matrix of second partial derivatives of the loglikelihood function, evaluated at $\hat{\theta}$. 213
8.4.4 When using numerical methods to implement ML estimation, some care is needed. 214
8.4.5 MLEs are sometimes obtained with the EM algorithm. 215
8.4.6 Maximum likelihood may produce bad estimates. 219

9 Propagation of Uncertainty and the Bootstrap 221
9.1 Propagation of Uncertainty 223
9.1.1 Simulated observations from the distribution of the random variable X produce simulated observations from the distribution of the random variable $Y = f(X)$. 223
9.1.2 In large samples, transformations of consistent and asymptotically normal random variables become approximately linear. 229
9.2 The Bootstrap 237
9.2.1 The bootstrap is a general method of assessing uncertainty. 237

9.2.2 The parametric bootstrap draws pseudo-data from an estimated parametric distribution. 239
9.2.3 The nonparametric bootstrap draws pseudo-data from the empirical cdf. 241
9.3 Discussion of Alternative Methods 245

10 Models, Hypotheses, and Statistical Significance 247
10.1 Chi-Squared Statistics 248
10.1.1 The chi-squared statistic compares model-fitted values to observed values. 249
10.1.2 For multinomial data, the chi-squared statistic follows, approximately, a χ^2 distribution. 250
10.1.3 The rarity of a large chi-squared is judged by its p-value. 253
10.1.4 Chi-squared may be used to test independence of two traits. 254
10.2 Null Hypotheses 256
10.2.1 Statistical models are often considered null hypotheses. 256
10.2.2 Null hypotheses sometimes specify a particular value of a parameter within a statistical model. 257
10.2.3 Null hypotheses may also specify a constraint on two or more parameters. 257
10.3 Testing Null Hypotheses 258
10.3.1 The hypothesis $H_0: \mu = \mu_0$ for a normal random variable is a paradigm case. 258
10.3.2 For large samples the hypothesis H_0: $\theta = \theta_0$ may be tested using the ratio $(\hat{\theta} - \theta_0)/SE(\hat{\theta})$. 260
10.3.3 For small samples it is customary to test $H_0: \mu = \mu_0$ using a t statistic. 262
10.3.4 For two independent samples, the hypothesis H_0: $\mu_1 = \mu_2$ may be tested using the t-ratio. 264
10.3.5 Computer simulation may be used to find p-values. 266
10.3.6 The Rayleigh test can provide evidence against a uniform distribution of angles. 268
10.3.7 The fit of a continuous distribution may be assessed with the Kolmogorov-Smirnov test. 270
10.4 Interpretation and Properties of Tests 271
10.4.1 Statistical tests should have the correct probability of falsely rejecting H_0, at least approximately. 271
10.4.2 A confidence interval for θ may be used to test H_0: $\theta = \theta_0$. 274

10.4.3 Statistical tests are evaluated in terms of their probability of correctly rejecting H_0.. 276
10.4.4 The performance of a statistical test may be displayed by the ROC curve. 278
10.4.5 The p-value is not the probability that H_0 is true.. . . . 281
10.4.6 Borderline p-values are especially worrisome with low power. 282
10.4.7 The p-value is conceptually distinct from type one error.. 283
10.4.8 A non-significant test does not, by itself, indicate evidence in support of H_0.. 283
10.4.9 One-tailed tests are sometimes used.. 285

11 General Methods for Testing Hypotheses 287
11.1 Likelihood Ratio Tests 288
11.1.1 The likelihood ratio may be used to test $H_0: \theta = \theta_0$. 288
11.1.2 P-values for the likelihood ratio test of $H_0: \theta = \theta_0$ may be obtained from the χ^2 distribution or by simulation. 290
11.1.3 The likelihood ratio test of H_0: $(\omega, \theta) = (\omega, \theta_0)$ plugs in the MLE of ω, obtained under H_0.. 291
11.1.4 The likelihood ratio test reproduces, exactly or approximately, many commonly-used significance tests. 293
11.1.5 The likelihood ratio test is optimal for simple hypotheses. 293
11.1.6 To evaluate alternative non-nested models the likelihood ratio statistic may be adjusted for parameter dimensionality.. 294
11.2 Permutation and Bootstrap Tests 297
11.2.1 Permutation tests consider all possible permutations of the data that would be consistent with the null hypothesis. 297
11.2.2 The bootstrap samples with replacement.. 300
11.3 Multiple Tests 301
11.3.1 When multiple independent data sets are used to test the same hypothesis, the p-values are easily combined. 301
11.3.2 When multiple hypotheses are considered, statistical significance should be adjusted. 302

12 Linear Regression 309
12.1 The Linear Regression Model 310
12.1.1 Linear regression assumes linearity of $f(x)$ and independence of the noise contributions at the various observed x values. 315
12.1.2 The relative contribution of the linear signal to the total response variation is summarized by R^2. 316
12.1.3 Theory shows that if the model were correct then the least-squares estimate would be likely to be accurate for large samples. 318
12.2 Checking Assumptions 319
12.2.1 Residuals should represent unstructured noise. 319
12.2.2 Graphical examination of (x, y) data can yield crucial information.. 320
12.2.3 Failure of independence among the errors can have substantial consequences. 321
12.3 Evidence of a Linear Trend 323
12.3.1 Confidence intervals for slopes are based on SE, according to the general formula. 323
12.3.2 Evidence in favor of a linear trend can be obtained from a t-test concerning the slope.. 325
12.3.3 The fitted relationship may not be accurate outside the range of the observed data. 326
12.4 Correlation and Regression 327
12.4.1 The correlation coefficient is determined by the regression coefficient and the standard deviations of x and y. 327
12.4.2 Association is not causation. 328
12.4.3 Confidence intervals for ρ may be based on a transformation of r. 328
12.4.4 When noise is added to two variables, their correlation diminishes. 330
12.5 Multiple Linear Regression 332
12.5.1 Multiple regression estimates the linear relationship of the response with each explanatory variable, while adjusting for the other explanatory variables. 334
12.5.2 Response variation may be decomposed into signal and noise sums of squares. 335
12.5.3 Multiple regression may be formulated concisely using matrices. 339
12.5.4 The linear regression model applies to polynomial regression and cosine regression. 346

12.5.5 Effects of correlated explanatory variables cannot be interpreted separately.. 350
12.5.6 In multiple linear regression interaction effects are often important. 352
12.5.7 Regression models with many explanatory variables often can be simplified. 353
12.5.8 Multiple regression can be treacherous.. 358

13 Analysis of Variance. 361
13.1 One-Way and Two-Way ANOVA 361
13.1.1 ANOVA is based on a linear model.. 363
13.1.2 One-way ANOVA decomposes total variability into average group variability and average individual variability, which would be roughly equal under the null hypothesis. 365
13.1.3 When there are only two groups, the ANOVA F-test reduces to a t-test.. 368
13.1.4 Two-way ANOVA assesses the effects of one factor while adjusting for the other factor.. 369
13.1.5 When the variances are inhomogeneous across conditions a likelihood ratio test may be used.. 371
13.1.6 More complicated experimental designs may be accommodated by ANOVA.. 371
13.1.7 Additional analyses, involving multiple comparisons, may require adjustments to p-values. 372
13.2 ANOVA as Regression 374
13.2.1 The general linear model includes both regression and ANOVA models. 374
13.2.2 In multi-way ANOVA, interactions are often of interest. 377
13.2.3 ANOVA comparisons may be adjusted using analysis of covariance. 380
13.3 Nonparametric Methods 381
13.3.1 Distribution-free nonparametric tests may be obtained by replacing data values with their ranks. 382
13.3.2 Permutation and bootstrap tests may be used to test ANOVA hypotheses.. 385
13.4 Causation, Randomization, and Observational Studies. 385
13.4.1 Randomization eliminates effects of confounding factors. 385
13.4.2 Observational studies can produce substantial evidence. 387

14 Generalized Linear and Nonlinear Regression 391
14.1 Logistic Regression, Poisson Regression, and Generalized Linear Models 392
14.1.1 Logistic regression may be used to analyze binary responses.. 392
14.1.2 In logistic regression, ML is used to estimate the regression coefficients and the likelihood ratio test is used to assess evidence of a logistic-linear trend with x. 395
14.1.3 The logit transformation is one among many that may be used for binomial responses, but it is the most commonly applied. 398
14.1.4 The usual Poisson regression model transforms the mean λ to $\log \lambda$. 400
14.1.5 In Poisson regression, ML is used to estimate coefficients and the likelihood ratio test is used to examine trends. 401
14.1.6 Generalized linear models extend regression methods to response distributions from exponential families. 402
14.2 Nonlinear Regression 405
14.2.1 Nonlinear regression models may be fitted by least squares.. 405
14.2.2 Generalized nonlinear models may be fitted using maximum likelihood. 409
14.2.3 In solving nonlinear optimization problems, good starting values are important, and it can be helpful to reparameterize. 411

15 Nonparametric Regression 413
15.1 Smoothers 414
15.1.1 Linear smoothers are fast. 415
15.1.2 For linear smoothers, the fitted function values are obtained via a "hat matrix," and it is easy to apply propagation of uncertainty. 415
15.2 Basis Functions 416
15.2.1 Splines may be used to represent complicated functions.. 418
15.2.2 Splines may be fit to data using linear models. 418
15.2.3 Splines are also easy to use in generalized linear models. 421
15.2.4 With regression splines, the number and location of knots controls the smoothness of the fit. 422

15.2.5 Smoothing splines are splines with knots at each x_i, but with reduced coefficients obtained by penalized ML. 423
15.2.6 A method called BARS chooses knot sets automatically, according to a Bayesian criterion. 424
15.2.7 Spline smoothing may be used with multiple explanatory variables. 425
15.2.8 Alternatives to splines are often used in nonparametric regression. 427
15.3 Local Fitting 429
15.3.1 Kernel regression estimates $f(x)$ with a weighted mean defined by a pdf. 430
15.3.2 Local polynomial regression solves a weighted least squares problem with weights defined by a kernel. 432
15.3.3 Theoretical considerations lead to bandwidth recommendations for linear smoothers. . 434
15.4 Density Estimation 435
15.4.1 Kernels may be used to estimate a pdf. 435
15.4.2 Other nonparametric regression methods may be used to estimate a pdf. 436

16 Bayesian Methods. 439
16.1 Posterior Distributions. 440
16.1.1 Bayesian inference equates descriptive and epistemic probability. 441
16.1.2 Conjugate priors are convenient. 442
16.1.3 For exponential families with conjugate priors the posterior mean is a weighted combination of the MLE and the prior mean. 443
16.1.4 There is no compelling choice of prior distribution. . . 447
16.1.5 For large samples, posteriors are approximately normal and centered at the MLE. 448
16.1.6 Powerful methods exist for computing posterior distributions. 450
16.2 Latent Variables 457
16.2.1 Hierarchical models produce estimates of related quantities that are pulled toward each other. 459
16.2.2 For hierarchical models, posterior distributions are often computed by Gibbs sampling. 466
16.2.3 Penalized regression may be viewed as Bayesian estimation. 469

16.2.4 State-space models allow parameters to evolve dynamically. . . . 470
16.2.5 The Kalman filter may be used to estimate state variables for linear Gaussian state-space models. . . . 473
16.3 Bayes Factors. . . . 475
16.3.1 Bayes factors can provide evidence in favor of hypotheses. . . . 477
16.3.2 Bayes factors provide an interpretation of scientific progress. . . . 480
16.3.3 Bayes factors can be difficult to use when there is little information about unknown parameters. . . . 481
16.3.4 Bayes factors can be used to calibrate p-values. . . . 482
16.4 Derivations of Results on Latent Variables . . . 482

17 Multivariate Analysis . . . 491
17.1 Introduction . . . 491
17.2 Multivariate Analysis of Variance . . . 492
17.2.1 MANOVA provides a multivariate extension of ANOVA. . . . 492
17.2.2 When the variance matrices across conditions are unequal, the likelihood ratio test may be applied. . . . 496
17.3 Dimensionality Reduction . . . 498
17.3.1 A variance matrix may be decomposed into principal components. . . . 498
17.3.2 Methods other than PCA may be used to reduce dimensionality. . . . 503
17.4 Classification and Clustering . . . 505
17.4.1 Bayes classifiers for multivariate normal distributions take a simple form. . . . 505
17.4.2 Bayes classifiers are not always practical. . . . 506
17.4.3 Multivariate observations may be clustered into groups. . . . 510

18 Time Series . . . 513
18.1 Introduction . . . 513
18.2 Time Domain and Frequency Domain. . . . 518
18.2.1 Fourier analysis is one of the great achievements of mathematical science. . . . 522
18.2.2 The periodogram is both a scaled representation of contributions to R^2 from harmonic regression and a scaled power function associated with the discrete Fourier transform of a data set. . . . 525
18.2.3 Autoregressive models may be fitted by lagged regression. . . . 530

18.3 The Periodogram for Stationary Processes 535
18.3.1 The periodogram may be considered an estimate of the spectral density function. 535
18.3.2 For large samples, the periodogram ordinates computed from a stationary time series are approximately independent of one another and chi-squared distributed. 537
18.3.3 Consistent estimators of the spectral density function result from smoothing the periodogram. 539
18.3.4 Linear filters can be fast and effective. 541
18.3.5 Frequency information is limited by the sampling rate. 544
18.3.6 Tapering reduces the leakage of power from non-Fourier to Fourier frequencies. 546
18.3.7 Time-frequency analysis describes the evolution of rhythms across time. 548
18.4 Propagation of Uncertainty for Functions of the Periodogram 550
18.4.1 Confidence intervals and significance tests may be carried out by propagating the uncertainty from the periodogram. 550
18.4.2 Uncertainty about functions of time series may be obtained from time series pseudo-data. 552
18.5 Bivariate Time Series 553
18.5.1 The coherence $\rho_{XY}(\omega)$ between two series X and Y may be considered the correlation of their ω-frequency components. 555
18.5.2 In examining cross-correlation or coherence of two time series it is advisable first to pre-whiten the series. 557
18.5.3 Granger causality measures the linear predictability of one time series by another. 559

19 Point Processes 563
19.1 Point Process Representations 566
19.1.1 A point process may be specified in terms of event times, inter-event intervals, or event counts. 566
19.1.2 A point process may be considered, approximately, to be a binary time series. 567
19.1.3 Point processes can display a wide variety of history-dependent behaviors 568

19.2 Poisson Processes 570
19.2.1 Poisson processes are point processes for which event probabilities do not depend on occurrence or timing of past events. 570
19.2.2 Inhomogeneous Poisson processes have time-varying intensities. 573
19.3 Non-Poisson Point Processes 578
19.3.1 Renewal processes have i.i.d. inter-event waiting times. 578
19.3.2 The conditional intensity function specifies the joint probability density of spike times for a general point process. 582
19.3.3 The marginal intensity is the expectation of the conditional intensity. 584
19.3.4 Conditional intensity functions may be fitted using Poisson regression. 586
19.3.5 Graphical checks for departures from a point process model may be obtained by time rescaling. . . . 594
19.3.6 There are efficient methods for generating point process pseudo-data. 597
19.3.7 Spectral analysis of point processes requires care. . . . 599
19.4 Additional Derivations 601

Appendix: Mathematical Background 605

References 623

Example Index 635

Index 639

Chapter 1
Introduction

1.1 Data Analysis in the Brain Sciences

The brain sciences seek to discover mechanisms by which neural activity is generated, thoughts are created, and behavior is produced. What makes us see, hear, feel, and understand the world around us? How can we learn intricate movements, which require continual corrections for minor variations in path? What is the basis of memory, and how do we allocate attention to particular tasks? Answering such questions is the grand ambition of this broad enterprise and, while the workings of the nervous system are immensely complicated, several lines of now-classical research have made enormous progress: essential features of the nature of the action potential, of synaptic transmission, of sensory processing, of the biochemical basis of memory, and of motor control have been discovered. These advances have formed conceptual underpinnings for modern neuroscience, and have had a substantial impact on clinical practice. The method that produced this knowledge, the scientific method, involves both observation and experiment, but always a careful consideration of the data. Sometimes results from an investigation have been largely qualitative, as in Brenda Milner's documentation of implicit memory retention, together with explicit memory loss, as a result of hippocampal lesioning in patient H.M. In other cases quantitative analysis has been essential, as in Alan Hodgkin and Andrew Huxley's modeling of ion channels to describe the production of action potentials. Today's brain research builds on earlier results using a wide variety of modern techniques, including molecular methods, patch clamp recording, two-photon imaging, single and multiple electrode studies producing spike trains and/or local field potentials (LFPs), optical imaging, electroencephalography (producing EEGs), and functional imaging—positron emission tomography(PET), functional magnetic resonance imaging (fMRI), magnetoencephalography (MEG)—as well as psychophysical and behavioral studies. All of these rely, in varying ways, on vast improvements in data storage, manipulation, and display technologies, as well as corresponding advances in analytical techniques. As a result, data sets from current investigations are often much larger, and more com-

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_1,

plicated, than those of earlier days. For a contemporary student of neuroscience, a working knowledge of basic methods of data analysis is indispensable.

The variety of experimental paradigms across widely ranging investigative levels in the brain sciences may seem intimidating. It would take a multi-volume encyclopedia to document the details of the myriad analytical methods out there. Yet, for all the diversity of measurement and purpose, there are commonalities that make analysis of neural data a single, circumscribed and integrated subject. A relatively small number of principles, together with a handful of ubiquitous techniques—some quite old, some much newer—lay a solid foundation. One of our chief aims in writing this book has been to provide a coherent framework to serve as a starting point in understanding all types of neural data.

In addition to providing a unified treatment of analytical methods that are crucial to progress in the brain sciences, we have a secondary goal. Over many years of collaboration with neuroscientists we have observed in them a desire to learn all that the data have to offer. Data collection is demanding, and time-consuming, so it is natural to want to use the most efficient and effective methods of data analysis. But we have also observed something else. Many neuroscientists take great pleasure in displaying their results not only because of the science involved but also because of the *manner in which* particular data summaries and displays are able to shed light on, and explain, neuroscientific phenomenon; in other words, they have developed a refined appreciation for the data-analytic process itself. The often-ingenious ways investigators present their data have been instructive to us, and have reinforced our own aesthetic sensibilities for this endeavor. There is deep satisfaction in comprehending a method that is at once elegant and powerful, that uses mathematics to describe the world of observation and experimentation, and that tames uncertainty by capturing it and using it to advantage. We hope to pass on to readers some of these feelings about the role of analytical techniques in illuminating and articulating fundamental concepts.

A third goal for this book comes from our exposure to numerous articles that report data analyzed largely by people who lack training in statistics. Many researchers have excellent quantitative skills and intuitions, and in most published work statistical procedures appear to be used correctly. Yet, in examining these papers we have been struck repeatedly by the absence of what we might call statistical thinking, or application of *the statistical paradigm*, and a resulting loss of opportunity to make full and effective use of the data. These cases typically do not involve an incorrect application of a statistical method (though that sometimes does happen). Rather, the lost opportunity is a failure to follow the *general approach* to the analysis of the data, which is what we mean by the label "the statistical paradigm." Our final pedagogical goal, therefore, is to lay out the key features of this paradigm, and to illustrate its application in diverse contexts, so that readers may absorb its main tenets.

To begin, we will review several essential points that will permeate the book. Some of these concern the nature of neural data, others the process of statistical reasoning. As we go over the basic issues, we will introduce some data that will be used repeatedly.

1.1.1 Appropriate analytical strategies depend crucially on the purpose of the study and the way the data are collected.

The answer to the question, "How should I analyze my data?" always depends on what you want to know. Convenient summaries of the data are used to convey apparent tendencies. Particular summaries highlight particular aspects of the data—but they ignore other aspects. At first, the purpose of an investigation may be stated rather vaguely, as in "I would like to know how the responses differ under these two experimental conditions." This by itself, however, is rarely enough to proceed. Usually there are choices to be made, and figuring out what analysis should be performed requires a sharpening of purpose.

Example 1.1 SEF neural activity under two conditions Olson et al. (2000) examined the behavior of neurons in the supplementary eye field (SEF), which is a frontal lobe region anterior to, and projecting to, the eye area in motor cortex. The general issue was whether the SEF merely relays the message to move the eyes, or whether it is involved in some higher-level processing. To distinguish these two possibilities, an experiment was devised in which a monkey moved its eyes in response to either an explicit external cue (the point to which the eyes were to move was illuminated) or an internally-generated translation of a complex cue (a particular pattern at fixation point determined the location to which the monkey was to move its eyes). If the SEF simply transmits the movement message to motor cortex and other downstream areas, one would expect SEF neurons to behave very similarly under the two experimental conditions. On the other hand, distinctions between the neural responses in the two conditions would indicate that the SEF is involved in higher-level cognitive processing. While an individual neuron's activity was recorded from the SEF of an alert macaque monkey, one of the two conditions was chosen at random and applied. This experimental protocol was repeated many times, for each of many neurons. Thus, for each recorded neuron, under each of the two conditions, there were many *trials*, which consist of experimental repetitions designed to be as close to identical as possible.

Results for one neuron are given in Fig. 1.1. The figure displays a pair of raster plots and peri-stimulus time histograms (PSTHs). Each line in each raster plot contains results from a single trial, which consist of a sequence of times at which action potentials or *spikes* occur. The sequence is usually called a *spike train*. Note that for each condition the number and timing of the spikes, displayed on the many lines of each raster plot, vary from trial to trial. The PSTH is formed by creating time bins (here, each bin is 10 ms in length), counting the total number of spikes that occur across all trials within each bin, and then normalizing (by dividing by the number of trials and the length of each bin in seconds) to convert count to firing rate in units of spikes per second. The PSTH is used to display firing-rate trends across time, which are considered to be common to[1] the many separate trials.

[1] One source of variation across trials is that the behavior of the monkey is not identical on every trial. For instance, the eyes may move along slightly different paths and at different rates. Even

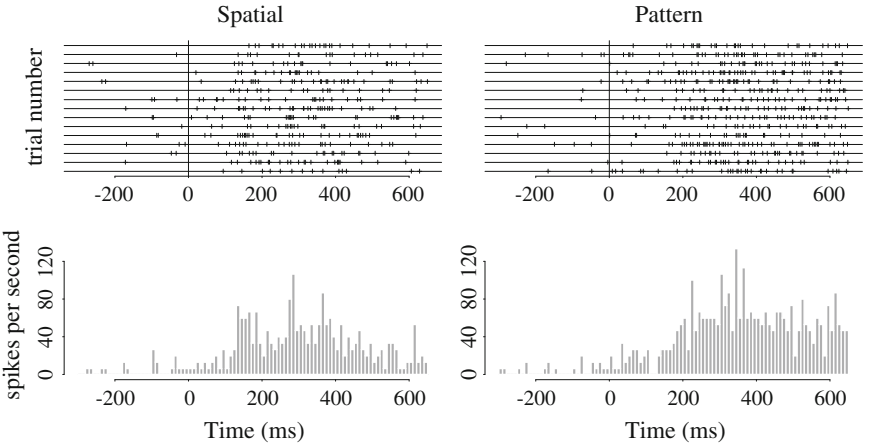

Fig. 1.1 Raster plot (*Top*) and PSTH (*Bottom*) for an SEF neuron under both the external-cue or "spatial" condition (*Left*) and the complex cue or "pattern" condition (*Right*). Each line in each raster plot contains data from a single trial, that is, a single instance in which the condition was applied. (There are 15 trials for each condition.) The tick marks represent spike times, i.e., times at which the neuron fired. The PSTH contains normalized spike counts within 10 ms time bins: for each time bin the number of spikes occurring in that bin is counted across all trials; this count is then divided by the number of trials, and the width of the time bin in seconds, which results in firing rate in units of spikes per second. Time is measured relative to presentation of a visual cue, which is considered time $t = 0$. This neuron is more active under the pattern condition, several hundred milliseconds post cue. The increase in activity may be seen from the raster plots, but is more apparent from comparison of the PSTHs.

Visual comparison of the two raster plots and two PSTHs in Fig. 1.1 indicates that this neuron tends to respond more strongly under the pattern condition than under the spatial condition, at least toward the end of the trial. But such qualitative impressions are often insufficiently convincing even for a single neuron; furthermore, results for many dozens of neurons need to be reported. How should they be summarized? Should the firing rates be averaged over a suitable time interval, and then compared? If so, which interval should be used? Might it be useful to display the firing-rate histograms on top of each other somehow, for better comparison, and might the distinctions between them be quantified and then summarized across all neurons? Might it be useful to compare the peak firing rates for the two neurons, or the time at which the peaks occurred? All of these variations involve different ways to look at the data, and each effectively defines differently the purpose of the study.

The several possible ways of examining firing rate, just mentioned, have in common the aggregation of data across trials. A quite different idea would be to examine the relationship of neural spiking activity and reaction time, on a trial-by-trial

(Footnote 1 continued)
in preparations in vitro, however, identical current inputs to a neuron do not necessarily produce identical spiking outputs. This is due, at least in part, to the stochastic behavior of the movements of ions and molecules that govern the spiking mechanism.

basis, and then to see how that changes across conditions. This intriguing possibility, however, would require a different experiment: in the experiment of Olson et al. the eye movement occurred long after[2] the cue, so there was no observed behavior corresponding to reaction time. This is an extreme case of the way analytical alternatives depend on the purpose of the experiment. □

Example 1.1 illustrates the way a particular purpose shaped the design of the experiment, the way the data were collected, and the possible analytic strategies. In thinking about the way the data are collected, one particular distinction is especially important: that of *steady-state* versus systematically evolving conditions. In many studies, an experimental manipulation leads to a measured response that evolves in a more-or-less predictable way over time. In Example 1.1 the neuronal firing rate, as represented by the PSTH, evolves over time, with the firing rate increasing roughly 200 ms after the cue. This may be contrasted with observation of a phenomenon that has no predictable time trend, experimentally-induced or otherwise. Typically, such situations arise when one is making baseline measurements, in which some indicator of neural activity is observed while the organism or isolated tissue is at rest and receives no experimental stimulus.[3] Sometimes a key piece of laboratory apparatus must be observed in steady state to establish background conditions. Here is an important example.

Example 1.2 MEG background noise Magnetoencephalography (MEG) is an imaging technique used to measure the magnetic fields produced by electrical activity in the brain. MEG recordings are used clinically to localize a brain tumor or to identify the site of an epileptic focus; they are used by neuroscientists to study such things as language production, memory formation, and the neurological basis of diseases such as schizophrenia.

The MEG signals are generated from the net effect of ionic currents flowing in the dendrites of cortical neurons during synaptic transmission. From Maxwell's equations, any electrical current produces a magnetic field oriented orthogonally (perpendicularly) to the current flow, according to the right-hand rule. MEG measures this magnetic field. Magnetic fields are relatively unaffected by the tissue through which the signal passes on the way to a detector, but the signals are very weak. Two things make detection possible. One is that MEG uses highly sensitive detectors called superconducting quantum interference devices (SQUIDs). The second is that currents from many neighboring neuronal dendrites have similar orientations, so that their magnetic fields reinforce each other. The dendrites of pyramidal cells in the cortex are generally perpendicular to its surface and, in many parts of the brain, their generated fields are oriented outward, toward the detectors sitting outside the head.

---

[2] They used a random delay followed by a separate cue to move; this helped ensure that movement and anticipatory effects would not contaminate the processing effects of interest.

[3] Analyses of brain activity when the subject is resting (e.g., during passive eye fixation or with eyes closed) have been reported by many groups. See, for example, Fox et al. (2005), who used fMRI to describe two distinct resting-state networks.

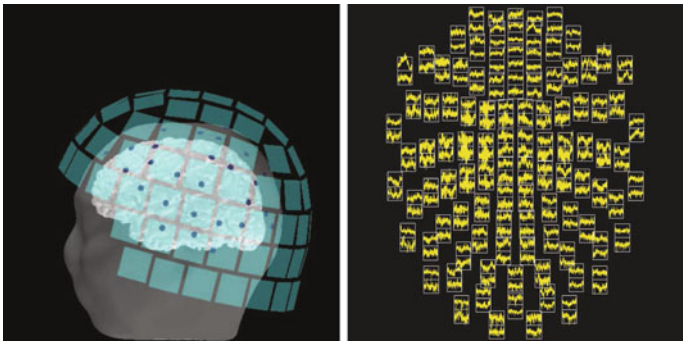

Fig. 1.2 MEG imaging. *Left* drawing of the way the SQUID detectors sit above the head in a MEG machine. *Right* plots of sensor signals laid out in a two-dimensional configuration to correspond, roughly, to their three-dimensional locations as shown in the *left* panel of the figure.

A detectable MEG signal is produced by the net effects of currents from approximately 50,000 active neurons. See Fig. 1.2.

Because the signals are weak, and the detectors extremely sensitive, it is important to assess MEG activity prior to imaging patients. Great pains are taken to remove sources of magnetic fields from the room in which the detector is located. Nonetheless, there remains a background signal that must be identified under steady-state conditions. □

Many analytical methods assume a steady state exists. The mathematical formulation of "steady state," based on *stationarity*, will be discussed in Chapter 18.

1.1.2 Many investigations involve a response to a stimulus or behavior.

In contrast to the steady state conditions in Example 1.2, many experiments involve perturbation or stimulation of a system, producing a temporally evolving response. This does *not* correspond to a steady state. The SEF experiment was a stimulus-response study. Functional imaging also furnishes good examples.

Example 1.3 fMRI in a visuomotor experiment Functional magnetic resonance imaging (fMRI) uses change in magnetic resonance (MR) to infer change in neural activity, within small patches (voxels) of brain tissue. When neurons are active they consume oxygen from the blood, which produces a local increase in blood flow after a delay of several seconds. Oxygen in the blood is bound to hemoglobin, and the magnetic resonance of hemoglobin changes when it is oxygenated. By using an appropriate MR pulse sequence, the change in oxygenation can be detected as the blood-oxygen-level dependent (BOLD) signal, which follows a few seconds after the increase in neural activity. The relationship between neural activity and BOLD is not

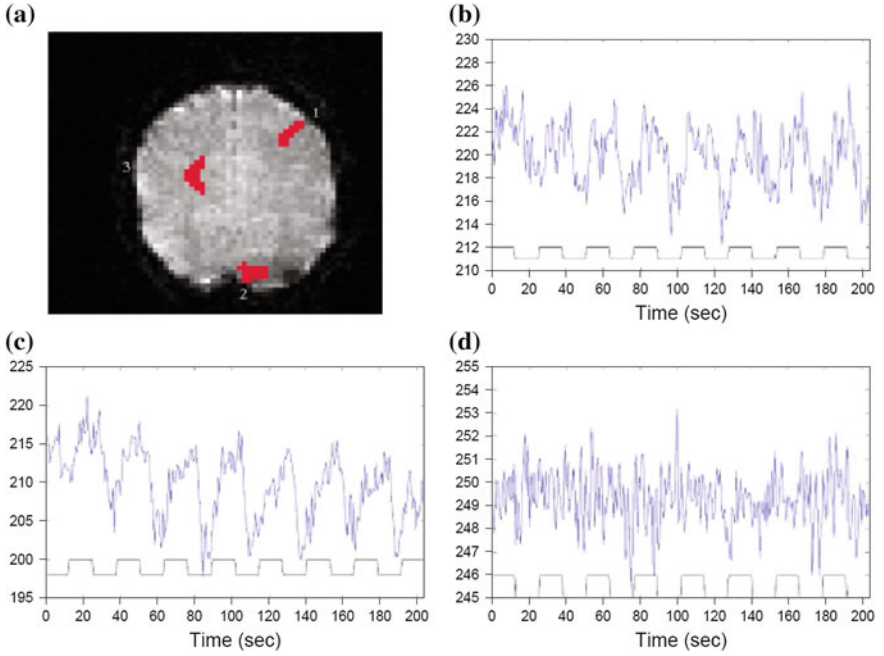

Fig. 1.3 An fMRI image with several traces of the signal across time. Panel A displays an image indicating three locations, shown in *red*, from which voxel signals were examined. Panels B-D display the signals themselves, averaged across the voxels. They correspond, respectively, to motor cortex, primary visual cortex, and white matter.

known in detail, but since the 1990s fMRI has been used to track changes in BOLD in relation to the execution of a task, giving at least a rough guide to the location of sustained functional neural activity.

Figure 1.3 displays images from one subject in a combined visual and motor fMRI experiment. The subject was presented with a full-field flickering checkerboard, in a repeating pattern of 12.8 s (seconds) OFF followed by 12.8 s ON. This was repeated 8 times. Alternating out of phase with the flickering checkerboard pattern the subject also executed a finger tapping task (12.8 s ON followed by 12.8 s OFF). The brain was imaged once every 800 ms for the duration of the experiment. The slice shown was chosen to transect both the visual and motor cortices. Three regions of interest have been selected, corresponding to (1) motor (2) visual cortex, and (3) white matter. Parts B through D of the figure illustrate the raw time series taken from each of these regions, along with timing diagrams of the input stimuli. As expected, the motor region is more active during finger tapping (but the BOLD signal responds several seconds after the tapping activity commences) while the visual region is more active during the flickering visual image (again with several seconds lag). The response within white matter serves as a control. □

The focus of stimulus-response experiments is usually the relationship between stimulus and response. This may suggest strategies for analysis of the data. If we

let X denote the stimulus and Y the response, we might write the relationship as follows:

$$Y \longleftarrow X \tag{1.1}$$

where the arrow indicates that X leads to Y. Chapters 12, 14, and 15 are devoted to regression methods, which are designed for situations in which X might lead to Y.

In Example 1.1, Y could be the average firing rate in a specified window of time, such as 200–600 ms following the cue, and X could represent the experimental condition. In other words, the particular experimental condition leads to a corresponding average firing rate. In Example 1.3, Y could be the value of the BOLD response, and X could represent whether the checkerboard was on or off 5 s prior to the response Y.

The arrow in (1.1) suggests a mechanistic relationship (the stimulus occurred, and that made Y occur), but it is often wise to step back and remain agnostic about a causal connection. A more general notion is that the variables X and Y are *associated*, meaning that they tend to vary together. A wide variety of neuroscientific studies seek to establish associations among variables. Such studies might relate a pair of behavioral measures, for example, or they might involve spike trains from a pair of neurons recorded simultaneously, EEGs from a pair of electrodes on the scalp, or MEG signals from a pair of SQUID detectors. Many statistical tools apply to both causal and non-causal relationships. Measures of association are discussed in Chapters 10 and 12. Chapter 13 also contains a brief discussion of the distinction between association and causation, and some issues to consider when one wishes to infer causation from an observed association.

1.2 The Contribution of Statistics

Many people think of statistics as a collection of particular data-analytic techniques, such as analysis of variance, chi-squared goodness-of-fit, linear regression, etc. And so it is. But the field of statistics, as an academically specialized discipline, strives for something much deeper, namely, the development and characterization of data collection and analysis methods according to well-defined principles, as a means of quantifying knowledge about underlying phenomena and rationalizing the learning and decision-making process. As we said above, one of the main pedagogical goals of this book is to impart to the reader some sense of the way data analytic issues are framed within the discipline of statistics. In trying to achieve this goal, we find it helpful to articulate the nature of the statistical paradigm as concisely as possible. After numerous conversations with colleagues, we have arrived at the conclusion that among many components of the statistical paradigm, summarized below, two are the most fundamental.

Two Fundamental Tenets of the Statistical Paradigm:

1. Statistical models are used to express knowledge and uncertainty about a signal in the presence of noise, via inductive reasoning.
2. Statistical methods may be analyzed to determine how well they are likely to perform.

In the remainder of this section we will elaborate, adding a variety of comments and clarifications.

1.2.1 Statistical models describe regularity and variability of data in terms of probability distributions.

When data are collected, repeatedly, under conditions that are as nearly identical as an investigator can make them, the measured responses nevertheless exhibit variation. The spike trains generated by the SEF neuron in Example 1.1 were collected under experimental conditions that were essentially identical; yet, the spike times, and the number of spikes, varied from trial to trial. The most fundamental principle of the statistical paradigm, its starting point, is that this variation may be described by probability. Chapters 3 and 5 are devoted to spelling out the details, so that it will become clear what we mean when we say that probability describes variation. But the idea is simple enough: probability describes familiar games of chance, such as rolling dice, so when we use probability also to describe variation, we are making an analogy; we do not know all the reasons why one measurement is different than another, so it is *as if* the variation in the data were generated by a gambling device. Let us consider a simple but interesting example.

Example 1.4 Blindsight in patient P.S. Marshall and Halligan (1988) reported an interesting neuropsychological finding from a patient, identified as P.S. This patient was a 49 year-old woman who had suffered damage to her right parietal cortex that reduced her capacity to process visual information coming from the left side of her visual space. For example, she would frequently read words incorrectly by omitting left-most letters ("smile" became "mile") and when asked to copy simple line drawings, she accurately drew the right-hand side of the figures but omitted the left-hand side without any conscious awareness of her error. To show that she could actually see what was on the left but was simply not responding to it—a phenomenon known as *blindsight*—the examiners presented P.S. with a pair of cards showing identical green line drawings of a house, except that on one of the cards bright red flames were depicted on the left side of the house. They presented to P.S. both cards, one above the other (the one placed above being selected at random), and asked her to choose which house she would prefer to live in. She thought this was silly "because they're the same" but when forced to make a response chose the non-burning house on 14 out of 17 trials. This would seem to indicate that she did, in

fact, see the left side of the drawings but was unable to fully process the information. But how convincing is it that she chose the non-burning house on 14 out of 17 trials? Might she have been guessing?

If, instead, P.S. had chosen 17 out of 17 trials there would have been very strong evidence that her processing of the visual information affected her decision-making, while, on the other hand, a choice of 9 out of 17 clearly would have been consistent with guessing. The intermediate outcome 14 out of 17 is of interest as a problem in data analysis and scientific inference precisely because it feels fairly convincing, but leaves us unsure: a thorough, quantitative analysis of the uncertainty would be very helpful.

The standard way to begin is to recognize the variability in the data, namely, that P.S. did not make the same choice on every trial; we then say that the choice made by P.S. on each trial was a random event, that the probability of her choosing the non-burning house on each trial was a value p, and that the responses on the different trials were independent of each other. These three assumptions use probability to describe the variability in the data. Once these three assumptions are made it becomes possible to quantify the uncertainty about p and the extent to which the data are inconsistent with the value $p = .5$, which would correspond to guessing. In other words, it becomes possible to make statistical inferences. □

The key step in Example 1.4 is the introduction of probability to describe variation. Once that first step is taken, the second step of making inferences about the phenomenon becomes possible. Because the inferences are statistical in nature, and they require the introduction of probability, we usually refer to the probability framework—with its accompanying assumptions—as a *statistical model.* Statistical models provide a simple formalism for describing the way the repeatable, regular features of the data are combined with the variable features. In Example 1.4 we may think of p as the propensity for P.S. to choose the non-burning house. According to this statistical model, p is a kind of regularity in the data in the sense that it is unchanging across trials. The variation in the data comes from the probabilistic nature of the choice: what P.S. will choose is somewhat unpredictable, so we attribute a degree of uncertainty to unknown causes and describe it as if predicting her choice were a game of chance. We elaborate on the statistical model, and the inferences drawn from the data of Example 1.4 in Chapters 5 and 7.

Probability is also often introduced to describe small fluctuations around a specified formula or "law." We typically consider such fluctuations "noise," in contrast to the systematic part of the variation in some data, which we call the "signal." For instance, as we explain in Chapter 12, when the underlying, systematic mathematical specification (the signal) has the form

$$y = f(x)$$

we will replace it with a statistical model having the form

$$Y = f(x) + \epsilon \tag{1.2}$$

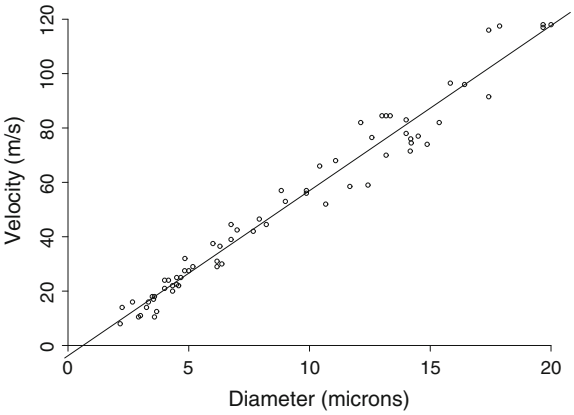

Fig. 1.4 Conduction velocity of action potentials, as a function of diameter. The x-axis is diameter in microns; the y-axis is velocity in meters per second. Based on Hursh (1939, Fig. 2). Also shown is the least-squares regression line.

where ϵ represents noise and the variable Y is capitalized to indicate its now-random nature: it becomes "signal plus noise." The simplest case occurs when $f(x)$ is a line, having the form $f(x) = \beta_0 + \beta_1 x$, where we use coefficients β_0 and β_1 (instead of writing $f(x) = a + bx$) to conform to statistical convention. Here is an example.

Example 1.5 Neural conduction velocity Hursh (1939) presented data on the relationship between a neuron's conduction velocity and its axonal diameter, in adult cats. Hursh measured maximal velocity among fibers in several nerve bundles, and then also measured the diameter of the largest fiber in the bundle. The resulting data, together with a fitted line, are shown in Fig. 1.4. In this case the line $y = \beta_0 + \beta_1 x$ represents the approximate linear relationship between maximal velocity y and diameter x. The data follow the line pretty closely, with the intercept β_0 being nearly equal to zero. This implies, for example, that if one fiber has twice the diameter of another, the first will propagate an action potential roughly twice as fast as the second. □

Before we conclude our introductory remarks about statistical models, by elaborating on (1.2), let us digress for a moment to discuss the method used to fit the line to the data in Fig. 1.4, which is called *least squares regression*, It is one of the core conceptions of statistics, and we discuss it at length in Chapter 12.

Suppose we have a line that is fit by some method, possibly least-squares or possibly another method, and let us write this line as $y = \beta_0^* + \beta_1^* x$. It is customary, in statistics, to use the notations β_0 and β_1 for the intercept and slope. Here we have included the asterisk * in β_0^* and β_1^* because it will simplify some additional notations later on. The important thing is that β_0^* and β_1^* are coefficients that define the line we fit to the data, using whatever method we might choose. Suppose there are n data pairs of the form (x, y) and let us label them with a subscript so that they take the form (x_i, y_i) with $i = 1, 2, \ldots, n$. That is, (x_1, y_1) would be the first data

pair, (x_2, y_2) the second, and so forth. The y-coordinate on the line $y = \beta_0^* + \beta_1^* x$ corresponding to x_i is

$$\hat{y}_i^* = \beta_0^* + \beta_1^* x_i.$$

The number $\hat{y}_i^*$ is called the *fitted value* at x_i and we may think of it as predicting y_i. We then define the ith *residual* as

$$e_i = y_i - \hat{y}_i^*.$$

The value e_i is the error at x_i in fitting, or the error of prediction, i.e., it is the vertical distance between the observation (x_i, y_i) and the line at x_i. We wish to find the line that best predicts the y_i values, which means we want to make the e_i's as small as possible, in aggregate. To do this, we have to minimize some measure of the size of all the e_i's taken together. In choosing such a measure we assume positive and negative values of the residuals are equally important. Two alternative aggregate measures that treat e_i and $-e_i$ equally are the following:

$$\begin{aligned} \text{sum of absolute deviations} &= \sum_{i=1}^{n} |e_i| \\ \text{sum of squares} &= \sum_{i=1}^{n} e_i^2. \end{aligned} \tag{1.3}$$

Data analysts sometimes choose β_0^* and β_1^* to minimize the sum of absolute deviations, but the solution can not be obtained in closed form, and it is harder to analyze mathematically. Instead, the method of least squares works with the sum of squares, where the solution may be found using calculus (see Chapter 12).

The least-squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are the values of β_0^* and β_1^* that minimize the sum of squares in (1.3). The least-squares line is then

$$y = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Having motivated least-squares with (1.2) let us return to that equation and note that it is not yet a statistical model. If we write

$$Y_i = f(x_i) + \epsilon_i, \tag{1.4}$$

take

$$f(x) = \beta_0 + \beta_1 x$$

and, crucially, let the noise term ϵ_i be a *random variable*, then we obtain a *linear regression model*. Random variables are introduced in Chapter 3. The key point in

the present discussion is that linear regression describes the regularity of the data by a straight line and the variability (the deviations from the line) by a probability distribution (the distribution of the noise random variable ϵ_i).

1.2.2 Statistical models are used to express knowledge and uncertainty about a signal in the presence of noise, via inductive reasoning.

The introduction of a statistical model not only provides guidance in determining fits to data, as in Example 1.5, but also assessments of uncertainty.

Example 1.4 (continued from page 9) Let us return to the question of whether the responses of P.S. were consistent with guessing. In this framework, guessing would correspond to $p = .5$ and the problem then becomes one of assessing what these data tell us about the value of p. As we will see in Chapter 7, standard statistical methods give an approximate 95 % confidence interval for p of (.64, 1.0). This is usually interpreted by saying we are 95 % confident the value of p lies in the interval (.64, 1.0), which is a satisfying result: while this interval contains a range of values, indicating considerable uncertainty, we are nonetheless highly confident that the value of p is inconsistent with guessing. □

The confidence interval we have just reported in Example 1.4 illustrates the expression of "knowledge and uncertainty." It is an example of *inductive reasoning* in the sense that we reason from the data back to the quantity p assumed in the model. Many mathematical arguments begin with a set of assumptions and *prove* some consequence. This is often called *deductive reasoning*. As described in Chapter 7, statistical theory uses deductive reasoning to provide the formalism for confidence intervals. However, when we interpret the result as providing *knowledge* about the unknown quantity p based on experience (the data), the argument is usually called "inductive." Unlike deductive reasoning, inductive reasoning is uncertain. We use probability to calibrate the degree to which a statement is likely to be true. In reporting confidence intervals, the convention is to use a probability of .95, representing a high degree of confidence.

In fact, as a conceptual advance, this expression of knowledge and uncertainty via probability is highly nontrivial: despite quite a bit of earlier mathematical attention to games of chance, it was not until the late 1700s that there emerged any clear notion of inductive, or what we would now call *statistical* reasoning; it was not until the first half of the twentieth century that modern methods began to be developed systematically; and it was only in the second half of the twentieth century that their properties were fully understood. From a contemporary perspective the key point is that the confidence interval is achieved by uniting two distinct uses of probability. The first is descriptive: saying P.S. will choose the non-burning house with probability p is analogous to saying the probability of rolling an even number with an apparently fair six-sided die is $1/2$. The second use of probability is often called "epistemic,"

and involves a statement of knowledge: saying we have 95 % confidence that p is in the interval (.64, 1.0) is analogous to someone saying they are 90 % sure that the capital of Louisiana is Baton Rouge. The fundamental insight, gained gradually over many years, is that the descriptive probability in statistical models may be used to produce epistemic statements for scientific inference. We will emphasize the contrast between the descriptive and epistemic roles of statistical models by saying that models describe the *variation* in data and produce *uncertain* inferences. Technically, there are alternative frameworks for bringing descriptive and epistemic probability together, the two principal ones being *Bayesian* and *frequentist*. We will discuss the distinction in Section 7.3.9, and develop the Bayesian approach to inference at greater length in Chapter 16.

While we wish to stress the importance of statistical models in data analysis, we also want to issue several qualifications and caveats: first, the notion of "model" we intend here is very general, the only restriction being that it must involve a probabilistic description of the data; second, modeling is done in conjunction with summaries and displays that do not introduce probability explicitly; third, it is very important to assess the fit of a model to a given set of data; and, finally, statistical models are mathematical abstractions, imposing structure on the data by introducing explicit assumptions. The next three subsections explain these points further.

1.2.3 Statistical models may be either parametric or nonparametric.

In emphasizing statistical models, our only restriction is that probability must be used to express the way regularity and variability in the data are to be understood. One very important distinction is that of *parametric* versus *nonparametric* models.

The terminology comes from the representation of a probability distribution in terms of an unknown parameter. A *parameter* is a number, or vector of numbers, that is used in the definition of the distribution; the probability distribution is characterized by the parameter in the sense that once the value of the parameter is known, the probability distribution is completely determined. In Example 1.4, p. 9, the parameter is p. In Example 1.5, p. 11, the parameter includes the pair (β_0, β_1), together with a noise variation parameter σ, explained in Chapter 12. In both of these cases the values of the unknown parameters determine the probability distribution of the random variables, as in (1.4). Parametric probability distributions are discussed in Chapter 5.

A related distinction arises in the context of y versus x models of the type considered in Example 1.5. That example involved a linear relationship. As we note in Chapters 14 and 15, the methods used to fit linear models can be generalized for nonlinear relationships. The methods in Chapter 15 are also called nonparametric because the fitted relationship is not required to follow a pre-specified form.

Example 1.6 Excitatory post-synaptic current As part of a study on spike-timing-dependent plasticity (Dr. David Nauen, personal communication), rat hippocampal

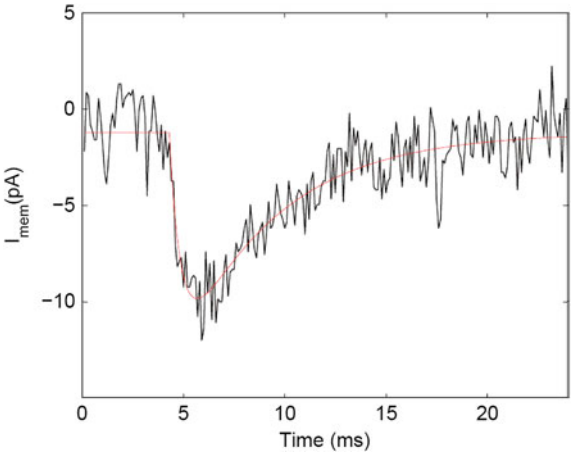

Fig. 1.5 Excitatory post-synaptic current. Current recorded from a rat hippocampal neuron, together with smoothed version (shown as the *thin red line* within the noisy current trace) obtained by fitting a suitable function of time, given in the text. The current values are connected by the *dark line*. When values recorded sequentially in time are plotted it is a common practice to connect them. (Figure courtesy of David Nauen.)

neurons were held in voltage clamp and post-synaptic currents were recorded following an action potential evoked in a presynaptic cell. Figure 1.5 displays a plot of membrane current as a function of time. One measurement of size of the current is found by integrating the current across time (which is implemented by summing the current values and multiplying by the time between observations), giving the total charge transmitted. Other quantities of interest include the onset delay, the rate at which the curve "rises" (here, a negative rise) from onset to peak current, and the rate at which the curve decays from peak current back toward steady state. The current trace is clearly subject to measurement noise, which would contaminate the calculations. A standard way to reduce the noise is to fit the data by a suitable function of time. Such a fit is also shown in the figure. It may be used to produce values for the various constants needed in the analysis. To produce the fit a statistical model of the form (1.4) was used where the function $y = f(x)$, with y being post-synaptic current and x being time, was defined as

$$f(x) = A_1(1-\exp((x-t_0)/\tau_1))\,(A_2\exp((x-t_0)/\tau_2)-(1-A_2)\exp((x-t_0)/\tau_3)).$$

This was based on a suggestion by Nielsen et al. (2004). Least squares was then applied, as defined in Section 1.2.1. The fit is good, though it distorts slightly the current trace in the dip and at the end. The advantage of using this function is that its coefficients may be interpreted and compared across experimental conditions. □

The simple linear fit in Example 1.5, p. 11, is an example of linear regression, discussed in Chapter 12, while the fit based on a combination of exponential functions

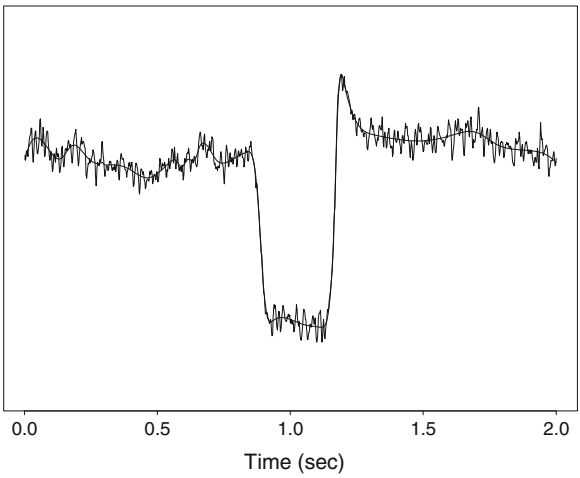

Fig. 1.6 Electrooculogram together with a smoothed, or "filtered" version that removes the noise. The method used for smoothing is an example of nonparametric regression.

in Example 1.6 is an example of *nonlinear regression* discussed in Section 14.2. Both are examples of *parametric regression* because both use specified functions based on formulas that involve a few parameters. In Example 1.5 the parameters were β_0 and β_1 while in Example 1.6 they were A_1, A_2, τ_1, τ_2, τ_3, t_0. *Nonparametric regression* is used when the formula for the function is not needed. Nonparametric regression is a central topic of Chapter 15. Here is an example.

Example 1.7 Electrooculogram smoothing for EEG artifact removal EEG recordings suffer from a variety of artifacts, one of which is their response to eye blinks. A good way to correct for eye-blink artifacts is to record potentials from additional leads in the vicinity of the eyes; such electrooculograms (EOGs) may be used to identify eye blinks, and remove their effects from the EEGs. Wallstrom et al. (2002, 2004) investigated methods for removing ocular artifacts from EEGs using the EOG signals. In Chapter 15 it will become clear how to use a general smoothing method to remove high-frequency noise. This does not require the use of a function having a specified form. Figure 1.6 displays an EOG recording together with a smoothed version of it, obtained using a nonparametric regression method known as BARS (Dimatteo et al. 2001). □

1.2.4 Statistical model building is an iterative process that incorporates assessment of fit and is preceded by exploratory analysis.

Another general point about the statistical paradigm is illustrated in Fig. 1.7. This figure shows where the statistical work fits in. Real investigations are far less sequential than depicted here, but the figure does provide a way of emphasizing two components of the process that go hand-in-hand with statistical modeling: exploratory analysis and assessment of fit. Exploratory analysis involves informal investigation of the data based on numerical or graphical summaries, such as a histogram. Exploratory results, together with judgment based on experience, help guide construction of an initial probability model to represent variability in observed data.

Every such model, and every statistical method, makes some assumptions, leading, as we have already seen, to a reduction of the data in terms of some small number of interpretable quantities. As shown in Fig. 1.7, the data may be used, again, to check the probabilistic assumptions, and to consider ramifications of departures from them. Should serious departures from the assumptions be found, a new model may be formed. Thus, probability modeling and model assessment are iterative, and only when a model is considered adequate are statistical inferences made. This process is embedded into the production of scientific conclusions from experimental results (Box et al. 1978).

1.2.5 All models are wrong, but some are useful.

The simple representation in Fig. 1.7 is incomplete and may be somewhat misleading. Most importantly, while it is true that there are standard procedures for model assessment, some of which we will discuss in Chapter 10, there is no uniformly-applicable rule for what constitutes a good fit. Statistical models, like scientific models, are abstractions and should not be considered perfect representations of the data. As examples of scientific models in neuroscience we might pick, at one extreme, the Hodgkin-Huxley model for action potential generation in the squid giant axon and, at the other extreme, being much more vague, the theory that vision is created via separate ventral and dorsal streams corresponding loosely to "what" and "where." Neither model is perfectly accurate—in fact, every scientific model fails[4] under certain conditions. Models are helpful because they capture important intuitions and can lead to specific predictions and inferences. The same is true of statistical models. On the other hand, statistical models are often driven primarily by raw empiricism—they are produced to fit data and may have little or no other justification or explanatory power. Thus, experienced data analysts carry with them a strong sense of both the

[4] For a discussion of some ways that great equations of physics remain fundamental while only approximating the real world, see Weinberg (2002). An entry into the philosophical literature on statistical inference and modeling is Mayo and Spanos (2010).

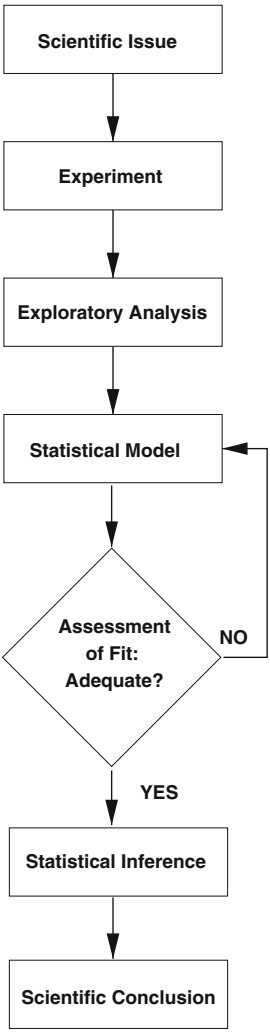

Fig. 1.7 Formal statistical inference within the process of drawing scientific conclusions. Statistical model building is a prerequisite to formal inference procedures. Model building is iterative in the sense that tentative models must be assessed and, if necessary, improved or abandoned. The figure is something of a caricature because the process is not as neat as depicted here. Furthermore, there are typically multiple aspects of the data, which bear on several different issues. A single scientific conclusion may rely on many distinct statistical inferences.

inaccuracies in statistical models and their lingering utility. This sentiment is captured well by the famous quote from George Box, "All models are wrong, but some are useful" (Box 1979).

To emphasize further the status of statistical models we have created Fig. 1.8. Pictured in the left column is the "real world" of data, i.e., the observables, obtained by recording in some form, often by measurement. In the right column is the

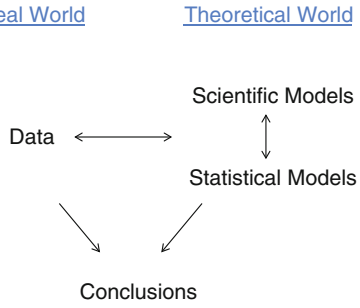

Fig. 1.8 The role of statistical models and methods in scientific inference. Statistical procedures are abstractly defined in terms of mathematics, but are used, in conjunction with scientific models and methods, to explain observable phenomena. Adapted from Kass (2011).

"theoretical world" where both scientific and statistical models live. Scientific models help us organize facts with explanations. They can be high-level or detailed, but they should not, at least in principle, be confused with the observations themselves. The theoretical world seeks to make statements and predictions, often using a precise but abstract mathematical framework, which may be applied to things in the real world that may be observed. In a domain where theory works well, the theoretical world would be judged to be very close to the real world and, therefore, its predictions would be highly trustworthy. Statistical models are used to describe the imperfect predictability of phenomena, the regularity and variability of data, in terms of probability distributions.

A second aspect of the flow diagram in Fig. 1.7 may be misleading. The diagram fails to highlight the way the judgment of adequate fit depends on context. When we say "All models are wrong, but some are useful," part of the point is that a model can be useful *for a specified inferential purpose*. Thus, in judging adequacy of a model, one must ask, "How might the reasonably likely departures from this model affect scientific conclusions?"

We illustrate the way statistical models lead to scientific conclusions in numerous examples throughout this book.

1.2.6 Statistical theory is used to understand the behavior of statistical procedures under various probabilistic assumptions.

The second of the two major components of the statistical paradigm is that methods may be *analyzed* to determine how well they are likely to perform. As we describe briefly in Sections 4.3.4 and 7.3.9, and more fully in Chapters 8 and 11, a series of general principles and criteria are widely used for this purpose. Statistical theory has been able to establish good performance of particular methods under certain probabilistic assumptions. In Chapters 3–6 we provide the necessary background for

the theory we develop. When we wish to add arguments that are not essential to the flow of material we highlight them as *details* and indent them, as follows.

> *Details*: We indent, like this, the paragraphs containing mathematical details we feel may be safely skipped. □

One easy and useful method of checking the effectiveness of a procedure, which is applicable in certain predictive settings, is *cross-validation*. The simplest form of cross-validation involves splitting the data set into two subsets, applying and refining a method using one of the subsets, and then judging its predictive performance (predicting the value of some response) on the second subset. Sometimes the second subset involves entirely new data. For example, in a behavioral study, a new set of subjects may be recruited and examined. Methods that perform well with this kind of cross-validation are often quite compelling. In addition to being intuitive, cross-validation has a theoretical justification discussed briefly in Chapter 12.

1.2.7 Important data analytic ideas are sometimes implemented in many different ways.

The usual starting point in books about data analysis is measures of central tendency, such as mean and median, which we review in Section 2.1.1. There are three reasons for putting a discussion of central tendency at the beginning. First, the use of a single representative value (such as the mean) to summarize a bunch of numbers is ubiquitous. Second, it is an excellent example of the process of data summary; data analysis as a whole may be considered a kind of generalization of this simple method. Third, the mean and median are both single-number summaries but they behave very differently. This last point, that it matters how a general data analytic idea (a single-number summary of central tendency) is defined (mean or median), has become ingrained into teaching about statistical reasoning. The crucial observation[5] is that it can be important to separate the general idea from any specific implementation; as a useful concept, the general idea may transcend any specific definition. For example, in Section 4.3.2 we discuss the deep notion that information represents reduction of uncertainty. As we explain there, the general idea of information could be defined, technically, in terms of a quantity called *mutual information*, but it could also be defined using the squared correlation. Mutual information and squared correlation have very different properties. The definition matters, but with either definition we can think of information as producing a reduction of uncertainty.

1.2.8 Measuring devices often pre-process the data.

Measurements of neural signals are often degraded by noise. A variety of techniques are used to reduce the noise and increase the relative strength of the signal, some

[5] This point was emphasized by Mosteller and Tukey (1977, Section 1F).

of which will be discussed in Chapter 7. In many cases, methods such as these are applied by the measurement software to produce the data the investigator will analyze. For example, fMRI data are acquired in terms of frequency and software is used to reconstruct a signal in time; MEG sensors must be adjusted to ensure detection above background noise; and extracellular electrode signals are thresholded and filtered to isolate action potentials, which then must be "sorted" to identify those from particular neurons. In each of these cases the data that are to be analyzed are not in the rawest form possible. Such pre-processing may be extremely useful, but its effects are not necessarily benign. Inaccurate spike sorting, for example, is a notorious source of problems in some contexts. (See Bar-Gad et al. (2001) and Wood et al. (2004).) The wise analyst will be aware of possible distortions that might arise before the data have been examined.

1.2.9 Data analytic techniques are rarely able to compensate for deficiencies in data collection.

A common misconception is that flaws in experimental design, or data collection, can be fixed by statistical methods after the fact. It is true that an alternative data analytic technique may be able to be help avoid some presumed difficulty an analyst may face in trying to apply a particular method—especially when associated with a particular piece of software. But when a measured variable does not properly capture the phenomenon it is supposed to be measuring, post hoc manipulation will be almost never be able to rectify the situation; in the rare cases that it can, much effort and very strong assumptions will typically be required. For example, we already mentioned that inaccurate spike sorting can create severe problems. When these problems arise, no post-hoc statistical manipulation is likely to fix them.

1.2.10 Simple methods are essential.

Another basic point concerning analytical methods is that simple, easily-understood data summaries, particularly visual summaries such as the PSTH, are essential components of analysis. These fit into the diagram of Fig. 1.7 mainly under the heading of exploratory data analysis, though sometimes inferential analyses from simple models are also used in conjunction with those from much more elaborate models. When a complicated data-analytic procedure is applied, it is important to understand the way results agree, or disagree, with those obtained from simpler methods.

1.2.11 It is convenient to classify data into several broad types.

When spike train data, like those in Example 1.1, are summarized by spike counts occurring in particular time intervals, the values taken by the counts are necessarily

non-negative integers. Because the integers are separated from each other, such data are called *discrete*. On the other hand, many recordings, such as MEG signals, or EEGs, can take on essentially all possible values within some range—subject only to the accuracy of the recording instrument. These data are called *continuous*. This is a very important distinction because specialized analytical methods have been developed to work with each kind of data.

Count data form an important subclass within the general category of discrete data. Within count data, a further special case occurs when the only possible counts are 0 or 1. These are *binary* data. The key characterization is that there are only two possible values; it is a matter of analytical convenience to consider the two values to be 0 or 1. As an example, the response of patient P.S. on each trial was binary. By taking the response "non-burning house" to be 1 and "burning house" to be zero, we are able to add up all the coded values (the 1 and 0 s) to get the total number of times P.S. chose the non-burning house. This summation process is easy to deal with mathematically. A set of binary data would almost always be assumed to consist of 0 and 1 s.

Two other kinds of data arising in neuroscience deserve special mention here. They are called *time series* and *point processes*. Both involve sequential observations made across time. MEG signals, EEGs, and LFPs are good examples of time series: at each of many successive points in time, a measurement is recorded. Spike trains are good examples of point processes: neuronal action potentials are recorded as sequences of event times. In each case, the crucial fact is that an observation at time t_1 is related to an observation made at time t_2 whenever t_1 and t_2 are close to each other. Because of this temporal relationship time series and point process data must be analyzed with specialized methods. Statistical methods for analyzing time series and point processes are discussed in Chapters 18 and 19.

Chapter 2
Exploring Data

Data analysis involves both manipulation, via formulas and computations, and interpretation of the results. During the period immediately following World War II, particularly in the United States, statistical theory was consumed with the logic of statistical inference and decision-making. Against this backdrop, John Tukey revolted. In addition to coining the term *data analysis* (see Brillinger 2002, Appendix D) Tukey emphasized the distinction between formal methods, based on the logic of statistical inference, and informal manipulations—which he called *exploratory*, having a role we indicated in Section 1.2.4. The informality of exploratory data analysis (EDA), however, should not be confused with mathematical simplicity. As we indicate in Section 2.1.2, the manipulations behind many EDA methods are complicated. Tukey's large and lingering influence came from demonstrating the power of mathematical, computational, and statistical insight in producing useful displays and summaries of data. We describe a few basic ideas below.

2.1 Describing Central Tendency and Variation

2.1.1 Alternative displays and summaries provide different views of the data.

Different displays and summaries may emphasize different aspects of the data. While certain data summaries may be well suited for particular purposes, there is never a uniquely "right" way to collapse the data. A multiplicity of possible data features is inherent to the data analytic process. Furthermore, the details of data summary can be important. A simple example is that the mean, i.e., the arithmetic average, of the numbers 2, 3, 10 is 5 while the median is 3. Similarly, a histogram displays the distribution of data values, but the way it does so depends on the way its bins are defined. This is illustrated in the next example.

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_2,

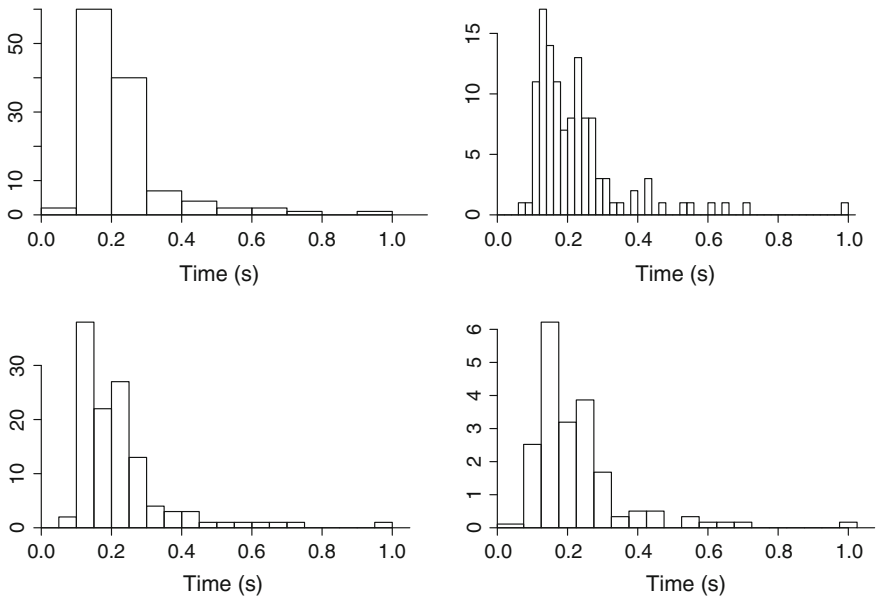

Fig. 2.1 Four histograms of saccadic reaction time data. The same data are used in each histogram. The appearance of the data distribution depends on details of histogram creation. The first three histograms have different bin sizes. The fourth histogram (*bottom right*) uses the same bin size as the third (*bottom left*) but shifts the bin locations slightly.

Example 2.1 Saccadic reaction time in hemispatial neglect Let us consider saccadic reaction times from a single patient in the study of hemispatial neglect by Behrmann et al. (2002). Each measured value is the time (in seconds) to complete an eye saccade. The data have been aggregated across several conditions for pedagogical purposes. There are 119 reaction times, which range from 72 to 988 ms (.072–.988 s (seconds)). The lower quartile (below which lie 25 % of the data) is 140 ms, the median (below which lie 50 % of the data) is 188 ms, and the upper quartile (below which lie 75 % of the data) is 252 ms. Thus, the fast reaction times (72–140 ms) are bunched relatively close to the middle reaction of 188 ms, while the slow reaction times (252–988 ms) are spread out and include some comparatively large values. We refer to this feature of the distribution as *skewness* and say the data are *skewed* toward high values.

Four histograms of the data are shown in Fig. 2.1. Although the same 119 values are used in each, the four histograms give different impressions of the data. In particular, the first histogram (top left) makes the distribution look *unimodal*, i.e., it looks like it has a single peak, while the second (top right) makes the distribution look *bimodal* (two peaks) or even *multimodal* (multiple peaks). However, all four give the clear impression of skewness toward high values. □

In discussing histograms it is important to distinguish this informal use of "distribution" from the mathematical use when we speak about a *probability distribution*.

We will, beginning in Chapter 3, use probability distributions to describe data, but that should be recognized as a conceptual leap: data are observed, and part of the "real world" of Fig. 1.8, while probability distributions are part of the "theoretical world." The word "distribution" is used in both contexts, and we typically hope that a particular probability distribution will do a good job describing a data distribution. As an example, sometimes data distributions—as represented by histograms—are unimodal and more-or-less symmetrical about the median, i.e., the relative frequency of data higher than the median is about the same as that of corresponding data lower than the median by an equal amount. Symmetric and unimodal data distributions are easier to describe concisely with probability distributions and the *normal distribution*, discussed in Chapter 5, is[1] unimodal and symmetrical (it is often called "the bell-shaped curve"). It is very rare to find a set of data that, on close inspection, may be described accurately by a normal distribution, but it is common to find unimodal and symmetric data distributions that are roughly normal-looking. A great deal of emphasis is placed on the normal distribution, in large part because of its appearance as a basic assumption of many formal statistical procedures and because such statistical procedures typically remain useful for modest departures from normality, due to[2] the Central Limit Theorem (Section 6.3.1). When departures from normality become large, however, they can materially affect the behavior of the procedures. A standard practice, therefore, is to examine data via displays such as histograms, looking especially for substantial skewness.

In talking about a single set of numbers, such as the saccadic reaction times, it is useful to have a word for the complete set of values. We will follow Tukey by using the word *batch*, i.e., we will speak of the batch of 119 saccadic reaction times.

The saccadic reaction time data are substantially skewed. One effect of this is that the mean (the arithmetic average) is substantially higher than the median: the mean reaction time is 226 ms, while the median is 188 ms. This is because the mean is affected much more strongly by values that are far away from the middle of the distribution. Data values that are very far from the middle of the distribution are called *outliers*, and the sensitivity of the mean to outliers is one reason it is often replaced by the median as a summary of central tendency, that is, a single number that represents a center among all the values.

In addition to the mean and median, the *mode*, which is the value occurring most frequently, is sometimes mentioned in this context. However, the term "mode" is not used in a precise way very often when describing a batch of numbers. The concept of a mode applies better to the theoretical setting of probability densities, where it is the value at which the density is maximized. For a batch of numbers we typically speak, instead, informally and approximately, of "the mode" as being the rough location of the peak of the distribution.

---

[1] The normal distribution is also called the *Gaussian distribution*.

[2] As we point out in Chapters 7–10, statistical procedures often require statistical summaries (such as the least-squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ on p. 12) to be normally distributed rather than the data themselves.

Just as central tendency in data may be summarized by mean or median, variability may be summarized by more than one measure. We might ask, for example, how much the saccade times vary. For instance, if we were to look at a control subject might we expect less variability? How should we quantify this?

The most widely used summary of variability is the *standard deviation*:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where $x_1, x_2, \ldots, x_n$ are the observations and $\bar{x}$ is their mean. We may think of s as an "average deviation" of the values from their mean. The square-root is used so that the units of standard deviation agree with the units of the data themselves. For the saccadic reaction time data we find the standard deviation to be $s = 134$ milliseconds ($s = .134$ s). The use of $n - 1$ rather than n in the formula for s comes from certain theoretical arguments given in Chapter 8.

An alternative to the standard deviation would be the mean absolute deviation $\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$ but this turns out to be mathematically less convenient. In some contexts the median absolute deviation is used as this is not affected by outliers. If $\tilde{x} = median(x_1, x_2, \ldots, x_n)$ is the median of the n data values x_i then the median absolute deviation is $median(|x_1 - \tilde{x}|, |x_2 - \tilde{x}|, \ldots, |x_n - \tilde{x}|)$. Sometimes the difference between the upper and lower quartiles (see p. 24) is used. This is called the *interquartile range*.

In this section we have reviewed several very basic methods of data summary and display while trying to illustrate the general notion that alternative measures and displays can produce differing impressions of the data. An additional concern is that perception of data may depend on aspects of the way the data are displayed that have nothing to do with choices of data features. For scatterplots of a variable y against another variable x, Cleveland et al. (1982) showed that a subject's perception of association depends on the size of the scatterplot within the frame created by the axes. In choosing data displays it is worth keeping such perceptual issues in mind.

2.1.2 Exploratory methods can be sophisticated.

As we said in Section 1.2.4, exploratory data analysis (EDA) refers to the collection of methods that are relatively informal, based not on a cohesive logical framework built around statistical models but rather on tools that help illuminate interesting features of the data. The informal methods of EDA can be extremely useful. In this section we have mentioned a couple of very elementary descriptive methods, but in some cases informal techniques can draw on quite sophisticated ideas. The next example involves a method we will discuss in Chapter 18.

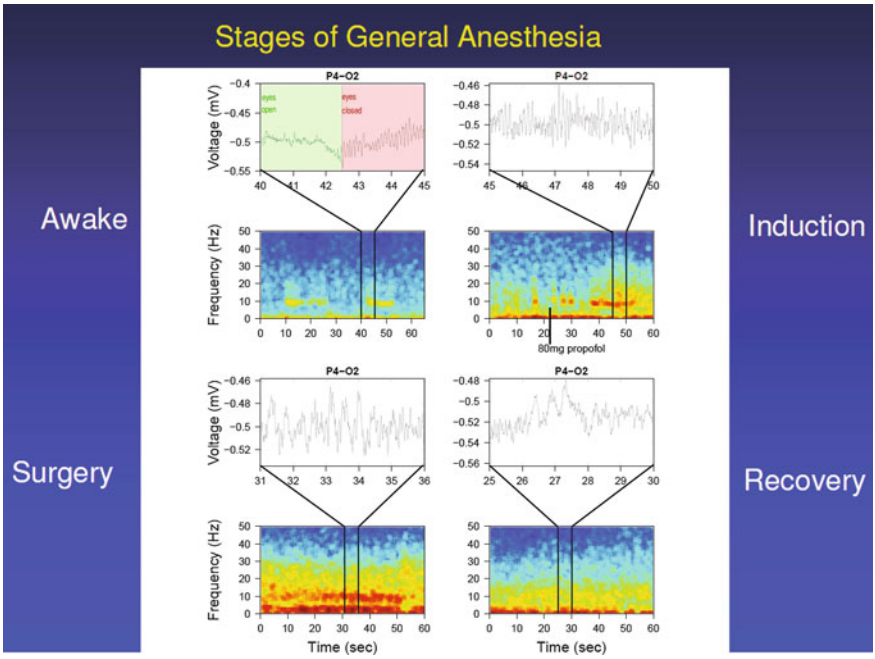

Fig. 2.2 EEG spectrograms for a subject in various stages of general anesthesia. In each of four stages an EEG voltage tracing is shown, and below it a spectrogram. The EEG tracings are for the P4 (right parietal) lead in an array of 16 leads (it is taken with O2 as reference lead). The spectrogram decomposes the voltage signal into frequency components across successive time bins. *Red* indicates high magnitudes, *yellow* medium magnitudes, and *blue* low magnitudes. Each displayed trace corresponds to several successive time bins in the spectrogram, as indicated by the *black lines*. Two prominent features are the alpha rhythm, at roughly 10 Hz, and the slower delta rhythm, below 4 Hz. Both sets of oscillations are visible in the EEG tracings, and their temporal presence or absence is indicated in the spectrogram. During the awake phase the alpha rhythm is absent when the eyes are open and present when the eyes are closed; the delta rhythm is also present, but only weakly. During surgery the delta rhythm is very strong, and the alpha rhythm is also stronger than in the awake phase.

Example 2.2 EEG spectrogram under general anesthesia When patients undergo general anesthesia for certain surgical procedures EEGs are recorded to monitor brain activity. These recordings provide a comparison of various states of unconsciousness. A set of EEG traces for a patient during carotid endarterectomy surgery at the Massachusetts General Hospital is displayed in Fig. 2.2. The figure shows EEGs and spectrograms during an initial awake phase, a general anesthesia induction phase, the surgical phase, and the recovery phase. Spectrograms are made by taking the signal within successive time bins (here, 1 s bins) and using *Fourier analysis* to decompose the signal into oscillatory components at varying frequencies. On the x-axis is time and on the y-axis is the frequency. The plotted spectrogram is the resulting power (a measure of the strength of a particular frequency component of the

signal) at each frequency, for each time bin, indicated in the figure by three different colors representing low, medium, and high power. In Fig. 2.2 the most easily visible oscillations are the alpha rhythm (roughly 8–13 Hz) in the second half of the EEG trace in the awake phase (when the eyes are closed) and the delta rhythm (below 4 Hz) during the surgical phase. Precise scientific statements often require statistical inferences (indications of uncertainty or tests of hypotheses), but spectrograms are very useful in displaying time-frequency information even without formal inferential assessments. □

2.2 Data Transformations

2.2.1 Positive values are often transformed by logarithms.

Measurement scales arise from convenience, and need not be considered in any way absolute or immutable; changing the scale often produces a more elegant description. A canonical example involves the acidity of a dilute aqueous solution, which is determined by the concentration of hydrogen ions. The larger the concentration $[H^+]$ of hydrogen ions, the more acidity. Rather than using $[H^+]$ to measure acidity, we use its logarithm, which is known as pH. Specifically, $pH = -\log_{10}([H^+])$, so that an increase in $[H^+]$ corresponds to a decrease in pH. Because the defining property of the logarithm is

$$\log ab = \log a + \log b, \tag{2.1}$$

log transformations are used when multiplicative effects seem more natural than additive. In the case of pH, a solution having a hydrogen ion concentration of 10^{-5} mol l^{-1} is 1 unit greater pH (less acidic) than a solution having a concentration of 10^{-4} mol l^{-1}. Similarly, a solution having a hydrogen ion concentration of 10^{-9} mol l^{-1} is 1 unit greater pH than a solution having a concentration of 10^{-8} mol l^{-1}. In both cases, a 1 unit increase in pH corresponds to a 10-fold decrease in hydrogen ion concentration, regardless of the concentration we started with. In chemical calculations, the log concentration scale is simpler to work with than the concentration scale.

Many other familiar scales are logarithmic. One example is the use of decibels to measure the strength of an auditory signal. Not only are log scales familiar and intuitive but, in addition, some batches of data look more nearly like observations from a normal distribution following a log transformation. In particular, it frequently happens that a batch of data look highly skewed in a given measurement scale, but are much closer to being symmetric in the log scale.

Example 2.1 (continued from p. 24) Figure 2.3 displays the saccadic reaction time data in both the original scale and the log transformed scale. To transform the data to the log scale we have replaced x = reaction time by $\log_{10}(x)$ for each of the 119

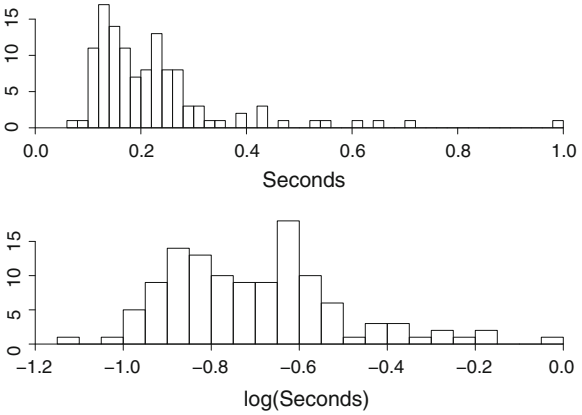

Fig. 2.3 Histograms of eye saccade data. *Top display* is for data in the original scale, *bottom display* is for the same data after being transformed by $\log_{10}$. The data are distributed more symmetrically in the log scale.

values. In the log scale the distribution is more symmetric. In addition, the potential bimodality, or possibly even multimodality of the distribution is also evident in the log scale. The data shown here were aggregated by combining conditions in which the eyes began fixating centrally, to the right, or to the left, which may explain why the histogram does not appear to be unimodal. When the data are disaggregated into single conditions, in the log scale they do appear unimodal and roughly symmetrical. For this reason, Behrmann et al. (2002) chose to perform many of their analyses in the log scale. □

Example 2.3 High-Field BOLD signal Lewis et al. (2005) have argued that for some purposes it may be advantageous to transform the BOLD signal in fMRI data by taking logarithms, at least in the case of high-field signals. Those authors examined the BOLD intensity for subjects during 4 T imaging, with a simple visual stimulus. Figure 2.4 displays a histogram (with dots replacing bin heights) of the BOLD values collected from 19,000 voxels for each of 15 subjects and 15 images under their control condition, during which the subjects were fixating on a central spot on the screen they were watching. It is apparent that this distribution across voxels is quite skewed. The authors produced various plots aimed at suggesting the log transformation could be useful. □

The way we usually think of the log transformation is that it produces a more "natural" scale for measurements whenever they are necessarily positive and might reasonably be compared in proportional relationships. We have already mentioned that normal distributions for data are assumed by standard statistical procedures, that data distributions are rarely very close to normal, but that mild departures from normality are generally tolerable. Such mild departures are common: once we transform the data to a suitable scale, distributions are often unimodal and more-or-less

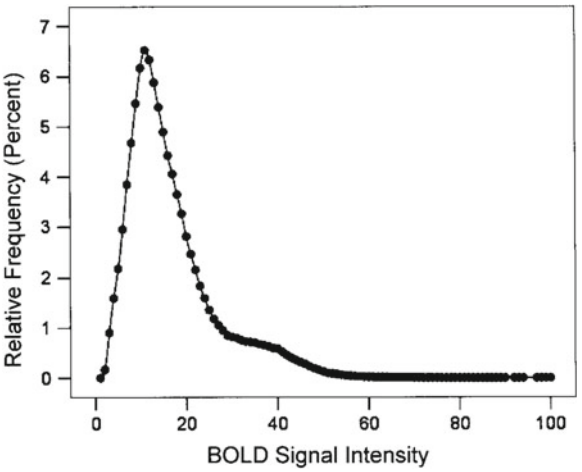

Fig. 2.4 High field of BOLD signal intensities. The frequencies are plotted as *dots*, rather than bin heights. The distribution across voxels is skewed toward high values. Reprinted with permission from Lewis et al. (2005).

symmetrical. Presumably, this has to do with effects of the Central Limit Theorem. We will discuss this great theorem in Chapter 6. For now let us be content to state it this way: if we add up many small, independent effects their sum will be approximately normally distributed. The empirical observation of approximate normality may then be interpreted as follows: *if we choose the right scale*, the data values may be considered sums of many small, independent effects.

We can understand this a little more deeply by returning to the logarithmic relationship in Eq. (2.1), and considering the role it may play when many small effects are combined to produce variability. The cases where the log transformation is valuable are those where it is natural to think in terms of proportionality. So suppose the reason that two measurements are different is that many small *proportional* effects, of somewhat different sizes in the two measurements, have been combined. For example, the length of a dendritic spine may depend on contributions to the cell membrane and its contents by vast numbers of lipid and protein molecules. If we break the growth process into many thousands of pieces, each might be considered a small effect, so that the net result is a composition of many, many small effects. When we see that one spine is longer than another, we might imagine that the many small effects in the longer spine tended to be *proportionally* larger than those in the shorter spine. Now consider two such small growth effects x_1 and x_2, occurring, respectively, in the shorter and longer dendrites. If we think of the variation as proportional, we may relate the values x_1 and x_2 by writing $x_2 = x_1(1+\epsilon)$, where ϵ is a small number representing the proportional change (e.g., $\epsilon = .05$, or 5 %) in going from x_1 to x_2. From Eq. (2.1) together with a little calculus, for small ϵ we have $\log(1+\epsilon) \approx \epsilon$ (see Section A.4 of the Appendix). We then have

$$\begin{aligned}\log x_2 - \log x_1 &= \log(1+\epsilon)\\ &\approx \epsilon.\end{aligned}$$

In other words, when we add a small perturbation ϵ to $\log x_1$ we get $\log x_2$. Thus, if we wish to think of ϵ as a small random quantity that creates variability in the data multiplicatively, it does so additively on the log scale. If the length of a dendritic spine is the result of thousands of small growth processes that act multiplicatively, the *log* of the length will be the sum of many thousands of ϵs and, therefore, according to the Central Limit Theorem, will be described, approximately, by a normal distribution. The same would be true of other measurements that are necessarily positive and might reasonably be expected to follow proportional variation. From this argument, one would expect that a batch of such numbers might look more symmetric and unimodal following a log transformation.

All of this is heuristic; there is no argument here that can be claimed formally correct—the Central Limit Theorem applies not to data but to mathematical quantities that live in the "theoretical world" described in Section 1.2.5. We are simply trying to provide a plausible explanation for the empirical fact that log-transformed growth measurements usually have fairly symmetric distributions.

In transforming data by logs it does not matter what base is used. In mathematics it is common to use[3] the "natural" log (base e). In applications, where answers are expressed and interpreted in decimal expansions, the "common" log (base 10) is often used. (Sometimes binary expansions are most intuitive and $\log_2(x)$ is used.) These transformations have a simple relationship:

$$\log_e(x) = \log_e(10)\log_{10}(x).$$

This implies a batch of numbers transformed by $\log_e$ will look essentially the same as the batch transformed by $\log_{10}$. The only distinction is multiplication by the constant $\log_e(10)$ applied to each value. Thus, for data analytic purposes it does not matter which scale is used. Of course, to interpret the results in a meaningful way, based on relevant physiological units, one must know which logarithmic base has been applied. The statistics literature follows the mathematics convention in using $\log_e$ unless otherwise noted. We follow this convention here.

Another motivation for logarithmic transformations is that they convert power laws, which are useful in describing many neuroscientific phenomena, to simpler linear forms. Power laws have the form

$$w = cv^b \tag{2.2}$$

and may be summarized by saying that a proportional change in v produces a proportional change in w. If we let $y = \log w$ and $x = \log v$ then

[3] Of all the values A in the function $f(x) = A^x$, the value $A = e$ makes the derivative of $f(x)$ exactly equal to $f(x)$ itself. For other values of A a constant must be introduced, which would make calculus-based formulas more complicated.

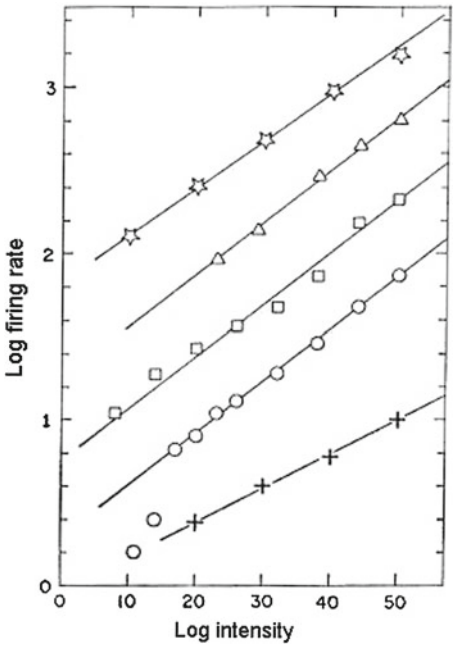

Fig. 2.5 Power function fits to firing-rate data, shown on log–log scale. On the y-axis are log firing rates, and on the x-axis is log intensity of light. The data are from three different sources, using three distinct methods of collection. Except for the deviation from the line at low intensities for the data set indicated by circles, the fits are quite good. Adapted from Stevens (1970).

$$y = a + bx,$$

where $a = \log c$.

Example 2.4 Stimulus-response power laws Power laws may be used to describe the way increases in stimulus intensity produce increased magnitudes of sensation Stevens (1961) (where they replace the "Weber–Fechner" law $w = a + d \log v$), or increased neural firing rate Stevens (1970). For example, Fig. 2.5 displays five classic sets of data on neural responses from the eye of the horseshoe crab *Limulus*. For each data set, the log of neural firing rate is plotted as a function of log of light intensity. In each case the function is approximately linear. In other words, in each case the relationship of firing rate to stimulus intensity follows, approximately, a power law. □

Example 2.5 Power law for skill acquisition Power laws also arise in describing the effects of practice on recall or reaction time in memory and skill acquisition (Anderson 1990). An interesting set of data comes from Kolers (1976) who investigated the learned skill of reading inverted text.[4] As shown in Fig. 2.6, he found

[4] See also the related work on power laws by Anderson and Schooler (1991).

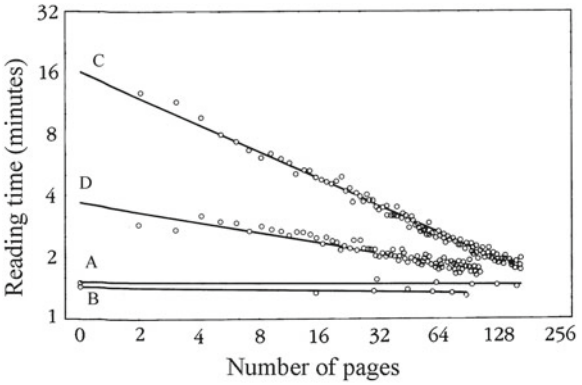

Fig. 2.6 Skill learning described by a power law, shown on a log–log scale. On the y-axis is the log (base 2) of the time taken to read a passage of inverted text (in minutes), and on the x-axis is log practice time (in pages). Four sets of data from multiple subjects are displayed. Data were obtained on two occasions, separated by a year, on both ordinary text and inverted text (creating a total of four conditions). *Line A* is fit to data based on ordinary text on the first occasion and *line B* is fit to data based on ordinary text on the second occasion. There is essentially no training effect. *Lines C* and *D* are the fits for inverted text. In both cases there is a clear power-law relationship, indicated by the good fit of the lines to the data. Substantively, after the delay by a year the subjects again improved with practice, but they had retained much of the skill of reading inverted text (*line D* is below *line C*) and needed only about 100 pages of training to reach the proficiency previously obtained after 200 pages. Adapted from Kolers (1976).

two things. First, a decreasing power law describes the relationship of reading time to amount of practice. Second, when subjects were tested a year later, they had lost some of their ability to read the inverted text, and then regained it again according to a power law, though at a slower rate. The two relationships are shown in Fig. 2.6 as a pair of lines with differing slopes and intercepts. These studies are of great interest for education: they suggest that retained learning may be quantified by the decrease in training time required to achieve proficiency following re-training, compared to the original training time. □

2.2.2 Non-logarithmic transformations are sometimes applied.

The log is by far the most common transformation, but there are others, too. The general method of transformations is to replace a measured variable x, such as reaction time, with some $f(x)$ for every value of x. For example, reaction times and other time measurements are sometimes analyzed on the reciprocal scale $1/x$: the reciprocal transforms time to something proportional to speed (speed is distance/time). Square-root transformations are also sometimes used, especially for spike counts because the square-root can be a so-called *variance-stabilizing transformation*, as discussed in Chapter 9. Square-roots are also sometimes used for measurements of area and

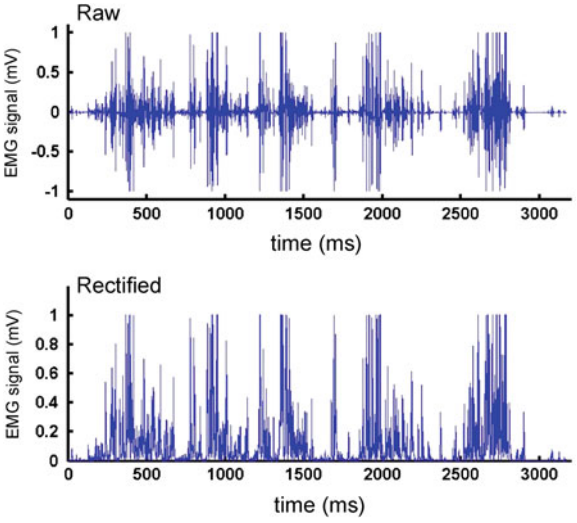

Fig. 2.7 EMG from the leg of a frog during a swimming motion. *Top panel* shows raw signal. *Bottom panel* shows the rectified signal.

cube-root transformations are occasionally used for volumetric measurements. We may order these transformations by letting, for the moment, the symbol $<$ stand for "less strong than" and then writing them as follows:

$$x < x^{1/2} < x^{1/3} < \log(x) < 1/x.$$

In each case we strengthen the transformation (make it pull in the right-hand tail more) as we decrease the power to which we raise x. Note that $1/x = x^{-1}$ and that, in this context, the log corresponds to using the power 0, so that increasing the strength of transformation corresponds to decreasing the exponent.

> *Details:* We may imbed the log in the power family of transformations by putting the power transformations in the normalized form
>
> $$f(x) = (x^{\alpha} - 1)/\alpha.$$
>
> By calculus (L'Hôpital's rule) it then follows that the log corresponds to $\alpha = 0$. □

In general, both distributional symmetry and interpretability are important in determining a scale for analysis.

These "power transformations" are all monotonic. Occasionally, non-monotonic transformations are used, as in the analysis of EMG recordings.

Example 2.6 EMG in frog movement An electromyogram (EMG) is a recording of the electrical impulses transmitted through a group of muscle fibers, recorded as electrical potentials. Because the instantaneous potential is generated from both

agonist and antagonist muscle fibers, it is recorded as both positive and negative. This is shown in the top panel of Fig. 2.7, which is a display of an EMG taken from a frog during a leg extension. Because the force generated by a muscle is only positive, the standard convention is to analyze the full-wave rectified signal, i.e., the absolute value of the potential. This is shown in the bottom panel of Fig. 2.7. □

Chapter 3
Probability and Random Variables

Probability is a rich and beautiful subject, a discipline unto itself. Its origins were concerned primarily with games of chance, and many lectures on elementary probability theory still contain references to dice, playing cards, and coin flips. These lottery-style scenarios remain useful because they are evocative and easy to understand. On the other hand, they give an extremely narrow and restrictive view of what probability is about: lotteries are based on elementary outcomes that are equally likely, but in many situations where quantification of uncertainty is helpful there is no compelling way to decompose outcomes into equally-likely components. In fact, the focus on equally-likely events is characteristic of pre-statistical thinking.[1] The great advance toward a more general notion of probability was slow, requiring over 200 years for full development.[2] This long, difficult transition involved a deep conceptual shift. In modern texts equally-likely outcomes are used to illustrate elementary ideas, but they are relegated to special cases. It is sometimes possible to compute the probability of an event by counting the outcomes within that event, and dividing by the total number of outcomes. For example, the probability of rolling an even number with a fair six-sided die, i.e., of rolling any of the three numbers 2, 4, or 6, out of the 6 possibilities, is $\frac{3}{6} = \frac{1}{2}$. In many situations, however, such reasoning is at best a loose analogy. To quantify uncertainty via statistical models a more general and abstract notion of probability must be introduced.

This chapter begins with the axioms and elementary laws of probability, and then discusses the way probability is used to describe variability. The key concept of *independence* is defined in Section 3.1.3. Quantities that are measured but uncertain are formalized in probability theory as *random variables*. More specifically, we set up a theoretical framework for understanding variation based on probability distributions of random variables, and the variation of random variables is supposed to be similar

[1] See Stigler (1986).

[2] Its beginning point is usually traced to a text by Jacob Bernoulli, posthumously-published in 1713 (Bernoullli 1713), and its modern endpoint was reached in 1933, with the publication of a text by Kolmogorov (1933).

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_3,

to real-world variation observed in data. Many families of probability distributions are used throughout the book. The most common ones are discussed in Chapter 5.

One quick note on terminology: the word *stochastic* connotes variation describable by probability. Within statistical theory it is often used in specialized contexts, but it is almost always simply a synonym for "probabilistic."

3.1 The Calculus of Probability

3.1.1 Probabilities are defined on sets of uncertain events.

The calculus of probability is defined for *sets*, which in this context are called *events*. That is, we speak of "the probability of the event A" and we will write this as $P(A)$. Events are considered to be composed of *outcomes* from some experiment or observational process. The collection of all possible outcomes (and, therefore, the union of all possible events) is called the *sample space* and will be denoted by Ω. Because Ω is a set, we also say that Ω is made up of elements (each of which is an outcome) and to indicate that ω is an element of Ω we write $\omega \in \Omega$. Recall the definitions of *union* and *intersection*: for events A and B the union $A \cup B$ consists of all outcomes that are either in A or in B or in both A and B; the intersection $A \cap B$ consists of all outcomes that are in both A and B. The *complement* A^c of A consists of all outcomes that are *not* in A. We say two events are *mutually exclusive* or *disjoint* if they have empty intersection.

Example 3.1 Two neurons from primary visual cortex In an experiment on response properties of cells in primary visual cortex, Dr. Ryan Kelly and colleagues recorded approximately 100 neurons simultaneously from an anesthetized macaque monkey while the animal's visual system was stimulated by highly irregular random visual input (Kelly et al. 2007). The stimulus they used is known as *white noise*, which will be defined in Chapter 18. Kelly examined the response of two neurons during 100 ms of the stimulus. Let A be the event that the first neuron fires at least once within the 100 ms time interval and B the event that the second neuron fires at least once during the same time interval. Here, $A \cup B$ is the event that at least one of the 2 neurons fires at least once, while $A \cap B$ is the event that both neurons fire at least once. Because it is possible that both neurons will fire during the time interval, the events A and B are *not* mutually exclusive. □

We now state the axioms of probability.

Axioms of Probability:

1. For all events A, $P(A) \geq 0$.
2. $P(\Omega) = 1$.
3. If $A_1, A_2, \ldots, A_n$ are mutually exclusive events, then $P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$.

If we let $\cup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \cdots \cup A_n$ then Axiom 3 may be written instead in the form

3. If $A_1, A_2, \ldots,$ are mutually exclusive events, then $P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$.

A technical point is that in advanced texts, Axiom 3 would instead involve infinitely many events, and an infinite sum:
$3'$. If $A_1, A_2, \ldots,$ are infinitely many mutually exclusive events, then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Regardless of whether one worries about the possibility of infinitely many events, it is easy to deduce from the axioms the elementary properties we need.

Theorem: Three Properties of Probability For any events A and B we have

(i) $P(A^c) = 1 - P(A)$, where A^c is the complement of A.
(ii) If A and B are mutually exclusive, $P(A \cap B) = 0$.
(iii) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof: To prove (i) we simply note that $\Omega = A \cup A^c$. From axiom (2) we then have $P(A \cup A^c) = 1$ and because A and A^c are mutually exclusive axiom (3) gives $P(A) + P(A^c) = 1$, which is the same as (i). It is similarly easy to prove (ii) and (iii). □

To illustrate, suppose we pick at random a playing card from a standard 52-card deck. We may compute the probability of drawing a spade or a face card, meaning either a spade that is not a face card, or a face card that is not a spade, or a face card that is also a spade. We take A to be the event that we draw a spade and B to be the event that we draw a face card. Then, because there are 3 face cards that are spades we have $P(A \cap B) = \frac{3}{52}$, and, applying the last formula above, we get $P(A \cup B) = \frac{1}{4} + \frac{3}{13} - \frac{3}{52} = \frac{11}{26}$. This matches a simple enumeration argument: there are 13 spades and 9 non-spade face cards, for a total of 22 cards that are either a spade or a face card, i.e., $P(A \cup B) = \frac{22}{52} = \frac{11}{26}$. The main virtue of such formulas is that they also apply to contexts where probabilities are determined without reference to a decomposition into equally-likely sub-components.

Example 3.1 (continued from p. 38) From 1,200 replications of the 100 ms stimulus Kelly calculated the probability that the first neuron would fire at least once was $P(A) = .13$ and the probability that the second neuron would fire at least once was $P(B) = .22$, while the probability that both would fire at least once was $P(A \cap B) = .063$. Applying the formula for the union (property (iii) above), the probability that at least one neuron will fire is $P(A \cup B) = .13 + .22 - .063 = .287$. □

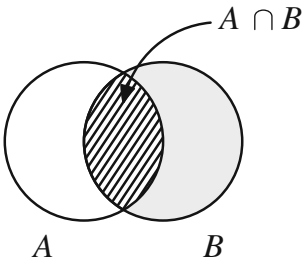

Fig. 3.1 Venn diagram showing the intersection of A and B. The events A and B are depicted as open and filled-in *circles*, respectively, while $A \cap B$, the portion of B that is also in A, is shown with diagonal lines. The conditional probability of A given B is the relative amount of probability assigned to A within the probability assigned to B, i.e., the probability assigned to the region having diagonal lines divided by the probability assigned to the whole of B.

3.1.2 The conditional probability P(A|B) is the probability that A occurs given that B occurs.

We often have to compute probabilities under an assumption that some event has occurred. For instance, one may be interested in the probability that a neuron will fire in an interval of time $(t, t + \Delta t)$ given that it has already fired at a previous time t_0. If we let A be the event we are interested in and B the event that is assumed to have occurred, then we write[3] $P(A|B)$ for the *conditional probability of A given B*. From a Venn diagram (see Fig. 3.1) it is easy to visualize the calculation required: we limit the universe to B and ask for the relative probability assigned to the part of A that is contained in B. Algebraicly, the formula is the following:

Definition: Conditional Probability Assume $P(B > 0)$. The conditional probability of A given B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Again, using draws from a deck of cards, the probability of drawing a Jack given that we draw a face card is $P(A|B) = \frac{4/52}{12/52} = \frac{1}{3}$.

A rewriting of the definition of conditional probability is also sufficiently useful to have a name:

Multiplication rule If $P(B) > 0$ we have $P(A \cap B) = P(A|B) \cdot P(B)$.

Although conditional probability calculations are pretty straightforward, problems involving conditioning can be confusing. The trick to keeping things straight is to be clear about the event to be conditioned upon. Here is one standard example.

[3] This notation is due to Jeffreys (1931); see his p. 15.

Illustration: The boy next door Suppose a family moves in next door to you and you know they have two children, but you do not know whether the children are boys or girls. Let us assume the probability that either particular child is a boy is $\frac{1}{2}$. We might label them Child 1 and Child 2 (e.g., Child 1 could be the older of the two). Thus, $P(\text{Child 1 is a boy}) = P(\text{Child 2 is a boy}) = \frac{1}{2}$. Now suppose you find out that one of the children is a boy. What is the probability that the other child is also a boy?

It may seem that the answer is $\frac{1}{2}$ but, if we assume that "you find out one of the children is a boy" means *at least one of the children is a boy*, then the correct answer is $\frac{1}{3}$. Here is the argument. When you find out that one of the children is a boy you do not know whether Child 1 is a boy, nor whether Child 2 is a boy, but you do know that one of them is a boy—and possibly both are boys. This information amounts to telling you it is impossible that both are girls. Let A be the event that both children are boys and B the event that at least one child is a boy. We want $P(A|B)$. Note that there are four equally-likely possibilities:

$$\begin{aligned} &P(\text{Child 1 is a boy and Child 2 is a boy}) \\ &\quad = P(\text{Child 1 is a boy and Child 2 is a girl}) \\ &\quad = P(\text{Child 1 is a girl and Child 2 is a boy}) \\ &\quad = P(\text{Child 1 is a girl and Child 2 is a girl}). \end{aligned}$$

Thus, we compute $P(A \cap B) = P(A) = \frac{1}{4}$ and $P(B) = \frac{3}{4}$. Plugging these numbers into the formula for conditional probability we get $P(A|B) = \frac{1}{3}$. □

3.1.3 Probabilities multiply when the associated events are independent.

Intuitively, two events are *independent* when the occurrence of one event does not change the probability of the other event. This intuition is captured by conditional probability: the events A and B are independent when knowing that B occurs does not affect the probability of A, i.e., $P(A|B) = P(A)$. This statement of independence is symmetrical: A and B are also independent if $P(B|A) = P(B)$. However, these statements are not usually taken as the definition of independence because they require the events to have nonzero probabilities (otherwise, conditional probability is not defined). Instead, the following is used as a definition.

Definition: Independence Two events A and B are independent if and only if $P(A \cap B) = P(A) \cdot P(B)$.

Note that from this definition, when A and B are independent and $P(B) > 0$ we have, as a consequence,

$$P(A|B) = P(A \cap B)/P(B) = P(A)P(B)/P(B) = P(A).$$

Multiplication of probabilities should be very familiar. If a coin has probability .5 of coming up heads when flipped, then we usually say the probability of getting two heads is $.25 = .5 \times .5$, because we usually assume that the two flips are independent.

Example 3.1 (continued from p. 39) For the probabilities $P(A)$, $P(B)$ given on p. 39 we have $P(A)P(B) = .029$ while the probability of the intersection was reported to be $P(A \cap B) = .063$. The latter is more than double the product $P(A)P(B)$. We conclude that the two neurons are not independent. Their tendency to fire much more often together than they would if they were independent could be due to their being connected, to their having similar response properties, or to their both being driven by network fluctuations (see also Kelly et al. 2010). □

The definition of independence extends immediately to more than two events: if $A_1, A_2, \ldots, A_n$ are independent then

$$P(\cap_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$$

where $\cap_{i=1}^{n} A_i = A_1 \cap A_2 \cap \cdots \cap A_n$.

Independence is extremely useful. Without it, dependencies represented by conditional probabilities can become very complicated. Independence simplifies calculations and is often assumed in statistical models and methods. On the other hand, as illustrated in Example 3.1, above, if the assumption of independence is wrong, the calculations can be way off: in Example 3.1 the probability $P(A \cap B)$ predicted by independence would be too small by a factor of more than 2. In many situations independence is the most consequential statistical assumption, and therefore must be considered carefully.

3.1.4 Bayes' theorem for events gives the conditional probability P(A|B) in terms of the conditional probability P(B|A).

Bayes' theorem is a very simple identity, which we derive easily below. Yet, it has profound consequences. We can state its purpose formally, without regard to its applications: Bayes' theorem allows us to compute $P(A|B)$ from the reverse conditional probability $P(B|A)$, if we also know $P(A)$. As we will see below, and in Chapter 16, there are more complicated versions of the theorem, and it is especially those that produce the wide range of applications. But the power of the result becomes apparent immediately when we take B to be some data and A to be a scientific hypothesis. In this case, we can use the probability $P(\text{data}|\text{hypothesis})$ from the statistical model to obtain the scientific inference $P(\text{hypothesis}|\text{data})$. In the words used in Chapter 1, p. 14, Bayes' theorem provides a vehicle for obtaining epistemic probabilities from

descriptive probabilities (see Section 16.1.1). The inverting of conditional probability statements, together with the recognition that a different notion of probability was involved, led to the name "inverse probability" during the early 1800s. This has been replaced by the name "Bayes" in the theorem, and the adjective "Bayesian" to describe many of its applications.[4] To derive the theorem we need a preliminary result which is also important.

Theorem: Law of Total Probability For events A and B we have

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c).$$

Proof : We begin by decomposing B into two pieces: $B = (B \cap A) \cup (B \cap A^c)$. Because A and A^c are disjoint, $(B \cap A)$ and $(B \cap A^c)$ are disjoint. We then have $P(B) = P(B \cap A) + P(B \cap A^c)$. Applying the multiplication rule to $P(B \cap A)$ and $P(B \cap A^c)$ gives the result. □

Bayes' Theorem in the Simplest Case If $P(B) > 0$ then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}. \tag{3.1}$$

Proof : We begin with the definition of conditional probability and then use the multiplication rule in the numerator and the law of total probability in the denominator:

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}. \end{aligned}$$

□

The "simplest case" modifier here refers to the statement of the theorem in which the law of total probability is applied to the denomoninator $P(B)$ by decomposing B by intersection with only two events, A and A^c. We discuss other versions of the theorem below.

One interesting class of problems where this simple case is useful is in the interpretation of clinical diagnostic screening tests. These tests are used to indicate that a patient may have a particular disease A, based on a test outcome B, but they are not definitive. The probability $P(B|A)$ that a patient having the disease tests positively is known as the *sensitivity* of the test, the probability $P(B^c|A^c)$ that a patient who does *not* have the disease tests negatively is known as the *specificity* of the test, and the probability $P(A)$ that a patient drawn randomly from the population has the disease

[4] For historical comments see Stigler (1986) and Fienberg (2006).

is known as the *prevalence* of the disease. Good diagnostic screening tests have sensitivity and specificity close to 1 but, as we will describe, Bayes' Theorem serves as a quantitative reminder that when a disease is rare, screening tests are preliminary, and other information will be needed to provide a diagnosis. Specifically, if we let $PPV = P(A|B)$, which stands for *positive predictive value*, we get

$$PPV = \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})} \tag{3.2}$$

and, when the prevalence is small, the value of PPV will also typically be small—sometimes surprisingly small.

A famous example involves screening for prostate cancer based on the radioimmunoassay prostatic acid phosphatase (PSA). Even though the test is reasonably accurate, the disease remains sufficiently rare among young men that a random male who tests as positive will still have a low probability of actually having prostate cancer. An application of Bayes' Theorem (with A being the event that a randomly chosen man will have the disease and B the event that he tests positive) to data from Watson and Tang (1980), places the probability of disease given a positive test at about 1/125. The intuition comes from recognizing that, among men under age 65 in the United States, the disease has a prevalence of about 1/1,500. Suppose we were to examine 1,500 men, 1 of whom actually had the disease. If the screening test were 90 % accurate, a 10 % false positive rate would mean that about 150 men would test positively. In other words, about 1/150 of the positively tested men would actually have the disease. Bayes' Theorem refines this very crude calculation. Here is an example drawn from neurology.

Example 3.2 Diagnostic test for vascular dementia Vascular dementia (VD) is the second leading cause of dementia. It is important that it be distinguished from Alzheimer's disease because the prognosis and treatments are different. In order to study the effectiveness of clinical tests for vascular dementia, Gold et al. (1997) examined 113 brains of dementia patients postmortem. One of the clinical tests these authors considered was proposed by the National Institute of Neurological Disorders and Stroke (NINDS, an institute of NIH). Gold et al. found that the proportion of patients with VD who were correctly identified by the NINDS test, its sensitivity, was .58, while the proportion of patients who did not have VD who were correctly so identified by the NINDS test, its specificity, was .80. Using these results, let us consider an elderly patient who is identified as having VD by the NINDS test, and compute the probability that this person will actually have the disease. Let A be the event that the person has the disease and B the event that the NINDS test is positive. We want $P(A|B)$, and we are given $P(B|A) = .58$ and $P(B^c|A^c) = .8$. To apply Bayes' Theorem we need the disease prevalence $P(A)$. Let us take this probability to be $P(A) = .03$ (which seems a reasonable value based on Hébert and Brayne 1995). We then also have $P(A^c) = .97$ and, in addition, $P(B|A^c) = 1 - P(B^c|A^c) = .2$. Plugging these numbers into the formula gives us

$$P(A|B) = \frac{(.58)(.03)}{(.58)(.03) + (.2)(.97)} = .082$$

or, approximately, 1/12. Thus, based on the Gold et al. study, because VD is a relatively rare disease, without additional evidence, even when the NINDS test is positive it remains unlikely that the patient has VD. □

As in Example 3.2, this form of Bayes' Theorem requires probabilities $P(B|A)$, $P(B|A^c)$ and $P(A)$ which must come from some background information. All applications of Bayes' Theorem are analogous in needing background information as inputs in order to get the desired conditional probability as output.

To generalize Bayes' Theorem from the simplest case we need the law of total probability, which gives a formula for $P(B)$ in terms of a decomposition of Ω: Given mutually exclusive events $A_1, A_2, \ldots, A_n$ that are exhaustive in the sense that $\Omega = A_1 \cup A_2 \cup \cdots \cup A_n$, we have

$$B = (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_n)$$

with the sets $B \cap A_i$ being mutually exclusive. We then have

$$\begin{aligned} P(B) &= P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_n) \\ &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_n)P(A_n) \\ &= \sum_{i=1}^{n} P(B|A_i)P(A_i). \end{aligned}$$

From this we obtain a more general form of the theorem.

Bayes' Theorem Suppose $A_1, A_2, \ldots, A_n$ are mutually exclusive with $P(A_i) > 0$, for all i, and $A_1 \cup A_2 \cup \cdots \cup A_n = \Omega$. If $P(B) > 0$ then

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_n)P(A_n)}.$$

Example 3.3 Decoding of saccade direction from SEF spike counts Bayes' Theorem is frequently used to study the ability of the relatively small networks of neurons to identify a stimulus or determine a behavior. As an example, Olson et al. (2000) reported results from a study of supplementary eye field neurons during a delayed-saccade task. In this study, described in Example 1.1 on p. 3, there were four possible saccade directions: up, right, down, and left. For each direction, and for each neuron, spike counts in fixed pre-saccade time intervals were recorded across multiple trials. From a combination of data analysis, and assumptions, the probability distribution of various spike counts could be determined for each of the four directions. If we consider a single neuron, we may then let B be the event that a particular spike count occurs, and the events A_1, A_2, A_3, and A_4 be the saccade directions up, right, down, left. Assuming the four directions are equally likely, from the probabilities $P(B|A_k)$

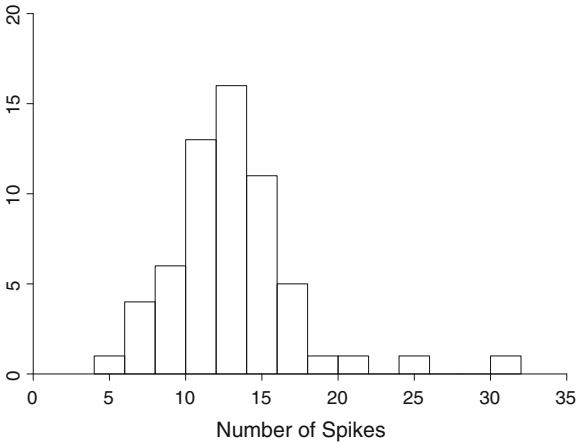

Fig. 3.2 Histogram of spike counts from a motor cortical neuron. The histogram displays 60 spike counts from a particular neuron recorded in primary motor cortex across 60 repetitions of the practiced condition.

together with Bayes' Theorem, we may determine from the spike count B the probability that the saccade will be in each of the four directions. In Bayesian decoding, the signals from many neurons are combined, and the direction A_k having the largest probability $P(A_k|B)$ is considered the "predicted" direction. In unpublished work, our colleague Dr. Valérie Ventura found that, from 55 neurons, Bayesian decoding was able to predict the correct direction more than 95 % of the time. □

3.2 Random Variables

So far we have discussed the basic rules of probability, which apply to sets representing uncertain events. A far more encompassing framework is obtained when we consider quantities measured from those events. For example, the number of times a neuron fires during a particular task may be observed, yielding a spike count. When the behavior is repeated across many trials, the spike counts will vary.

Example 3.4 Spike counts from a motor cortical neuron Matsuzaka et al. (2007) studied cortical correlates of practicing a movement repeatedly by comparing the firing of neurons in primary motor cortex during two sequential button-pressing tasks: one in which the sequence was highly practiced, and the other in which the sequence was determined at random. Figure 3.2 displays spike counts from a single neuron across 60 repetitions of the practiced condition. The histogram displays substantial variation among the counts. □

To describe variation among quantitative measurements, such as that seen in Fig. 3.2, we need to introduce mathematical objects called *random variables*, which assign to each outcome (e.g., neuronal spiking behavior on a particular trial) a number

(the spike count). The variation in data may be summarized by a histogram, as in Fig. 3.2. The uncertainty in a random variable is described by[5] its *probability distribution*. In this section we develop some of the basic attributes and properties of random variables, and their probability distributions.

At the outset it is important to emphasize the abstraction involved in using a random variable to describe observed data. Strictly speaking, random variables and their probability distributions live in the theoretical world of mathematics, while data live in the real world of observations (as depicted by Fig. 1.8). When we speak of the distribution of some data, as in the histogram in Fig. 3.2, we are talking about observed variation. On the other hand, if we use a probability distribution, such as a normal (or Gaussian) distribution or a Poisson distribution, both discussed in Chapter 5, to describe some data, we are imposing a mathematical structure. To be useful, such a structure must capture dominant features that drive scientific inferences, and a fundamental part of data analytic expertise involves appreciation of the ways inaccuracies in probabilistic description may or may not lead to misleading inferences. We discuss assessments of probability distributions, and consequences of incorrect assumptions, throughout the book. In this chapter we concentrate on essential mathematical definitions and results.

3.2.1 Random variables take on values determined by events.

Let us start by returning to the framework of Example 1.4, in which patient P.S. made a choice between two drawings on each of many trials. Suppose that the probability of her choosing the non-burning house on each trial was p, and let us consider the possibilities for two trials, assuming the outcomes were independent. For a given trial, let A be the binary (i.e., two-choice) event that she chooses the non-burning house, so that $p = P(A)$ and $P(A^c) = 1 - p$. For two trials, let us write the four possible outcomes as AA, AA^c, A^cA, A^cA^c. From independence, the probabilities of these events are

$$\begin{aligned} P(AA) &= p^2, \\ P(AA^c) &= p(1-p) \\ P(A^cA) &= (1-p)p \\ P(A^cA^c) &= (1-p)^2. \end{aligned}$$

Now take X to be the number of times, out of 2, that she chooses the non-burning house. We have

[5] We often shorten "probability distribution" to "distribution." The word *distribution* is sometimes also applied to data, where it describes the variation among the numbers. However, a *probability distribution* can refer only to a random variable.

$$P(X = 2) = p^2,$$
$$P(X = 1) = p(1-p) + (1-p)p = 2p(1-p)$$
$$P(X = 0) = (1-p)^2.$$

In this situation X is a random variable and it has a *binomial distribution*. More generally, given a sample space Ω, a *random variable* is a mapping that assigns to every element of Ω a real number. That is, if $\omega \in \Omega$ (see p. 38) then $X(\omega) = x$ is the value of the random variable X when ω occurs. In the context above, $\Omega = \{AA, AA^c, A^cA, A^cA^c\}$ and $X(AA) = 2, X(AA^c) = 1, X(A^cA) = 1, X(A^cA^c) = 0$.

In Chapter 1 we discussed the distinction between continuous and discrete data. We may similarly distinguish continuous and discrete random variables: a random variable is continuous if it can take on all values in some interval (A, B), where it is possible that either $A = -\infty$ or $B = \infty$ or both. The mathematical distinctions between discrete and continuous distributions are that (i) discrete distributions assign probabilities to specific values (such as non-negative integers) that can be separated from each other, but continuous distributions assign probabilities to intervals of non-separable numbers (such as numbers in the interval $(0, 1)$) and (ii) wherever summation signs appear for discrete distributions, integrals replace them for continuous distributions.

3.2.2 *Distributions of random variables are defined using cumulative distribution functions and probability density functions, from which theoretical means and variances may be computed.*

There are several definitions we need, which will apply to other probability distributions besides the binomial. In the case of two trials from patient P. S., discussed on p. 47, the probabilities $P(X = 0)$, $P(X = 1)$, and $P(X = 2)$ form the *probability mass function*. For convenience, as indicated in Section 3.2.3, we generally instead call the probability mass function a *probability density function* (*pdf*). We would typically write $P(X = x)$, with x taking the values 0, 1, 2, and we also use the notation $f(x) = P(X = x)$. The function $F(x) = P(X \leq x)$ is called the *cumulative distribution function (cdf)*. Thus, in the case of two trials from patient P.S. we have $F(0) = P(X = 0)$, $F(1) = P(X \leq 1) = P(X = 0) + P(X = 1)$, and $F(2) = P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$. From the pdf we can obtain the cdf, and vice-versa. When we speak loosely of the "probability distribution of X," or the "distribution of X," we will be referring generically to the range of probabilities attached to X, which could be specified by either the pdf or the cdf.

Illustration: Litter sizes of mice As a simple non-binomial example, useful for pedagogical purposes, suppose that 50 female mice were maintained in a facility, that each gave birth to a litter, and that the litter sizes may be summarized in the following table:

size	3	4	5	6	7	8
count	3	7	12	14	10	4

Let us consider choosing a mouse at random from among the 50 that gave birth, and let X be the litter size for that mouse. By dividing each count in the table above by 50 we get the following table for the probability distribution of X:

x	3	4	5	6	7	8
$f(x)$	.06	.14	.24	.28	.20	.08

Thus, $f(3) = 3/50 = .06$ signifies the probability that a randomly drawn mouse will have litter size 3. □

Notice that a plot of the counts (against x) would be a histogram of the 50 litter sizes. Aside from the divisor of 50 used in getting each probability from the corresponding count, a plot of $f(x)$ against x would look the same as the histogram of the counts; this would, instead, be a plot of the *relative frequencies*.[6] More generally, a plot of a pdf looks something like a histogram, except that the total amount of probability must equal 1.

One way to understand *any* specification of probabilities $f(x)$ is to consider them to represent relative frequencies among a population of individuals. However, in many cases the idea of a random drawing from a population is an abstraction, and may be rather unrealistic. This is actually an important philosophical point that has been argued about a great deal, but we will not go into it.

> *Details:* In experimental settings, it is quite artificial to imagine that the repeated measurements (trials) of an experiment are being drawn at random from some population of such things. Similarly, when there is a single unique event, such as the outcome of a football game, or the flip of a fair coin, we can be comfortable speaking about the probability of the outcome without any need for a population. In the case of the coin, suppose we let $X = 1$ if it comes up heads and $X = 0$ if it comes up tails, and take $f(1) = P(X = 1) = .5$ and $f(0) = P(X = 0) = .5$. We could, if we wished, imagine some very large population of fair coins, just like the one we are going to flip, among which, if flipped in just the same way, half would come up heads and half would come up tails. But we do not really need this imaginary device: thinking only about one single coin it remains easy enough to understand the idea that it is "fair" precisely when $f(1) = .5$ and $f(0) = .5$. That is, the

[6] In this context terminology is inconsistent: "frequency" can mean either "count" or "relative frequency."

notion that it is equally likely to be heads and tails does not require further elaboration. If we wished to have an operational meaning to "fair" we could take it to mean that we are willing to accept a fair bet, i.e., one in which we would win the same amount if heads as we would lose if tails. □

For our purposes, what is important is that relative frequencies sometimes define probabilities, and more generally provide a useful analogy for thinking about probability.

Now, let us go on to the concepts of mean and variance. For the 50 litter sizes in the table on p. 49 we would compute the mean as

$$\text{mean} = \frac{3(3) + 7(4) + 12(5) + 14(6) + 10(7) + 4(8)}{50} = 5.66.$$

Alternatively, we could write

$$\text{mean} = 3(\frac{3}{50}) + 4(\frac{7}{50}) + 5(\frac{12}{50}) + 6(\frac{14}{50}) + 7(\frac{10}{50}) + 8(\frac{4}{50}) = 5.66$$

which, from the table on p. 49 is the same as

$$\text{mean} = 3 \cdot f(3) + 4 \cdot f(4) + 5 \cdot f(5) + 6 \cdot f(6) + 7 \cdot f(7) + 8 \cdot f(8) = 5.66.$$

This latter form may be interpreted as the litter size we would "expect" to see ("on average") for a randomly drawn mouse, and it is an instance of the general expression for the *mean* or *expected value* or *expectation* of the random variable X:

$$\mu_X = E(X) = \sum_x x \cdot f(x). \tag{3.3}$$

Correspondingly, the *variance* of X is

$$\sigma_X^2 = V(X) = \sum_x (x - \mu_X)^2 \cdot f(x)$$

and the *standard deviation* is $\sigma_X = \sqrt{\sigma_X^2}$. The subscript X is often dropped, leaving simply μ and σ. The standard deviation summarizes the magnitude of the deviations from the mean; roughly speaking, it may be considered an average amount of deviation from the mean. It is thus a measure of the spread, or variability, of the distribution. There are alternative measures (such as $\sum_x |x - \mu| f(x)$), and these are used in special circumstances, but the standard deviation is the easiest to work with mathematically. It is, therefore, the most common measure of spread.

Note that μ_X and σ_X are theoretical quantities defined for *distributions*, and are analogous to the mean and standard deviation defined for data. In fact, if there are n values of x and we plug into (3.3) the special case $f(x) = \frac{1}{n}$ (which states

that all n values of x are equally likely) we get back[7] $\mu_X = \bar{x}$. Because data are often called *samples*, the data-based mean and standard deviation are often called the *sample mean* and the *sample standard deviation* to differentiate them from μ_X and σ_X, which are often called the *population mean and standard deviation.* This terminology distinguishes samples from "populations," rather than distributions, with the word "sample" connoting a batch of observations randomly selected from some large population. Sometimes there is a measurement process that corresponds to such random selection. However, as we have already mentioned, probability is much more general than the population/sample terminology might lead one to expect; specifically, we do not need to have a well-defined population from which we are randomly sampling in order to speak of a probability distribution. So, at least in principle, we might rather avoid calling μ_X a population mean. On the other hand, the "sample" terminology is useful for emphasizing that we are dealing with the observations, as opposed to the theoretical distribution, and it is deeply imbedded in statistical jargon. Similarly, the "population" identifier is frequently used rather than "theoretical." The crucial point is that one must be careful to distinguish between a theoretical distribution and the actual distribution of some sample of data. Many analyses assume that data follow some particular theoretical distribution, and in doing so *hope* that the match between theory and reality is pretty good. We will look at ways of assessing this match in Section 3.3.1.

The following properties are often useful.

Theorem For a discrete random variable X with mean μ_X and standard deviation σ_X we have

$$E(a \cdot X + b) = a \cdot \mu_X + b \tag{3.4}$$

$$\sigma^2_{aX+b} = a^2 \cdot \sigma^2_X \tag{3.5}$$

$$\sigma_{aX+b} = |a| \cdot \sigma_X. \tag{3.6}$$

Proof : We have

$$\begin{aligned} E(aX + b) &= \sum_x (ax + b) f(x) \\ &= a(\sum_x x f(x)) + b \sum_x f(x) \\ &= aE(X) + b \end{aligned}$$

which is the same as (3.4). The derivation of (3.5) is similar, and taking square-roots gives (3.6). □

[7] We also get $\sigma_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)^2}$ which, when we replace μ_X with $\bar{X}$, is not quite the same thing as the *sample standard deviation*; the latter requires a change from n to $n-1$ as the divisor for certain theoretical reasons, including that the sample variance then becomes an *unbiased* estimator of σ^2_X. See p. 183.

3.2.3 *Continuous random variables are similar to discrete random variables.*

Suppose X is a continuous random variable on an interval (A, B), with $A = -\infty$ and $B = \infty$ both being possible. The *probability density function* (pdf) of X will be written as $f(x)$ where now

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

and, because (from Axiom 2 on p. 38) the total probability is 1, we have

$$\int_A^B f(x)dx = 1.$$

Note that in this continuous case there is no distinction between $P(a \leq X)$ and $P(a < X)$ (we have $P(X = a) = 0$). We may think of $f(x)$ as the probability per unit of x; $f(x)dx$ is the probability that X will lie in an infinitesimal interval about x, that is, $f(x)dx = P(x \leq X \leq x + dx)$. In some contexts there are various random variables being considered and we write the pdf of X as $f_X(x)$.

A technical point is that when either $A > -\infty$ or $B < \infty$ or both, by convention, the pdf $f(x)$ is extended to $(-\infty, \infty)$ by setting $f(x) = 0$ outside (A, B). When we say that X is a continuous random variable on an interval (A, B) we will mean that $f(x) > 0$ on (A, B) and, if either A or B is a number, $f(x) = 0$ outside of (A, B). We next give several examples of continuous distributions.

Illustration: Uniform distribution Perhaps the simplest example is the *uniform distribution*. For instance, if the time of day at which births occurred followed a uniform distribution, then the probability of a birth in any given 30 min period would be the same as that for any other 30 min period throughout the day. In this case the pdf $f(x)$ would be constant over the interval from 0 to 24 h (hours). Because it must integrate to 1, we must have $f(x) = 1/24$ and the probability of a birth in any given 30 min interval starting at a hours is $\int_a^{a+.5} f(x)dx = 1/48$. When a random variable X has a uniform distribution on a finite interval (A, B) we write this as $X \sim U(A, B)$ and the pdf is $f(x) = \frac{1}{B-A}$. □

In this illustration above we have introduced a convention that is ubiquitous, both in this book and throughout statistics: the squiggle "∼" means "is distributed as."

Figure 3.3 displays pdfs for four common distributions. For the two in the top panels, exponential and gamma distributions, X may take on all positive values, i.e., values in $(0, \infty)$. The lower left panel shows a beta distribution, which is confined to the interval $(0, 1)$. A normal distribution, which ranges over the whole real line, is shown in the bottom right panel. We discuss the exponential and normal distributions briefly below and return to them, and to the beta and gamma distributions in Chapter 5.

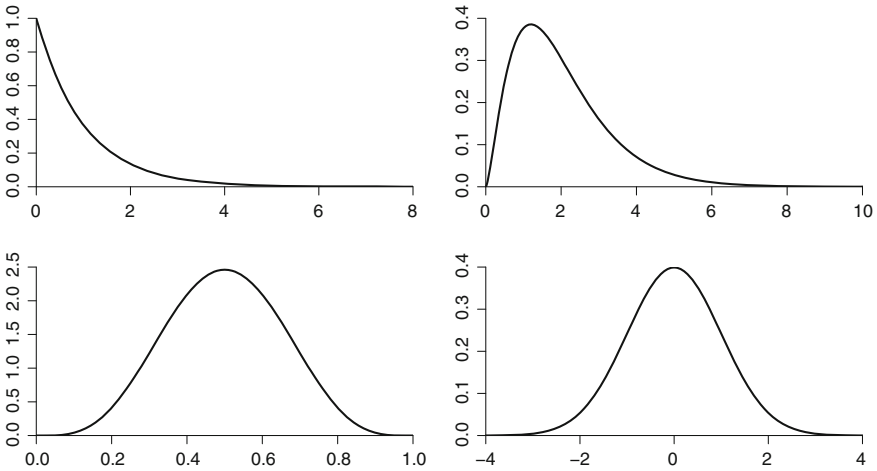

Fig. 3.3 Plots of pdfs for four continuous distributions. *Top left* Exponential. *Top right* Gamma. *Bottom left* Beta. *Bottom right* Normal. See Chapter 5 for the explanation of the latter three distributions.

Illustration: Normal distribution The normal distribution (also called the Gaussian distribution) is the most important distribution in statistical analysis. The reason for this, however, has little to do with its ability to describe data. Example 1.2, continued below, presents one of the few examples we know in which the data really appear normally distributed to a high degree of accuracy; it is rare for a batch of data *not* to be detectably non-normal. Instead, in statistical inference, the normal distribution is used to describe the variability in quantities *derived from* the data as functions of a sample mean. As we discuss in Chapter 6, according to the Central Limit Theorem, sample means are approximately normally distributed and, in Chapter 9, we will also see that functions of a sample mean are approximately normally distributed.

The normal distribution is characterized by two parameters: the mean and the standard deviation (or, equivalently, its square, the variance). When a random variable X is normally distributed we write $X \sim N(\mu, \sigma^2)$. Both in most software and in most applications, one speaks of the parameters μ and σ rather than μ and σ^2. The pdf for the normal distribution with mean μ and standard deviation σ is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2\right). \tag{3.7}$$

This pdf can be hard to use for analytic calculations because integrals such as

$$P(a \le X \le b) = \int_a^b \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2\right) dx$$

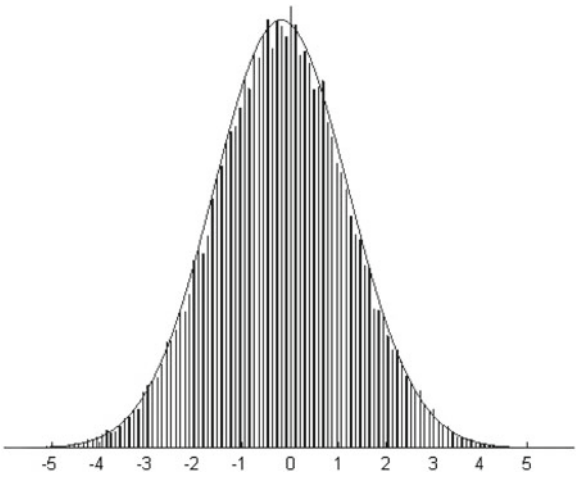

Fig. 3.4 A histogram of MEG noise at a SQUID sensor, overlaid with a normal density function (the "*bell-shaped curve*").

can not be obtained in closed form, meaning that there is no simple formula for the answer. Thus, probabilities for normal distributions are almost always obtained numerically. Because of its shape the normal pdf is often called "the bell-shaped curve." We exemplify this in the next example. □

Example 1.2 (continued from p. 5) We previously noted that the SQUID detectors in MEG are extremely sensitive, and there is nontrivial background noise that is detected in the absence of any brain signal. Figure 3.4 shows a histogram of the signal at one detector during a short period with nothing in the scanner. The noise histogram is very well approximated by a normal pdf. Indeed, this is one of the rare examples in which even on close inspection, a substantial batch of data appear to follow a normal distribution. □

In fact, the general bell shape of the distribution is not unique to the normal distribution. On the other hand, the normal is very special among bell-shaped distributions. The most important aspect of its being very special is its role in the Central Limit Theorem, which we'll come back to in Chapter 6. We also describe additional important properties of normal distributions on p. 63 and in Chapter 5.

The *cumulative distribution function*, or simply *distribution function*, is written again as $F(x)$ and is defined as in the discrete case: $F(x) = P(X \leq x)$. If $A = -\infty$ and $B = \infty$ this becomes

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

If A is a number, i.e., $-\infty < A$, then $F(x) = 0$ when $x < A$ and

$$F(x) = \int_{A}^{x} f(t)dt,$$

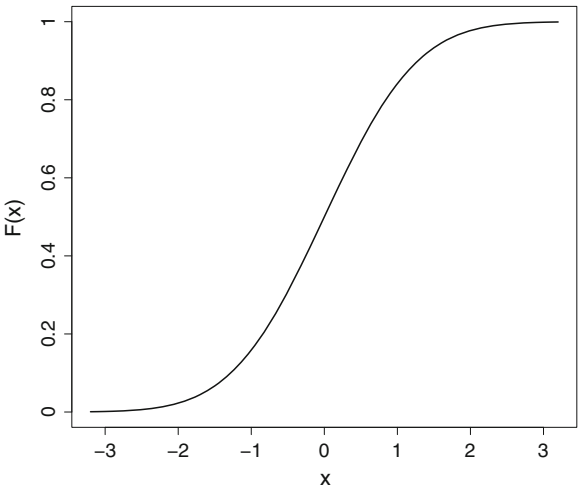

Fig. 3.5 The cdf of a $N(0, 1)$ random variable. The cdf of any other continuous distribution will, similarly, be continuous with asymptotes at 0 and 1.

while if B is a number ($B < \infty$) then $F(x) = 1$ when $x > B$. From the definition, the cumulative distribution function for a continuous distribution has a sigmoid appearance, as in Fig. 3.5, given by the following theorem.

Theorem Suppose $f(x)$ is a continuous pdf that is positive on (A, B). Then $F(x)$ is a non-decreasing function and it is strictly increasing ($F'(x) > 0$) on (A, B). In addition we have $F(x) \to 0$ as $x \to A$ and $F(x) \to 1$ as $x \to B$.

Proof : By differentiation (the Fundamental Theorem of Calculus) we have $F'(x) = f(x)$, which implies $F'(x) \geq 0$ and, by assumption, $F'(x) > 0$ on (A, B). Furthermore, because $F(x)$ is differentiable, it is also continuous. Because $f(x)$ integrates to 1 on the interval (A, B), when $A = -\infty$ we must have $F(x) \to 0$ as $x \to -\infty$ (otherwise the integral would be infinite) and when $B = \infty$ $F(x) \to 1$ as $x \to \infty$. When A is a number, from the integral form of $F(x)$, $F(A) = 0$ and $F(x) \to 0$ as $x \to A$. Similarly, when B is a number we get $F(B) = 1$ and then $F(x) \to 1$ as $x \to B$. □

In the continuous case, the *expected value* of X is

$$\mu_X = E(X) = \int_A^B xf(x)dx$$

and the *standard deviation* of X is $\sigma_X = \sqrt{V(X)}$ where

$$V(X) = \int_A^B (x - \mu_X)^2 f(x)dx$$

is the *variance* of X. Note that in each of these formulas we have simply replaced sums by integrals in the analogous definitions for discrete random variables. Note, too, that *pdf* and *cdf* values for certain continuous distributions may be computed with statistical software.[8] We again have

$$\mu_{a\cdot X+b} = a \cdot \mu_X + b \tag{3.8}$$

$$\sigma_{a\cdot X+b} = |a| \cdot \sigma_X. \tag{3.9}$$

These formulas are just as easy to prove as (3.4) and (3.6). Another formula is useful for certain calculations:

$$V(X) = E(X^2) - \mu^2 \tag{3.10}$$

and this, too, is easily verified. In many contexts the variation relative to the mean is summarized using the *coefficient of variation*, given by

$$\text{CV}(X) = \frac{\sigma}{\mu}. \tag{3.11}$$

The *quantiles* or *percentiles* are often used in working with continuous distributions: for p a number between 0 and 1 (such as .25), the p quantile or 100pth percentile (e.g., the .25 quantile or the 25th percentile) of a distribution having cdf $F(x)$ is the value η such that $p = F(\eta)$. Thus, we write the p quantile as $\eta_p = F^{-1}(p)$, where F^{-1} is the inverse cdf.

Illustration: Exponential distribution Let us illustrate these ideas in the case of the exponential distribution, which is special because it is easy to handle and also because of its importance in applications. We provide an application in Example 3.5

A random variable X is said to have an exponential distribution with parameter λ, with $\lambda > 0$, when its pdf is

$$f(x) = \lambda e^{-\lambda x} \tag{3.12}$$

for $x > 0$, and is 0 for $x \leq 0$. We will then say that X has an $Exp(\lambda)$ distribution and we will write $X \sim Exp(\lambda)$. The pdf of X when $X \sim Exp(1)$ is shown in Fig. 3.6. Also illustrated in that figure is computation of probabilities as areas under the pdf for the case

$$P(X > 2) = \int_2^\infty f(x)dx$$

which means we compute the area under the curve to the right of $x = 2$. For the exponential distribution this value is easy to compute using calculus. The cdf of an exponential distribution is

[8] The definitions of expectation and variance assume that the integrals are finite; there are, in fact, some important probability distributions that do not have expectations or variances because the integrals are infinite.

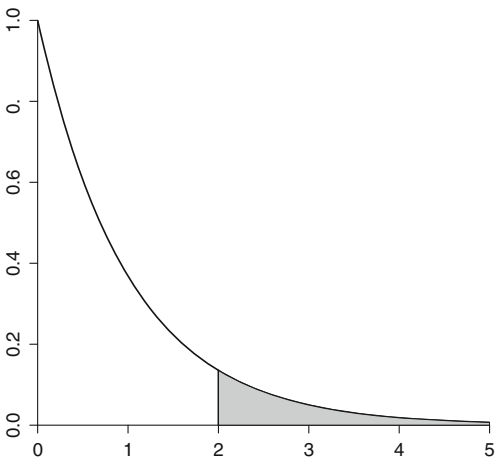

Fig. 3.6 The pdf of a random variable X having an exponential distribution with $\lambda = 1$. The *shaded area* under the pdf gives $P(X > 2)$.

$$\begin{aligned} F(x) &= \int_0^x \lambda e^{-\lambda t} dt \\ &= -e^{-\lambda t}\Big|_0^x \\ &= 1 - e^{-\lambda x}. \end{aligned}$$

Thus, when $X \sim Exp(\lambda)$, using $P(X > x) = 1 - F(x)$, we also have

$$P(X > x) = e^{-\lambda x} \tag{3.13}$$

and if $\lambda = 1$

$$P(X > 2) = 1 - F(2) = e^{-2}.$$

The quantiles are also easily obtained. For example, if $X \sim Exp(\lambda)$ the .95 quantile of X is the value $\eta_{.95}$ such that $P(X \leq \eta_{.95}) = F(\eta_{.95}) = .95$. We have

$$.95 = F(\eta_{.95}) = 1 - e^{-\lambda \eta_{.95}}$$

and we must solve $\eta_{.95}$. More generally, if we set $p = F(x)$ then

$$p = 1 - e^{-\lambda x}$$

so that

$$e^{-\lambda x} = 1 - p$$

and, therefore, $-\lambda x = \log(1 - p)$ so that

$$x = -\frac{\log(1-p)}{\lambda}.$$

Plugging in $p = .95$ gives $\eta_{.95} = -\log(.05)/\lambda$.

If $X \sim Exp(\lambda)$ then, by similar calculations, we obtain

$$\begin{aligned} E(X) &= 1/\lambda \\ V(X) &= 1/\lambda^2 \\ \sigma_X &= 1/\lambda. \end{aligned}$$

We omit the details. For future reference, we note that when we put the formulas for $E(X)$ and σ_X above in Eq. (3.11), we find the coefficient of variation of an exponentially-distributed random variable X to be

$$\mathrm{CV}(X) = 1. \tag{3.14}$$

□

If $X_1, X_2, \ldots, X_n$ are independently distributed as $Exp(\lambda)$ then their sum $Y = X_1 + X_2 + \cdots + X_n$ follows a *gamma* distribution with shape parameter n, written $Y \sim G(n, \lambda)$. The exponential is often used to describe event durations, and the gamma then becomes a sum of event durations, as illustrated in the next example.

Example 3.5 Duration of ion channel activation To investigate the functioning of ion channels, Colquhoun and Sakmann (1985) used patch-clamp methods to record currents from individual ion channels in the presence of various acetylcholine-like agonists; see also Colquhoun (2007). A set of their recordings is shown in Fig. 3.7. One of their main objectives was to describe the opening and closing of the channels in detail, and to infer mechanistic actions from the results. Colquhoun and Sakmann found that channels open in sets of activation "bursts" in which the channel may open, then shut again and open again in rapid succession, and this may be repeated, with small gaps of elapsed time during which the ion channel is closed. A burst may thus have 1 or several openings. As displayed in Fig. 3.8, Colquhoun and Sakmann examined separately the bursts having a single opening, then bursts with 2 openings, then bursts with 3, 4, and 5 openings. Panel B of Fig. 3.8 indicates that, for bursts with a single opening, the opening durations follow closely an exponential distribution. In the case of bursts with 2 openings, if each of the two opening durations were exponentially distributed, and the two were independent, then their sum—the total opening duration—would be gamma with shape parameter $\alpha = 2$. Panel C of Fig. 3.8 indicates the good agreement of the gamma with the data. The remaining panels show similar results for the other cases. □

The formulas and concepts that apply to random variables are usually stated with the notation of integrals rather than sums. This is partly because it is cumbersome to repeat everything for both continuous and discrete random variables, when the

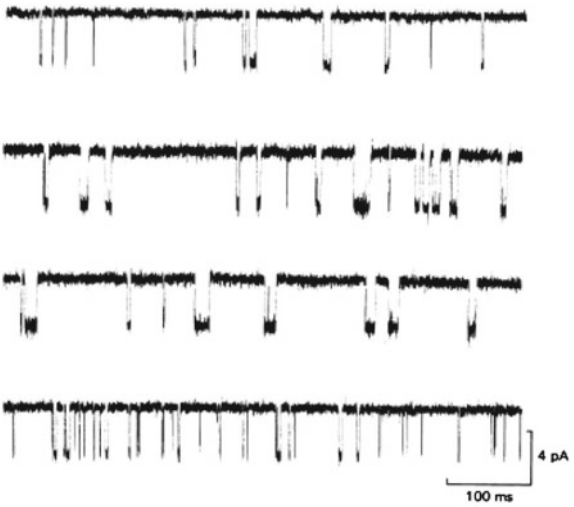

Fig. 3.7 Current recordings from individual ion channels in the presence of acetylcholine-type agonists. The records show the opening (*higher current levels*) and closing (*lower current levels*), with the timing of opening and closing being stochastic. Adapted from Colquhoun and Sakmann (1985).

results are in essence the same. In fact, there is an elegant theory of integration[9] that, among other things, treats continuous and discrete random variables together, with summations becoming special cases of integrals. Throughout our presentation we will, for the most part, discuss the continuous case with the understanding that the analogous results follow for discrete random variables. For example, we will freely use the terminology *pdf* for both continuous and discrete random variables, where for the latter it will refer to a probability mass function.

For many purposes we do not actually need formulas such as those derived for the exponential distribution. Most statistical software contains routines to generate random observations artificially[10] from standard distributions, such as those presented below, and the software will typically also provide pdf values, probabilities, and quantiles. Indeed, as we note below, random variables having essentially any continuous distribution may be generated on a computer from a program that generates $U(0, 1)$ random variables. In showing this we will have to use the cdf, which is given next.

[9] Lebesgue integration is a standard topic in mathematical analysis; see for example, Billingsley (1995).

[10] The numbers generated by the computer are really *pseudo*-random numbers because they are created by algorithms that are actually deterministic, so that in very long sequences they repeat and their non-random nature becomes apparent. However, good computer simulation programs use good random number generators, which take an extremely long time to repeat, so this is rarely a practical concern.

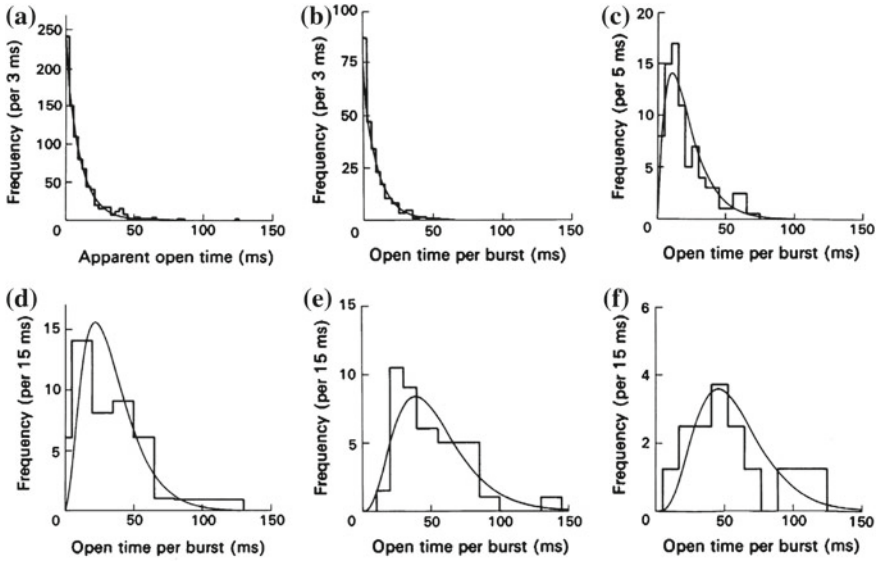

Fig. 3.8 Duration of channel openings. Panel **a** depicts the distribution of burst durations for a particular agonist. Panel **b** displays the distribution of bursts for which there was only 1 opening, with an exponential pdf overlaid. This illustrates the good fit of the exponential distribution to the durations of ion channel opening. Panel **c** displays the distribution of bursts for which there were 2 apparent openings, with a gamma pdf, with shape parameter 2, overlaid. Panel **c** again indicates good agreement. Panels **d–f** show similar results, for bursts with 3–5 openings. Adapted from Colquhoun and Sakmann (1985).

Illustration: Uniform distribution (continued from p. 52) If a continuous random variable X has cdf $F(x) = x$ on the interval $(0, 1)$ we may differentiate to get the $U(0, 1)$ pdf $f(x) = 1$. On the other hand, if $X \sim U(0, 1)$ we integrate $f(x) = 1$ to get

$$F(x) = \int_0^x 1 \cdot dx = x.$$

In other words, X has a $U(0, 1)$ distribution if and only if its cdf is $F(x) = x$ on the interval $(0, 1)$. □

Illustration: Normal distribution (continued from p. 53) When X is distributed normally with mean μ and standard deviation σ it has a pdf given by Eq. 3.7. Its cdf is given by

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2\right) dx.$$

This integral can not be evaluated in explicit form. Therefore, normal probabilities of the form $P(a \leq X \leq b)$ are obtained by numerical approximation. □

3.2.4 The hazard function provides the conditional probability of an event, given that it has not yet occurred.

Another useful characterization of a probability distribution arises in specialized contexts, including the analysis of spike train data, where a random variable X represents the waiting time until some event occurs. In the case of a spiking neuron, X would be the elapsed time since the neuron last fired, and the event of interest would be next time it fires. We want a formula for the instantaneous probability that the neuron will fire at time x, i.e., that it will fire in an interval $(x, x + dx)$, given that it has not yet fired in $(0, x)$. Assuming X is a continuous random variable, the event that the neuron has not yet fired in $(0, x)$ is the same as $X > x$. Recall that if $P(B) > 0$ then

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Applying this with A being the event $X \in (x, x+h)$ and B being the event that $X > x$ we have

$$P(X \in (x, x + h)|X > x) = \frac{F(x + h) - F(x)}{1 - F(x)}.$$

Passing to the limit as h vanishes gives

$$\lim_{h \to 0} \frac{P(X \in (x, x + h)|X > x)}{h} = \frac{f(x)}{1 - F(x)},$$

which we may interpret as the probability $X \in (x, x+dx)$ given $X > x$. The function

$$\lambda(x) = \frac{f(x)}{1 - F(x)}$$

is called the *hazard function* of X. For example, if X is the elapsed time that an ion channel is open, so that its values are times x, then $\lambda(x)dx$ becomes the probability the ion channel will close in the interval $(x, x+dx)$, given that it has remained open up to time x. Similarly, if X is the elapsed time since a neuron last fired an action potential then $\lambda(x)dx$ becomes the probability the neuron will fire in the interval $(x, x + dx)$, given that it has not yet fired again before elapsed time x. In spike train analysis, the hazard function for a neuron becomes its theoretical firing rate (its instantaneous probability of firing per unit time), which is known in general as the *intensity* or *conditional intensity* function. See Chapter 19.

The "hazard" terminology comes from lifetime analysis, where the random variable X is the lifetime (of a lightbulb or a person, etc) in units of time t and $\lambda(t)dt$ is the probability of failure (death) in the interval $(t, t + dt)$ given that failure has not yet occurred.

3.2.5 *The distribution of a function of a random variable is found by the change of variables formula.*

There are many situations in which we begin with a random variable X that has a particular distribution and we want, in addition, to obtain the distribution of another random variable $Y = g(X)$ for some function $g(x)$. This arises in the context of data transformations (discussed in Chapter 2) and it is also important in various theoretical derivations. In the simplest cases there is no need for any special formula.

Illustration: Two trials from patient P.S. Let us return to the framework on p. 47, where X is the number of times, out of 2, that P.S. chooses non-burning house, and $P(X = 2) = p^2$, $P(X = 1) = 2p(1 - p)$, $P(X = 0) = (1 - p)^2$. Suppose $Y = g(X) = 10^x$. Then we have $P(Y = 100) = p^2$, $P(Y = 10) = 2p(1 - p)$, $P(Y = 1) = (1-p)^2$. It would be easy to calculate the mean and variance of Y from these probabilities. □

When X has a continuous distribution we may obtain the pdf $f_Y(y)$ of $Y = g(X)$ using the change-of-variables formula from calculus—which follows from the chain rule.

Theorem: Pdf of a Function of a Random Variable Suppose X is a continuous random variable having pdf $f_X(x)$ for which $f_X(x) > 0$ on an interval (A, B) and $f_X(x) = 0$ otherwise; suppose further that $g(x)$ is a differentiable function and $g'(x) \neq 0$ for $x \in (A, B)$. Then the random variable $Y = g(X)$ has pdf given by

$$f_Y(y) = f_X(g^{-1}(y))|\frac{d}{dy}g^{-1}(y)|$$

wherever $y = g(x)$ for some x, and $f_Y(y) = 0$ elsewhere.

Proof details: Let us consider $x \in (A, B)$. Because $g'(x) \neq 0$, $g'(x)$ is either always positive, in which case $g(x)$ is monotonically increasing, or always negative in which case $g(x)$ is monotonically decreasing. Let us assume $g'(x) > 0$. Because $g(x)$ is monotonically increasing we then have $x \leq c \iff g(x) \leq g(c)$. We will obtain the pdf $f_y(y)$ by differentiating the cdf $F_Y(y)$, using $f_y(y) = F'_Y(y)$. Suppose $y = g(x)$ for some x. Then

$$\begin{aligned} F_Y(y) &= P(g(X) \leq y) \\ &= P(X \leq g^{-1}(y)) \\ &= F_X(g^{-1}(y)) \end{aligned}$$

where the second equality used $x \leq c \iff g(x) \leq g(c)$. Now, by the chain rule, differentiation gives

$$f_Y(y) = f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y).$$

Because we have assumed $g'(x) > 0$, this is the desired result. The case in which $g'(x) < 0$ requires a small modification of the argument above (which we leave to the attentive reader). □

Here is a simple consequence of the theorem above.

Theorem: Linear transformation of a normal random variable Suppose $X \sim N(\mu_X, \sigma_X^2)$ and let $g(x) = a + bx$ with $b \neq 0$. If $Y = g(X)$ then $Y \sim N(\mu_Y, \sigma_Y^2)$ where $\mu_Y = a + b\mu_X$ and $\sigma_Y = |b|\sigma_X$.

Proof : Notice first that the mean and standard deviation formulas follow from (3.8) and (3.9). Let us apply the transformation theorem above. We have $g^{-1}(y) = (y-a)/b$ and

$$|\frac{d}{dy}g^{-1}(y)| = \frac{1}{|b|}. \tag{3.15}$$

If we substitute $x = (y - a)/b$ into the pdf formula (3.7), multiply by the derivative factor $1/|b|$ from (3.15) as required by the theorem above, and simplify we obtain the pdf

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_Y}\exp(-\frac{1}{2}(\frac{y-\mu_Y}{\sigma_Y})^2)$$

in agreement with (3.7). □

Another result that will be used later in the book provides a way of reducing the distribution of X to a uniform distribution.

Theorem: The Probability Integral Transform, Part 1 Suppose X is a continuous random variable having pdf $f_X(x)$ and cdf $F_X(x)$, and suppose further that $f_X(x) > 0$ on an interval (A, B) and $f_X(x) = 0$ otherwise. The random variable Y defined by $Y = F_X(X)$ has a $U(0, 1)$ distribution.

Proof : First, let us note that $F_X(x)$ is strictly increasing on (A, B). It therefore has a well-defined, strictly increasing inverse function $F_X^{-1}(y)$ satisfying $F_X^{-1}(y) = x$ whenever $F_X(x) = y$. Furthermore, $x \leq c \iff F_X^{-1}(x) \leq F_X^{-1}(c)$ and $F_X(F_X^{-1}(y)) = y$. We must show that $P(Y \leq y) = y$ whenever $y \in (0, 1)$. We have

$$\begin{aligned} P(Y \leq y) = P(F_X(X) \leq y) &= P(X \leq F_X^{-1}(y)) \\ &= F_X(F_X^{-1}(y)) \\ &= y. \end{aligned}$$

□

Theorem: The Probability Integral Transform, Part 2 Suppose X is a continuous random variable having pdf $f_X(x)$ and cdf $F_X(x)$, and suppose further that $f_X(x) > 0$ on an interval (A, B) and $f_X(x) = 0$ otherwise. If $U \sim U(0, 1)$ then the random variable Y defined by $Y = F_X^{-1}(U)$ has the same distribution as X, i.e., its cdf F_Y satisfies $F_Y(y) = F_X(y)$ for all y.

Proof : The proof involves manipulations similar to those of part 1. □

This result gives a general method of generating a random variable that has a distribution with a given distribution function $F(x)$: we generate a $U(0, 1)$ random variable U and apply the transformation $F^{-1}(U)$.

We conclude with a technical result that provides transformations from one distribution to another in terms of CDFs.

Corollary to the Probability Integral Transform Suppose X and Y are continuous random variables having pdfs $f_X(x)$, $f_Y(y)$ and cdfs $F_X(x)$, $F_Y(y)$. Suppose further that $f_X(x) > 0$ on an interval (A, B) and $f_X(x) = 0$ otherwise and $f_Y(y) > 0$ on an interval (C, D) and $f_Y(y) = 0$ otherwise. Then the random variable W defined by $W = F_Y^{-1}(F_X(X))$ has the same distribution as Y, i.e., its cdf F_W satisfies $F_W(w) = F_Y(w)$ for all w.

Proof : This is simply a combination of parts 1 and 2 of the probability integral transform. □

3.3 The Empirical Cumulative Distribution Function

One way to check the accuracy with which a probability distribution fits the data is to overlay a pdf on a histogram, as in Figs. 3.4 and 3.8. (In Chapter 7 we discuss how to choose the parameter values for the pdf, e.g., the λ in an exponential.) In this section we consider another pair of graphical techniques, called P–P and Q–Q plots, which can be more sensitive than plotting the pdf.

The difficulty in examining the pdf is that its values cover a large range: it can be hard to judge deviations from a curving trend, especially when some of the values are close to zero. An alternative is to straighten things out so that a perfect fit is represented by a straight line. Both P–P and Q–Q plots accomplish this, and both are based on the cdf. We begin by defining the data-based counterpart of the theoretical cdf.

Let $X_1, \ldots, X_n$ be independent random variables all having the same distribution function $F(x)$. The *empirical cumulative distribution function*, written $\hat{F}_n(x)$, is the cdf for the discrete probability distribution that puts mass $1/n$ on each value $X_1, \ldots, X_n$, i.e.,

$$\hat{F}_n(x) = \frac{\text{number of indices } i \text{ for which } X_i \leq x}{n}.$$

That is, $\hat{F}_n(x)$ provides the proportion of the random variables, out of n, that are less than or equal to x. When n is large, we might expect this proportion to be close to the theoretical probability that each random variable is less than or equal to x, i.e., we might expect $\hat{F}_n(x)$ to be close to $F(x)$. We will see in Chapter 6 that this is necessarily so, for sufficiently large n. Figure 3.9 illustrates this in the case of a Gamma(2, 1) distribution, for samples of size $n = 10$ and $n = 200$. Specifically, to create the left panel in Fig. 3.9 we (i) used the computer to generate 10 observations $x_1, x_2, \ldots, x_{10}$ from a Gamma(2, 1) distribution, then (ii) plotted $\hat{F}_n(x)$ versus x and (iii) overlaid a plot (dashed line) of the theoretical Gamma(2, 1) cdf $F(x)$ versus x. In this case there is a reasonably close agreement between $\hat{F}_n(x)$ and $F(x)$. The agreement is much closer in the right panel, when $n = 200$.

The same procedure could be used for any set of observations $x_1, \ldots, x_n$ to check whether they seem to be consistent with random draws from a distribution with cdf $F(x)$, i.e., we could plot $F(x)$ versus x on together with a plot of $\hat{F}_n(x)$ versus x and see whether they agree well. A variation on this idea is to plot $\hat{F}_n(x)$ versus $F(x)$. This becomes a P–P plot, discussed in Section 3.3.1.

3.3.1 P–P and Q–Q plots provide graphical checks for gross departures from a distributional form.

Suppose we wish to compare a cdf $\tilde{F}(x)$ with another, similar cdf $F(x)$. If $\tilde{F}(x) \approx F(x)$, we could define $v = \tilde{F}(x)$ and $u = F(x)$, plot v against u over the range of values of x, and judge the accuracy of the approximation by the deviation of this plot from the line $v = u$. In other words, we could plot probabilities against probabilities. This is the idea behind the P-P plot (P–P for Probability-Probability), except that in examining data it is performed with the empirical cdf $\hat{F}_n(x)$ replacing $\tilde{F}(x)$. Specifically, to examine the fit of a theoretical cdf $F(x)$ to some data, we pick suitable values of x spanning the range of the data and compute $v = \hat{F}_n(x)$ and $u = F(x)$ and then plot v against u. Often, the "suitable values" of x are simply the data values themselves. In other words, for data values $x_1, \ldots, x_n$ we plot $\hat{F}_n(x_i)$ against $F(x_i)$, for $i = 1, \ldots, n$.

Example 1.2 (continued from p. 54) A P–P plot of the data shown in Fig. 3.4 is given in Fig. 3.10, where we have used a normal distribution as our theoretical $F(x)$. The plot follows extremely closely the line $y = x$. □

One difficulty with the P–P plot is that the range of both axes is [0, 1], which sometimes makes it difficult to see clearly the departures from the line $v = u$ for values of u near 0 or 1. An alternative is to pick suitable values of w between 0 and 1 and plot $\hat{F}_n^{-1}(w)$ versus $F^{-1}(w)$, both of which will be on the scale of the data. This is the idea behind the Q–Q plot, which is based on quantiles (Q–Q for Quantile-Quantile).

On p. 56 we defined the quantiles of a continuous probability distribution. The data quantiles (or observed quantiles, or sample quantiles) are analogous, but it turns out

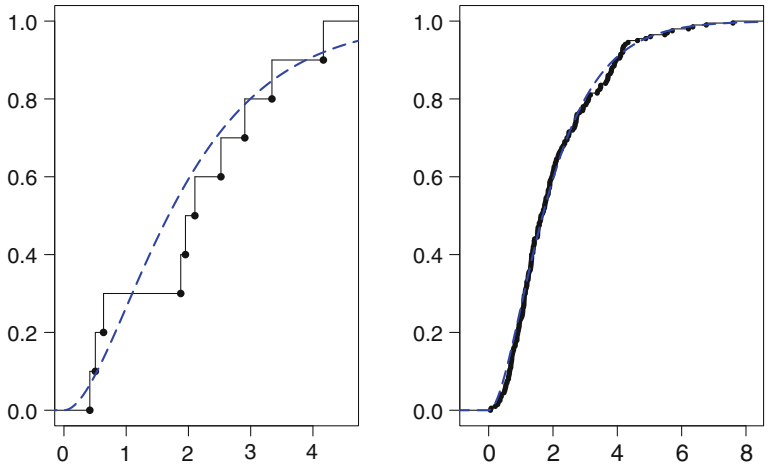

Fig. 3.9 Convergence of the empirical cdf to the theoretical cdf. The *left panel* displays the empirical cdf for a random sample of size 10 from the *Gamma* distribution whose pdf is in the *top right panel* of Fig. 3.3, together with the gamma cdf (*dashed blue line*). The *right panel* shows the empirical cdf for a random sample of size 200, again with the gamma cdf. In the *right panel* the empirical cdf is quite close to the theoretical gamma cdf.

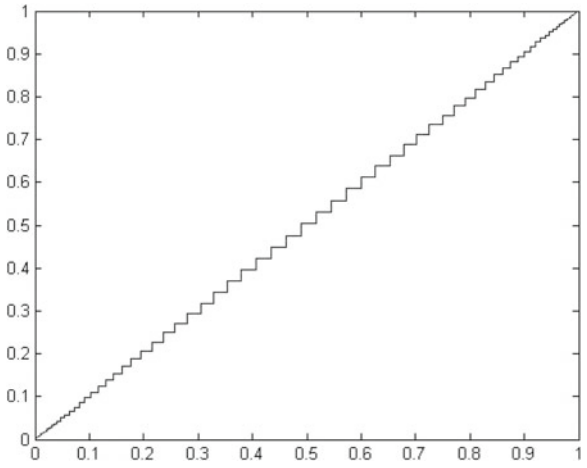

Fig. 3.10 A P–P plot of the MEG noise data from Fig. 3.4. The straightness of the plot indicates excellent agreement with the normal distribution.

that there is no unique analogue and instead one of several variants may be used. If we start from a sample of observations $x_1, x_2, \ldots, x_n$ we first put the data in ascending order according to the size of each observation: we write $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, where $x_{(1)}$ is the smallest value, $x_{(2)}$ is the second-smallest, and $x_{(n)}$ is the largest. Let us use

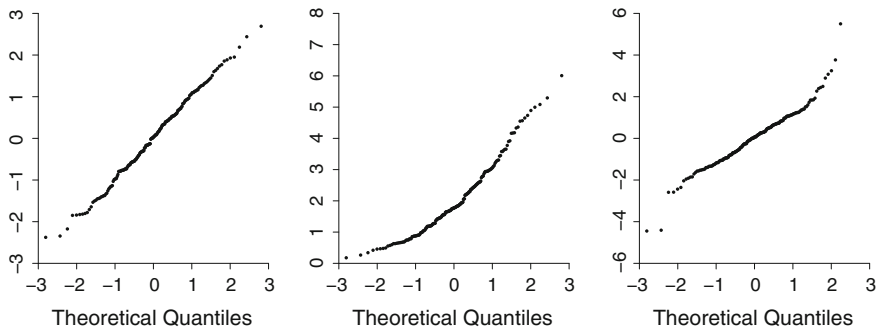

Fig. 3.11 Q–Q plots for 200 randomly-drawn observations from a three distributions. *Left* observations from a $N(0, 1)$ distribution; *middle* observations from a gamma distribution, whose pdf is shown in the *top right panel* of Fig. 3.3, which is skewed to toward high values; *right* observations from a t distribution (see Section 5.4.7), which is symmetric with heavy tails. In each case the theoretical quantiles come from a normal distribution.

r to denote the index of ordered values, meaning that $x_{(r)}$ is the rth smallest value. Working by analogy with the definition $\eta = F^{-1}(p)$ we could define the $\frac{r}{n}$ *sample quantile*, or the $100\frac{r}{n}$ *sample percentile*, by setting $p = \frac{r}{n}$ and replacing F with $\hat{F}_n$ to get $\hat{F}_n^{-1}(\frac{r}{n}) = x_r$. We then define

$$\eta_{(r)} = \tilde{F}^{-1}(\frac{r}{n})$$

for $r = 1, \ldots, n$ and plot the ordered data against these values. That is, we plot the points $(\eta_{(1)}, x_{(1)}), \ldots, (\eta_{(n)}, x_{(n)})$. Most software modifies the details of this procedure, but the idea remains the same.

> *Details:* A common variation is to take x_r to be the $100\frac{r-.5}{n}$ *sample percentile*. To see why this makes some sense, suppose we have $n = 7$ ordered observations. Then the 4th is the median. This divides the 7 numbers into the 3 smallest and the 3 largest and, effectively says that the 4th is part of both the smallest half of the numbers and the largest half of the numbers. It could therefore be considered the 3.5th ordered value. The reasoning behind the designation of $x_{(r)}$ as the $\frac{r-.5}{n}$ quantile is similar. Statistical software sometimes chooses alternative definitions based on expected values of $x_{(r)}$ under particular assumptions. Also, in creating a P–P plot, some software plots $\hat{F}(x_{(r)})$ against $\frac{r-.5}{n}$. □

Figure 3.11 displays three Q–Q plots, for which the theoretical quantiles are based on the normal distribution. Thus, we would make these plots in order to check whether the data could reasonably be described by a normal distribution. The three data sets were generated on the computer from three very different probability distributions.

The first comes from a normal distribution, the second from a gamma distribution, which is skewed toward high values, and the third from a t-distribution, which is symmetric but has heavy tails in both directions. The first plot shows adherence to a linear relationship between the observed and theoretical quantiles. The second, for skewed data, shows upward curvature: the points on the far right-hand side of the plot correspond to data values that are farther from the middle than would be expected if normal (the observed quantiles for those points are too large for the theoretical quantiles—the data should have been pulled in toward the middle—so the points appear too high) and those on the far left-hand side are too close to the middle (the observed quantiles are again too large—the data should now be pushed away from the middle—and the points are again too high). The third plot, for symmetrical but heavy-tailed data, has an S-shaped tendency (the observed quantiles are too large on the far right-hand side and too small on the left; on both extremes, to look more normal, the data should be pushed back toward the middle).

Although such plots are very useful for revealing serious departures from normality, small wiggles in these plots are very common even for computer-generated normal data. Thus, strong nonlinearities are what we look for, and even these are sometimes a bit subtle. Figure 3.12 shows Q–Q plots, where again the theoretical quantiles are based on the normal distribution, with data being 30 randomly drawn observations from a $N(0, 1)$ distribution. That is, these were computer-generated data from a $N(0, 1)$ distribution and one might expect Q–Q plots from the correct distribution to be nearly exactly linear. The 6 plots show 6 replications of this random number generation and plotting. The departures from linearity indicate that randomly drawn observations fluctuate; they do not conform perfectly to what is theoretically "expected." Or, put differently, what we *should expect* is that small samples of truly normal data will be somewhat erratic and less regular than the theoretical curve based on infinitely much data. This basic lesson applies to all probability distributions, and it also applies to many situations other than examination of Q–Q plots. It is something we must keep in mind when using our personal perceptions[11] to judge random quantities.

In a P–P plot we look for departures from the line $y = x$, and the same holds for a Q–Q plot (except that sometimes a scale factor changes the slope, so that departures from linearity are of interest). In either case it does not matter which of the variables is plotted on the x-axis and which is plotted on the y-axis. There is no fixed convention here and, in interpreting the plots, a data analyst must check which choice is made by the software being used.

[11] The cognitive psychology of perception of randomness has been studied quite extensively. See, for instance, Gilovich et al. (1985).

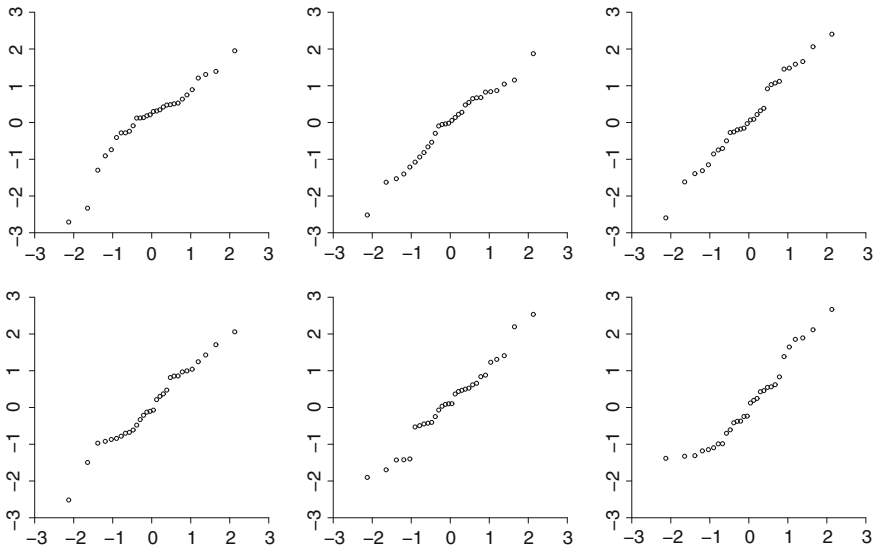

Fig. 3.12 Normal Q–Q plots for 30 randomly-drawn observations from a $N(0, 1)$, repeated six times. The plots are more or less linear, but display mild departures (wiggles, etc.) from linearity.

3.3.2 Q–Q and P–P plots may be used to judge the effectiveness of transformations.

In Chapter 2 we discussed transformations of data, especially to improve symmetry. There we used histograms as displays. An alternative is to use Q–Q or P–P plots.

Example 2.1 (continued from p. 24) Figure 3.13 provides Q–Q plots for the human eye saccade data shown in Chapter 2. The logarithm makes the distribution more symmetrical, and the reciprocal does an even better job. An unusually long delay in the saccade time becomes apparent as an outlier in the latter plot.

On the bottom right of Fig. 3.13 is a Q–Q plot from a different patient, for whom much of the data were unusable. We have included this because the plot has the classic S-shape, indicating a "heavy-tailed" distribution. Power transformations do not fix this problem. If one wishes to analyze data of this sort it is important to use a statistical procedure either specifically designed for such situations or having well-understood behavior in the presence of heavy-tailed distributions. We discuss nonparametric procedures in Chapters 9 and 11.

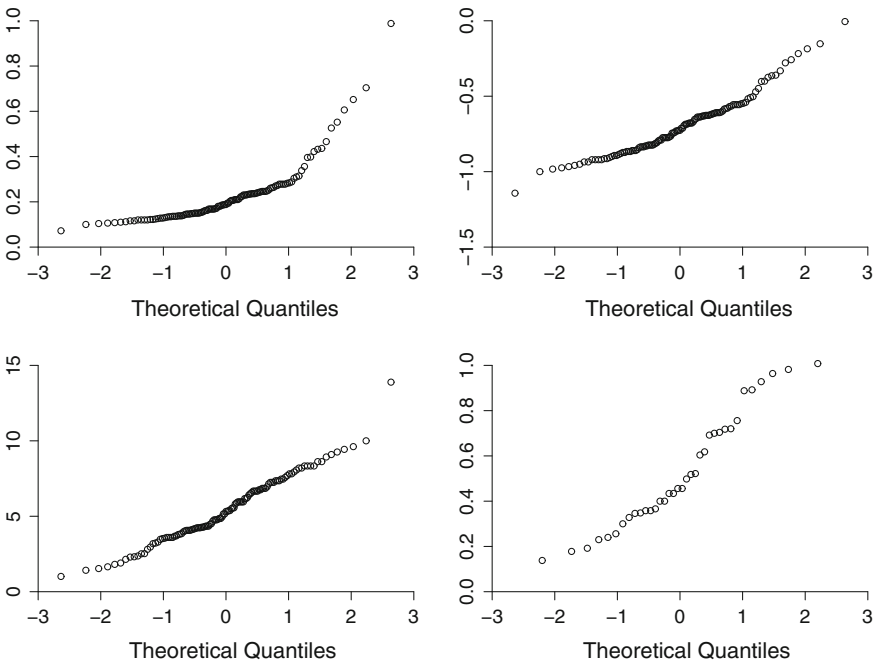

Fig. 3.13 Q–Q plots. *Upper left* Q–Q plot for the data from a particular patient, shown in Chapter 1, from the study by Behrmann et al. (2002); *upper right* Q–Q plot of the same data following a log transformation; *lower left* Q–Q plot following a reciprocal transformation. The plot for the log-transformed data is straighter than that for the raw data; the plot for the reciprocal-transformed data is straighter still. *Lower right* Q–Q plot of data from a different patient, which exhibits an S shape.

Chapter 4
Random Vectors

In most experimental settings data are collected simultaneously on many variables, and the statistical modeling problem is to describe their *joint* variation, meaning their tendency to vary together. The starting point involves m-dimensional *random vectors* (where m is some positive integer), which are the natural multivariate extension of random variables. The fundamental concepts of distribution, expectation, and variance discussed in Chapter 3 extend fairly easily to m-dimensions. We review the essential definitions in Section 4.1, then consider bivariate dependence in Section 4.2 and multivariate dependence in Section 4.3. The most commonly applied measure of association between two random variables is the correlation, defined in Section 4.2.1. As we explain, correlation is a measure of linear dependence. Nonlinear dependence is often quantified by mutual information, which we define in Section 4.3.2. In Section 4.3.4 we apply concepts of multivariate dependence to the problem of classification, and show that Bayes classifiers provide the best possible classification accuracy.

4.1 Two or More Random Variables

Let us begin our discussion of multivariate dependence with a motivating example.

Example 4.1 Tetrode spike sorting One relatively reliable method of identifying extracellular action potentials *in vivo* is to use a "tetrode." As pictured in panel A of Fig. 4.1, a tetrode is a set of four electrodes that sit near a neuron and record slightly different voltage readings in response to an action potential. The use of all four recordings allows more accurate discrimination of a particular neuronal signal from the many others that affect each of the electrodes. Action potentials corresponding to a particular neuron are identified from a complex voltage recording by first "thresholding" the recording, i.e., identifying all events that have voltages above the threshold. Each thresholded event is a four-dimensional vector (x_1, x_2, x_3, x_4), with x_i being the voltage amplitude (in millivolts) recorded at the ith electrode or "channel." Panels b-d display data from a rat hippocampal CA1 neuron. Because

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_4,

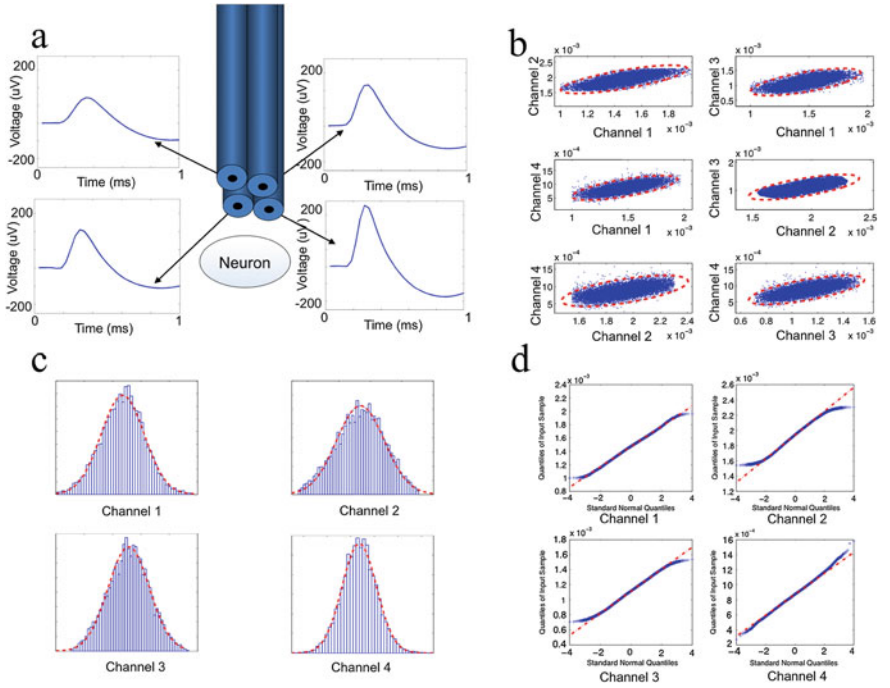

Fig. 4.1 Spike sorting from a tetrode recording. Panel **a** is a diagram of a tetrode, which is a set of four electrodes; also shown are signals recorded from a particular neuron (indicated as an elliptical disk) that is sitting near the tetrode. Panel **b** displays the six pairs of plots of event amplitudes. For instance, the *top left* plot in panel **b** shows the event amplitudes for channel 1 (x-axis) and channel 2 (y-axis). Also overlaid on the data in panel **b** are 95 % probability contours found from a suitable bivariate normal distribution. Panel **c** displays histograms for the event amplitudes on each channel, together with fitted normal pdfs, and panel **d** provides the corresponding normal Q-Q plots.

there are six pairs of the four tetrodes (channel 1 and channel 2, channel 1 and channel 3, etc.) six bivariate plots are shown in panel b. The univariate distributions are displayed in panel c and Q-Q plots are in panel d. We return to this figure in Chapter 5. □

Particularly when the number of dimensions m is greater than 2 it becomes hard to visualize multidimensional variation. Some form of one and two-dimensional visualization is usually the best we can do, as illustrated in Fig. 4.1. As we contemplate theoretical representations, the possibilities for interactions among many variables quickly become quite complicated. Typically, simplifications are introduced and an important challenge is to assess the magnitude of any distortions they might entail. We content ourselves here with a discussion of multivariate means and variances, beginning with the bivariate case.

4.1.1 The variation of several random variables is described by their joint distribution.

If X and Y are random variables, their *joint distribution* may be found from their *joint pdf*, which we write as $f(x, y)$:

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_c^d \int_a^b f(x, y) dx dy.$$

In the discrete case the integrals are replaced by sums. Each individual or *marginal* pdf is obtained from the joint pdf by integration (or, in the discrete case, summation): if $f_X(x)$ is the pdf of X then

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy.$$

Illustration: Spike Count Pairs Suppose we observe spike counts for two neurons recorded simultaneously over an interval of 100 ms. Let X and Y be the random variables representing the two spike counts. We may specify the joint distribution by writing down its pdf. Suppose it is given by the following table:

Y	X = 0	1	2
2	.03	.07	.10
1	.06	.16	.08
0	.30	.15	.05

This means the probability that the first neuron spikes once and the second neuron spikes twice, during the observation interval, is $P(X = 1, Y = 2) = .07$. We may compute from this table all of the marginal probabilities. For example, we have the following marginal probabilities: $P(X = 1) = .07 + .16 + .15 = .38$ and $P(Y = 2) = .03 + .07 + .10 = .2$. □

The example above explains some terminology. When we compute $P(Y = 2)$ we are finding a probability that would naturally be put in the *margin* of the table; thus, it is a marginal probability.

More generally, if $X_1, X_2, \ldots, X_n$ are continuous random variables their joint distribution may be found from their joint pdf $f(x_1, x_2, \ldots, x_n)$:

$$\begin{aligned} &P(a_1 \leq X_1 \leq b_1, a_2 \leq X_2 \leq b_2, \ldots, a_n \leq X_n \leq b_n) \\ &\quad = \int_{a_n}^{b_n} \cdots \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n \end{aligned}$$

and the marginal pdf of the ith random variable X_i is given by

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_{i-1} dx_{i+1} \cdots dx_n$$

where all the variables other than x_i are integrated out. The joint cdf is defined by

$$F(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n).$$

Once again, the formulas for discrete random variables are analogous.

Let us introduce a general notation. Sometimes we will write $X = (X_1, X_2, \ldots, X_n)$, so that X becomes a *random vector* with pdf (really, a joint pdf for its components) $f_X(x) = f_{(X_1, X_2, \ldots, X_n)}(x_1, x_2, \ldots, x_n)$. When we must distinguish row vectors from column vectors we will usually want X to be an $n \times 1$ column vector, so we would instead write $X = (X_1, X_2, \ldots, X_n)^T$, where the superscript T denotes the transpose of a matrix.

A very useful and important fact concerning two or more random variables is that their expectation is linear in the sense that the expectation of a linear combination of them is the corresponding linear combination of their expectations.

Theorem: Linearity of Expectation For random variables X_1 and X_2 we have

$$E(aX_1 + bX_2) = aE(X_1) + bE(X_2).$$

More generally, for random variables $X_1, X_2, \ldots, X_n$ we have

$$E(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i E(X_i). \tag{4.1}$$

Proof: Consider the case of two random variables and assume X_1 and X_2 are continuous. Let $f_1(x_1), f_2(x_2)$, and $f_{12}(x_1, x_2)$ be the marginal and joint pdfs of X_1 and X_2, and assume these random variables take values in the respective intervals (A_1, B_1) and (A_2, B_2) (which could be infinite). We have

$$\begin{aligned}
E(aX_1 + bX_2) &= \int_{A_2}^{B_2} \int_{A_1}^{B_1} (ax_1 + bx_2) f_{12}(x_1, x_2) dx_1 dx_2 \\
&= a \int_{A_2}^{B_2} \int_{A_1}^{B_1} x_1 f_{12}(x_1, x_2) dx_1 dx_2 \\
&\quad + b \int_{A_2}^{B_2} \int_{A_1}^{B_1} x_2 f_{12}(x_1, x_2) dx_1 dx_2 \\
&= a \int_{A_1}^{B_1} x_1 \int_{A_2}^{B_2} f_{12}(x_1, x_2) dx_2 dx_1 \\
&\quad + b \int_{A_2}^{B_2} x_2 \int_{A_1}^{B_1} f_{12}(x_1, x_2) dx_1 dx_2
\end{aligned}$$

$$= a \int_{A_1}^{B_1} x_1 f_1(x_1) dx_1 + b \int_{A_2}^{B_2} x_2 f_2(x_2) dx_2$$
$$= aE(X_1) + bE(X_2).$$

The proof in the discrete case would replace the integrals by sums, and the proof in the general case of n variables follows the same steps. □

4.1.2 *Random variables are independent when their joint pdf is the product of their marginal pdfs.*

We previously said that two events A and B are independent if $P(A \cap B) = P(A)P(B)$, and we used this in the context of random variables that identify dichotomous events. For example, if p is the probability that P.S. chooses the non-burning house on any given trial, we said that p^2 will be the probability she chooses the non-burning house on both of two trials. Generally, we say that two random variables X and Y are *independent* if

$$P(a \le X \le b \text{ and } c \le Y \le d) = P(a \le X \le b)P(c \le Y \le d) \tag{4.2}$$

for all choices of a, b, c, d. It follows that when X and Y are independent we also have

$$f(x, y) = f_X(x) f_Y(y) \tag{4.3}$$

for all x and y. Indeed, when X and Y are random variables with pdf $f(x, y)$, they are independent if and only if Eq. (4.3) holds. Thus, we may instead take (4.3) as the definition of independence of two random variables.

> *Details:* Suppose X and Y are continuous random variables. If (4.3) holds we may integrate both sides over the region $(a, b) \times (c, d)$ to obtain (4.2). If (4.2) holds we rewrite it in terms of integrals, set $b = x$ and $d = y$, and compute the mixed second partial derivatives with respect to x and y. This gives (4.3).
> If X and Y are discrete, the integrals are replaced by sums. If (4.2) holds then we set $a = b = x$ and $c = d = y$ to get (4.3). If (4.3) holds for all x and y then the double summation on the left-hand side of (4.2) factors as
>
> $$\sum_{a \le x \le b, c \le y \le d} f(x) f(y) = \sum_{a \le x \le b} f(x) \sum_{c \le y \le d} f(y)$$
>
> which is the right-hand side of (4.2). □

Illustration: Spike Count Pairs (continued from p. 73) We return once again to the joint distribution of spike counts for two neurons, given by the table on p. 73. Are X and Y independent?

The marginal pdf for X is $f_X(0) = .39$, $f_X(1) = .38$, $f_X(2) = .23$ and the marginal pdf for Y is $f_Y(0) = .50$, $f_Y(1) = .30$, $f_Y(2) = .20$. We thus obtain $f_X(0)f_Y(0) = .195 \neq .30 = f(0,0)$, which immediately shows that X and Y are not independent. □

We may generalize the definition of independence to multiple random variables: we say that $X_1, X_2, \ldots, X_n$ are independent random variables if their joint pdf $f_{(X_1,X_2,\ldots,X_n)}(x_1, x_2, \ldots, x_n)$ is equal to the product of their marginal pdfs,

$$f_{(X_1,X_2,\ldots,X_n)}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i).$$

In the previous subsection we showed that the expectation of a sum is always the sum of the expectations. In general, it is not true that the variance of a sum of random variables is the sum of their variances (the formula is instead more complicated; see (4.6)), but this *is* true under independence.

Theorem: Variance of a Sum of Independent Random Variables For independent random variables X_1 and X_2 we have

$$V(aX_1 + bX_2) = a^2V(X_1) + b^2V(X_2). \tag{4.4}$$

More generally, for independent random variables $X_1, X_2, \ldots, X_n$ we have

$$V\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 V(X_i). \tag{4.5}$$

Proof: The proof is similar to that of the theorem on linearity of expectations, except that the factorization of the joint pdf, due to independence, must be used. □

The formula (4.5) may fail if X_1 and X_2 are not independent. For example, if $X_2 = -X_1$ then $X_1 + X_2 = 0$ and $V(X_1 + X_2) = 0$. A general formula appears in Eq. (4.6).

4.2 Bivariate Dependence

In Section 4.1.2 we said that random variables $X_1, X_2, \ldots, X_n$ are independent if their joint pdf $f_{(X_1,X_2,\ldots,X_n)}(x_1, x_2, \ldots, x_n)$ is equal to the product of their marginal pdfs. We now consider the possibility that $X_1, X_2, \ldots, X_n$ are *not* independent and develop some simple ways to quantify their dependence. In the case of two random

variables the most common way to measure dependence is through their correlation, which is discussed in Section 4.2.1. We first interpret the correlation as a measure of linear dependence then, in Section 4.2.2, describe its role in the bivariate normal distribution. After we discuss conditional densities in Section 4.2.3 we re-interpret correlation using conditional expectation in Section 4.2.4. We then turn to the case of arbitrarily many random variables ($X_1, \ldots, X_n$ with $n \geq 2$), providing results in Section 4.3 that will be useful later on. We discuss general multivariate normal distributions later, in Section 5.5.

4.2.1 *The linear dependence of two random variables may be quantified by their correlation.*

When we consider X and Y simultaneously, we may characterize numerically their joint variation, meaning their tendency to be large or small together. This is most commonly done via the *covariance* of X and Y which, for continuous random variables, is

$$\begin{aligned} Cov(X, Y) &= E\left((X - \mu_X)(Y - \mu_Y)\right) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x, y) dx dy \end{aligned}$$

and for discrete random variables the integrals are replaced by sums. The covariance is analogous to the variance of a single random variable. We now generalize Eq. (4.5) to the case in which the random variables may not be independent.

Theorem: Variance of a Sum of Random Variables For random variables X_1 and X_2 we have

$$V(aX_1 + bX_2) = a^2 V(X_1) + b^2 V(X_2) + 2abCov(X_1, X_2).$$

More generally, for random variables $X_1, X_2, \ldots, X_n$ we have

$$V(\sum_{i=1}^{n} a_i X_i) = \left(\sum_{i=1}^{n} a_i^2 V(X_i)\right) + 2 \sum_{i<j} a_i a_j Cov(X_i, X_j). \tag{4.6}$$

Proof: The proof follows from the definition by straightforward algebraic manipulations and is omitted. □

The covariance depends on the variability of X and Y individually, as well as their joint variation, and therefore depends on scaling. For instance, as is immediately verified from the definition, $Cov(3X, Y) = 3Cov(X, Y)$. To obtain a measure of joint variation that does not depend on the variance of X and Y, we standardize. The *correlation* of X and Y is

$$Cor(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

where σ_X and σ_Y are the standard deviations of X and Y. This is also often called the *Pearson correlation*, after Karl Pearson who studied extensively this and other measures of association.[1] The correlation is also called the *correlation coefficient* and is commonly denoted by ρ, as in

$$\rho_{XY} = Cor(X, Y),$$

and when it is clear which random variables are being considered the subscript is omitted.

Let us emphasize that, just as a theoretical mean μ and standard deviation σ should be distinguished from the sample mean $\bar{x}$ and sample standard deviation s, the theoretical quantities $Cov(X, Y)$ and $Cor(X, Y)$ should be distinguished from the analogous quantities computed from data: if $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ are two batches of numbers their *sample correlation* is

$$r_{XY} = \frac{\frac{1}{n-1} \sum_{i=1}^{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \tag{4.7}$$

where s_x is the sample standard deviation of $x_1, \ldots, x_n$ and s_y is the sample standard deviation of $y_1, \ldots, y_n$. The numerator in (4.7) is the *sample covariance* of these two samples. The quantity r_{XY} in (4.7) is also often called the *sample Pearson correlation* and sometimes "Pearson correlation" may mean either ρ_{XY} or r_{XY}. The sample correlation is also often written using the alternate notation

$$\hat{\rho}_{XY} = r_{XY} \tag{4.8}$$

to indicate that ρ_{XY} is being estimated by the sample correlation. We discuss the sample correlation further in Chapter 12. In the remainder of this section we focus exclusively on $Cor(X, Y)$.

It is easy to check that $Cor(X, Y)$ is invariant to linear rescaling of X and Y and it may be shown that $-1 \leq Cor(X, Y) \leq 1$. The latter is an instance of what is known in mathematical analysis as the Cauchy-Schwartz inequality. When X and Y are independent their covariance, and therefore also their correlation, is zero.

> *Details:* This last fact follows from the definition of covariance: if X and Y are independent we have $f(x, y) = f_X(x) f_Y(y)$ and then

[1] The concept of association also played a prominent role in Pearson's influential book *The Grammar of Science*, the first edition of which appeared in 1892. For a discussion of Pearson's research see Stigler (1986).

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} & (x-\mu_X)(y-\mu_Y)f(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)f_X(x)(y-\mu_Y)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty}(x-\mu_X)f_X(x)dx\int_{-\infty}^{\infty}(y-\mu_Y)f_Y(y)dy\end{aligned}$$

but from the definition of μ_Y

$$\int_{-\infty}^{\infty}(y-\mu_Y)f_Y(y)dy = 0$$

(and similarly the integral over x is zero). □

We now illustrate the calculation of correlation in a simple example, introduced earlier.

Illustration: Spike count pairs (continued from p. 76) We return to the joint distribution of spike counts for two neurons, discussed on p. 73, with joint pdf given by the following table:

Y				
2	.03	.07	.10	
1	.06	.16	.08	
0	.30	.15	.05	
	0	1	2	X

We may compute the covariance and correlation of X and Y as follows:

$$\begin{aligned}\mu_X &= 0 + 1\cdot(.38) + 2\cdot(.23)\\ \mu_Y &= 0 + 1\cdot(.30) + 2\cdot(.2)\\ \sigma_X &= \sqrt{.39\cdot(0-\mu_X)^2 + .38\cdot(1-\mu_X)^2 + .23\cdot(2-\mu_X)^2}\\ \sigma_Y &= \sqrt{.5\cdot(0-\mu_Y)^2 + .3\cdot(1-\mu_Y)^2 + .2\cdot(2-\mu_Y)^2}\end{aligned}$$

which gives

$$\begin{aligned}\mu_X &= .84\\ \mu_Y &= .7\\ \sigma_X &= .771\\ \sigma_Y &= .781.\end{aligned}$$

We then get

$$\sum f(x,y)(x-\mu_X)(y-\mu_Y) = .272$$

and

$$\frac{\sum f(x,y)(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} = .452.$$

Thus, the correlation is $\rho \approx .45$. □

The correlation is undoubtedly the most commonly used measure of association between two random variables, but it is rather special. For one thing, $Cor(X, Y) = 0$ does *not* imply that X and Y are independent. Here is a counterexample.

Illustration: Dependent variables with zero correlation. Suppose X is a continuous random variable having a distribution that is symmetric about 0, meaning that for all x we have $f_X(-x) = f_X(x)$, and let us assume that $E(X^4)$ is finite (i.e., $E(X^4) < \infty$). From symmetry we have

$$\int_{-\infty}^{0} x f_X(x)dx = -\int_{0}^{\infty} x f_X(x)dx$$

so that

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x)dx \\ &= \int_{-\infty}^{0} x f_X(x)dx + \int_{0}^{\infty} x f_X(x)dx = 0 \end{aligned}$$

and, similarly, $E(X^3) = 0$. Now let $Y = X^2$. Clearly X and Y are not independent: given $X = x$ we know that $Y = x^2$. On the other hand,

$$Cov(X, Y) = E(X(Y - \mu_Y)) = E(X^3) - E(X)\mu_Y = 0.$$

Therefore, $Cor(X, Y) = 0$. □

A more complete intuition about correlation may be found from the next result. Suppose we wish to predict a random variable Y based on another random variable X. That is, suppose we take a function $g(x)$ and apply it to X to get $g(X)$ as our prediction of Y. To evaluate how well $g(X)$ predicts Y we can examine the average size of the error, letting under-prediction ($g(x) < y$) be valued the same as over-prediction ($g(x) > y$). A mathematically simple criterion that accomplishes this is expected squared error, or *mean squared error*, $E((Y - g(X))^2)$. We therefore pose the problem of finding the form of $g(x)$ that minimizes mean squared error. There is a general solution to this problem, which we give in Section 4.2.4. For now we consider the special case in which $g(x)$ is linear, and find the best linear predictor in the sense of minimizing mean squared error.

Theorem: Linear prediction Suppose X and Y are random variables having variances σ_X^2 and σ_Y^2 (with $\sigma_X^2 < \infty$ and $\sigma_Y^2 < \infty$). In terms of mean squared error, the best linear predictor of Y based on X is $\alpha + \beta X$ where

$$\beta = \rho \frac{\sigma_Y}{\sigma_X} \tag{4.9}$$

$$\alpha = \mu_Y - \beta \mu_X \tag{4.10}$$

and $\rho = Cor(X, Y)$. In other words, the values of α and β given by (4.10) and (4.9) minimize $E((Y - \alpha - \beta X)^2)$. With α and β given by (4.10) and (4.9) we also obtain

$$E\left((Y - \alpha - \beta X)^2\right) = \sigma_Y^2 (1 - \rho^2). \tag{4.11}$$

Proof details: Write

$$Y - \alpha - \beta X = (Y - \mu_Y) - (\alpha + \beta(X - \mu_X)) + \mu_Y - \beta \mu_X$$

then square both sides, take the expected value, and use the fact that for any constants c and d, $E(c(X - \mu_X)) = 0 = E(d(Y - \mu_Y))$. This leaves

$$E\left((Y - \alpha - \beta X)^2\right) = \sigma_Y^2 + \beta^2 \sigma_X^2 - 2\beta\rho\sigma_X\sigma_Y + (\mu_Y - \alpha - \beta\mu_X)^2. \tag{4.12}$$

Minimizing this quantity by setting

$$0 = \frac{\partial}{\partial \alpha} E\left((Y - \alpha - \beta X)^2\right)$$

and

$$0 = \frac{\partial}{\partial \beta} E\left((Y - \alpha - \beta X)^2\right)$$

and then solving for α and β gives (4.10) and (4.9). Inserting these into (4.12) gives (4.11). □

Let us now interpret these results by considering how well $\alpha + \beta X$ can predict Y. From (4.11) we can make the prediction error (the mean squared error) smaller simply by decreasing σ_Y. In order to standardize we may instead consider the ratio $E((Y - \alpha - \beta X)^2)/\sigma_Y^2$. Solving (4.11) for ρ^2 we get

$$\rho^2 = 1 - \frac{E\left((Y - \alpha - \beta X)^2\right)}{\sigma_Y^2}. \tag{4.13}$$

Expression (4.13) shows that the better the linear prediction is, the closer to 1 ρ^2 will be; and, conversely, the prediction error is maximized when $\rho = 0$. Furthermore, we have $\rho > 0$ for positive association, i.e, $\beta > 0$, and $\rho < 0$ for negative association, i.e, $\beta < 0$. Based on (4.13) we may say that correlation is a measure of linear association between X and Y. Note that the counterexample on p. 80, in which X and Y were perfectly dependent yet had zero correlation, is a case of nonlinear dependence.

4.2.2 A bivariate normal distribution is determined by a pair of means, a pair of standard deviations, and a correlation coefficient.

As you might imagine, to say that two random variables X and Y have a bivariate normal distribution is to imply that each of them has a (univariate) normal distribution and, in addition, they have some covariance. Actually, there is a mathematical subtlety here: the requirement of bivariate normality is much more than that each has a univariate normal distribution. We return to this technical point later in this section. For now, we will say that X and Y have a bivariate normal distribution when they have a joint pdf

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)}$$

where $\rho = Cor(X,Y)$ and we assume that $\sigma_X > 0$, $\sigma_Y > 0$, and $-1 < \rho < 1$. We may also write this pdf in the form

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} exp^{(-\frac{1}{2}Q(x,y))} \tag{4.14}$$

where

$$Q(x,y) = \frac{1}{1-\rho^2}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right).$$

Note that the factor multiplying the exponential in (4.14) does not depend on either x or y and that $Q(x,y)$ is a quadratic centered at the mean vector; we have inserted the minus sign as a reminder that the density has a maximum rather than a minimum. An implication involves the contours of the pdf. In general, a *contour* of a function $f(x,y)$ is the set of (x,y) points such that $f(x,y) = c$ for some particular number $c > 0$. When the graph $z = f(x,y)$ is considered, a particular contour represents a set of points for which the height of $f(x,y)$ is the same. The various contours of $f(x,y)$ are found by varying c. The contours of a bivariate normal pdf satisfy $Q(x,y) = c^*$, for some number c^*, and it may be shown that the set of points (x,y) satisfying such a quadratic equation form an ellipse (see Eq. (A.24) in the Appendix). Therefore, the bivariate normal distribution has elliptical contours. See Fig. 4.2. The orientation and narrowness of these elliptical contours are governed by σ_X, σ_Y, and ρ. If $\sigma_X = \sigma_Y$ the axes of the ellipse are on the lines $y = x$ and $y = -x$; as ρ increases toward 1 (or decreases toward -1) the ellipse becomes more tightly concentrated around $y = x$ (or $y = -x$); and when $\rho = 0$ the contours become circles. If, instead, $\sigma_X \neq \sigma_Y$ the axes of the ellipse rotate to $y = \frac{\sigma_Y}{\sigma_X}x$ and $y = -\frac{\sigma_X}{\sigma_Y}x$.

We have assumed here that $\sigma_X > 0$, $\sigma_Y > 0$, and $-1 < \rho < 1$, which corresponds to "positive definiteness" of the quadratic, a point we return to in Section 4.3.

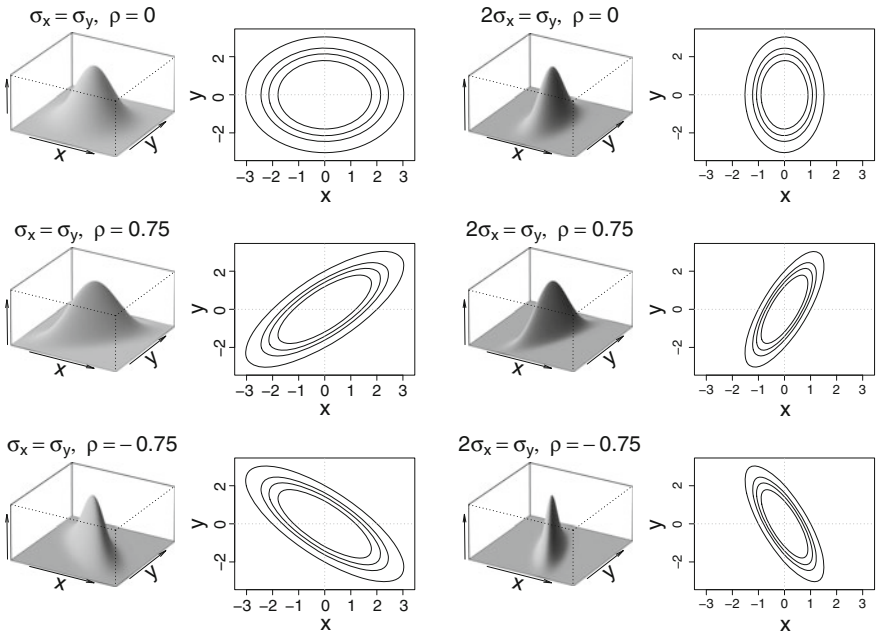

Fig. 4.2 The bivariate normal pdf. Perspective plots and contour plots are shown for various values of σ_X, σ_Y and ρ, with $(\mu_X, \mu_Y) = (0, 0)$. Left column has $\sigma_X = \sigma_Y$ and right column has $2\sigma_X = \sigma_Y$. First, second, and third rows correspond to $\rho = 0$, $\rho = .75$, $\rho = -.75$. Contours enclose probability equal to .8, .9, .95, and .99.

Sometimes a more general definition of bivariate normality is needed: we say that (X, Y) is bivariate normal if every nonzero linear combination of X and Y has a normal distribution, i.e., for all numbers a and b that are not both zero, $aX + bY$ is normally distributed. This covers additional cases, such as when $\rho = 1$, and we mention it again in Chapter 5 when we discuss the general multivariate normal distribution. An important point is that joint normality is a stronger requirement than normality of the individual components. It is not hard to construct a counterexample in which X and Y are both normally distributed but their joint distribution is not bivariate normal.

> *A detail:* Let U and V be independent $N(0, 1)$ random variables. Let $Y = V$ and for $U < 0, V > 0$ or $U > 0, V < 0$ take $X = -U$. This amounts to taking the probability assigned to (U, V) in the 2nd and 4th quadrants and moving it, respectively, to the 1st and 3rd quadrants. The distribution of (X, Y) is then concentrated in the 1st and 3rd quadrants ((X, Y) has zero probability of being in the 2nd or 4th quadrants), yet X and Y remain distributed as $N(0, 1)$. □

In practice, when we examine data $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ to see whether their variation appears roughly to follow a bivariate normal distribution, the general result

suggests one should plot them together as scatterplot pairs $(x_1, y_1), \ldots, (x_n, y_n)$, rather than simply examining $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ separately. In the multivariate case, however, one must rely on one-dimensional and two-dimensional visual representations of data, as in Fig. 4.1.

4.2.3 *Conditional probabilities involving random variables are obtained from conditional densities.*

We previously defined the probability of one event conditionally on another, which we wrote $P(A|B)$, as the ratio $P(A \cap B)/P(B)$, assuming $P(B) > 0$. When we have a pair of random variables X and Y with $f(y) > 0$, the conditional density of X given $Y = y$ is

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}. \tag{4.15}$$

For discrete random variables $f_{X|Y}(x|y)$ is the probability that $X = x$ given that $Y = y$. For continuous random variables, roughly speaking, $f(x, y)dxdy$ is the probability that X will lie in the infinitesimal interval $(x, x+dx)$ and Y will lie in the infinitesimal interval $(y, y+dy)$. We may thus think of $f_{X|Y}(x|y)dx$ as the probability that X will lie in the infinitesimal interval $(x, x + dx)$ given that Y lies in the infinitesimal interval $(y, y + dy)$.

Suppose we

1. Draw a random variable $Y^{(1)}$ from the marginal distribution with pdf $f_Y(y)$, and then
2. Draw another random variable $X^{(1)}$ from the distribution with pdf $f_{X|Y}(x|y)$.

In the discrete case we have

$$\begin{aligned} P(X^{(1)} = x, Y^{(1)} = y) &= P(X^{(1)} = x|Y^{(1)} = y)P(X^{(1)} = x) \\ &= P(X = x|Y = y)P(X = x) \\ &= P(X = x, Y = y). \end{aligned}$$

In other words, this two-step procedure produces a bivariate random vector $(X^{(1)}, Y^{(1)})$ having the joint distribution with pdf $f(x, y)$, which provides a very important intuition: a joint distribution may be considered to arise from a compound process of first drawing a random variable from one marginal distribution, and then drawing a second random variable from the resulting conditional distribution. The interpretation also holds in the continuous case, and the argument is analogous.

Note that when X and Y are independent we have

$$f_{X|Y}(x|y) = f_X(x).$$

This is immediate from (4.15).

Illustration: Spike count pairs (continued from p. 79) We return to the joint distribution of spike counts for two neurons. We may calculate the conditional distribution of X given $Y = 0$. We have $f_{X|Y}(0|0) = .30/.50 = .60$, $f_{X|Y}(1|0) = .15/.50 = .30$, $f_{X|Y}(2|0) = .05/.50 = .10$. Note that these probabilities are different than the marginal probabilities .39, .38, .23. In fact, if $Y = 0$ it becomes more likely that X will also be 0, and less likely that X will be 1 or 2. □

4.2.4 The conditional expectation $E(Y|X = x)$ is called the regression of Y on X.

The conditional expectation of $Y|X$ is

$$E(Y|X = x) = \int y f_{Y|X}(y|x)dy \tag{4.16}$$

where the integral is taken over the range of y.

Illustration: Spike count pairs (continued) Using the joint pdf table repeated on p. 79), we compute $E(X|Y = 0)$. This uses (4.16) except that the roles of X and Y are reversed and the integral is replaced by a sum. We previously found $f_{X|Y}(0|0) = .60$, $f_{X|Y}(1|0) = .30$, $f_{X|Y}(2|0) = .10$. Then

$$E(X|Y = 0) = 0(.6) + 1(.3) + 2(.1) = .5.$$

□

Note that $E(Y|X = x)$ is a function of x, so we might write $M(x) = E(Y|X = x)$ and thus $M(X) = E(Y|X)$ is a random variable. An important result concerning $M(X)$ is often called the law of total expectation.

Theorem: Law of total expectation. Suppose X and Y are random variables and Y has finite expectation. Then we have

$$E(E(Y|X)) = E(Y).$$

Proof: From the definition we compute

$$\begin{aligned} E(E(Y|X = x)) &= \int \left(\int y f_{Y|X}(y|x)dy \right) f_X(x)dx \\ &= \int \int y f_{Y|X}(y|x) f_X(x)dxdy \end{aligned}$$

$$= \int \int y f_{(X,Y)}(x, y) dx dy$$
$$= \int y f_Y(y) dy = E(Y). \qquad \square$$

There are also the closely-related law of total probability and law of total variance.

Theorem: Law of total probability. Suppose X and Y are random variables. Then we have

$$E(P(Y \leq y|X)) = F_Y(y).$$

Proof: The proof follows a series of steps similar to those in the proof of the law of total expectation. $\square$

We may also define the conditional variance of $Y|X$

$$V(Y|X = x) = \int (y - E(Y|X = x))^2 f_{Y|X}(y|x) dy$$

and then get the following, which has important applications.

Theorem: Law of total variance. Suppose X and Y are random variables and Y has finite variance. Then we have

$$V(Y) = V(E(Y|X)) + E(V(Y|X)). \tag{4.17}$$

Proof: The proof is similar to that of the law of total expectation. $\square$

Example 4.2 Decision-making and trial-to-trial variability of spike counts from LIP neurons When an experiment is run repeatedly across many experimental trials, as in Example 1.1, the spiking pattern will vary across trials, as is evident in Fig. 1.1. It is convenient to consider separately the variation in overall rate of firing across trials, which would operate on slow time scales on the order of the length of the trial, from the much faster variation of spike occurrences within trials. To do this we may introduce a random variable Y_i to represent the spike count on trial i and then consider its expectation as it varies across trials, which we write $X_i = E(Y_i)$. The random variable X_i represents the theoretical expectation of the spike count that strips away the variability of spike occurrences within trials but retains the variation across trials, on the slower trial-length time scale. Churchland et al. (2011) used this idea, and applied formula (4.17) to neural spike counts recorded from lateral intraparietal (LIP) cortex during a decision-making task. They argued that the results they obtained for their estimates of $V(E(Y_i|X_i))$ were consistent with a particular model of decision-making but not with competing models. $\square$

In the spike count pairs illustration, we computed the conditional expectation $E(X|Y = y)$ for a single value of y. We could evaluate it for each possible value

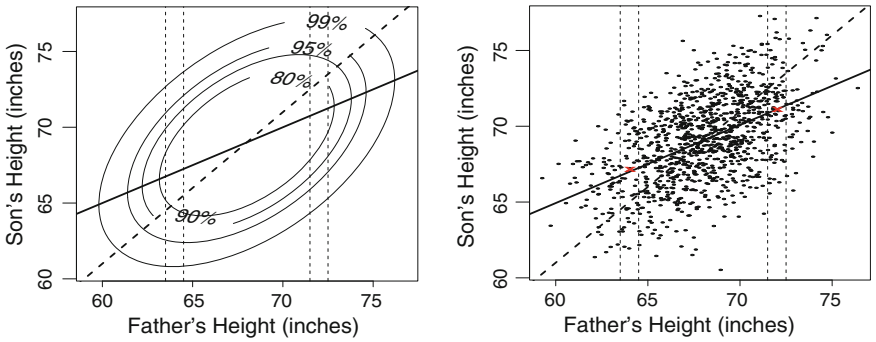

Fig. 4.3 Conditional expectation for bivariate normal data mimicking Pearson and Lee's data on heights of fathers and sons. *Left panel* shows contours of the bivariate normal distribution based on the means, standard deviations, and correlation in Pearson and Lee's data. The *dashed vertical lines* indicate the averaging process used in computing the conditional expectation when $X = 64$ or $X = 72$ inches: to get the average height of a son when the father has height $X = x$ we average y using the probability $f_{Y|X}(y|x)$, which is the probability, roughly, in between the *dashed vertical lines*, integrating across y. In the *right panel* we generated a sample of 1,078 points (the sample size in Pearson and Lee's data set) from the bivariate normal distribution pictured in the *left panel*. We then, again, illustrate the averaging process: when we average the values of y within the *dashed vertical lines* we obtain the two values indicated by a plotted *red x*. These fall very close to the least-squares line (*the solid line*). The *dashed diagonal line* is discussed in the text.

of y. When we consider $E(X|Y = y)$ as a function of y, this function is called the *regression* of X on Y. Similarly, the function $E(Y|X = x)$ is called the regression of Y on X. To understand this terminology, and the interpretation of the conditional expectation, consider the case in which (X, Y) is bivariate normal.

Example 4.3 Regression of son's height on father's height A famous data set, from Pearson and Lee (1903), has been used frequently as an example of regression (See Freedman et al. (2007).) Figure 4.3 displays both a bivariate normal pdf and a set of data generated from the bivariate normal pdf—the latter are similar to the data obtained by Pearson and Lee (who did not report the data, but only summaries of them). For a bivariate normal pair (X, Y), the left panel of Fig. 4.3 shows $E(Y|X = x)$, which is the regression line. The right panel shows a line fitted to the data by least squares, which was discussed briefly in Chapter 1 and will be discussed more extensively in Chapter 12. In a large sample like this one, the least-squares line (right panel) is close to the bivariate normal regression line (left panel). The purpose of showing both is to help clarify the averaging process represented by the conditional expectation $E(Y|X = x)$.

The terminology "regression" is illustrated in Figure 4.3 by the slope of the regression line being less than that of the dashed line. Here, $\sigma_Y = \sigma_X$, because the variation in sons' heights and fathers' heights was about the same, while $(\mu_X, \mu_Y) = (68, 69)$, so that the average height of the sons was about an inch more than the average height among their fathers. The dashed line has slope $\sigma_Y/\sigma_X = 1$ and it goes through the point (μ_X, μ_Y). Thus, the points falling on the dashed line in the left panel, for

example, would be those for which a theoretical son's height was exactly 1 inch more than his theoretical father. Similarly, in the plot on the left, any data points falling on the dashed line would correspond to a real son-father pair for which the son was an inch taller than the father. However, if we look at $E(Y|X = 72)$ we see that among these taller fathers, their son's height tends, on average, to be less than the 1 inch more than the father's predicted by the dashed line. In other words, if a father is 3 inches taller than average, his son will likely be *less than* 3 inches taller than average. This is the tendency for the son's height to "regress toward the mean." An explanation of the phenomenon is as follows. First, the father is tall partly for genetic reasons and partly due to environmental factors which pushed him to be taller. If we represent the effect due to the environmental factors as a random variable U, and assume its distribution follows a bell-shaped curve centered at 0, then for any positive u we have $P(U < u) > 1/2$. Thus, if u represents the effect due to environmental factors that the father received and U the effect that the son receives, the son's environmental effect will tend to be smaller than the father's whenever the father's effect is above average. For a tall father, while the son will inherit the father's genetic component, his positive push toward being tall from the environmental factors will tend to be somewhat smaller than his father's had been. This is *regression toward the mean.* The same tendency, now in the reverse direction, is apparent when the father's height is $X = 64$. Regression to the mean is a ubiquitous phenomenon found whenever two variables vary together. □

In general, the regression $E(Y|X = x)$ could be a nonlinear function of x but in Fig. 4.3 it is a straight line. This is not an accident: if (X, Y) is bivariate normal, the regression of Y on X is linear with slope $\rho \cdot \sigma_Y/\sigma_X$. Specifically,

$$E(Y|X = x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)\,. \tag{4.18}$$

We say that Y has a regression on X with regression coefficient

$$\beta_{Y|X} = \rho\frac{\sigma_Y}{\sigma_X}. \tag{4.19}$$

This means that when $X = x$, the *average* value of Y is given by (4.18). We should emphasize, again, that we are talking about random variables, which are *theoretical* quantities, as opposed to observed data. In data-analytic contexts the word "regression" almost always refers to least-squares regression, illustrated in the right panel of Fig. 4.3.

For later use let us note that when (X, Y) is bivariate normal we may also consider the regression of X on Y

$$E(Y|X = x) = \mu_Y + \beta_{X|Y}(y - \mu_Y)$$

where, as in (4.19),

$$\beta_{X|Y} = \rho\frac{\sigma_X}{\sigma_Y} \tag{4.20}$$

so that if we combine (4.19) and (4.20) we get the following expression for the correlation:

$$\rho = \text{sign}(\beta_{Y|X})\sqrt{\beta_{Y|X}\beta_{X|Y}} \tag{4.21}$$

where $\text{sign}(\beta_{Y|X})$ is -1 if $\beta_{Y|X}$ is negative and 1 if $\beta_{Y|X}$ is positive.

Compare Eq. (4.18) to Eqs. (4.9) and (4.10). From(4.9) and (4.10) we have that the best linear predictor of Y based on X is $f(X)$ where

$$f(x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)\,. \tag{4.22}$$

In general, we may call this the *linear regression* of Y on X. In the case of bivariate normality, the regression of Y on X is equal to the *linear* regression of Y on X, i.e., the regression is linear. We derived (4.22) as the best linear predictor of Y based on X by minimizing mean squared error. More generally, if we write the regression function as $M(x) = E(Y|X = x)$. Then $M(X)$ is the best predictor of Y in the sense of minimizing mean squared error.

Prediction Theorem The function $f(x)$ that minimizes $E((Y - f(X))^2)$ is the conditional expectation $f(x) = M(x) = E(Y|X = x)$.

Proof details: Note that $E(Y - M(X)) = E(Y) - E(E(Y|X))$ and by the law of total expectation (p. 85) this is zero. Now write $Y - f(X) = (Y - M(X)) + (M(X) - f(X))$ and expand $E((Y - f(X))^2)$ to get

$$\begin{aligned} E((Y - f(X))^2) = E((Y - M(X))^2) + 2E((Y - M(X)) \\ (M(X) - f(X))) + E((M(X) - f(X))^2). \end{aligned} \tag{4.23}$$

Applying the law of total expectation to the second term we get

$$\begin{aligned} E((Y - M(X))(M(X) - f(X))) = E(E((Y - M(X))(M(X) \\ - f(X))|X))) \end{aligned}$$

but for every x we have

$$\begin{aligned} E((Y - M(X))(M(X) - f(X))|X = x) = (M(x) - M(x))(M(x) \\ - f(x)) = 0 \end{aligned}$$

so that the second term in (4.23) is 0. The third term $E((M(X) - f(X))^2)$ is always non-negative and it is zero when $f(x)$ is chosen

to equal $M(x)$. Therefore the whole expression is minimized when $f(x) = M(x)$. □

Let us also note the following, which may be viewed as a special case of the prediction theorem.

Theorem: Optimality of the Mean In predicting a random variable X, the number d that minimizes $E((X-d)^2)$ is the mean $d = E(X)$.

Proof: As in the proof of the prediction theorem, we expand the expectation to get $E((X-d)^2) = E(X^2) - 2d\mu + d^2$ where $\mu = E(X)$. The derivative of the expression $d^2 - 2d\mu$ is $2(d-\mu)$, so the expectation is minimized when $d = \mu$. □

4.3 Multivariate Dependence

4.3.1 The mean of a random vector is a vector and its variance is a matrix.

Now suppose we wish to consider the way m random variables $X_1, \ldots, X_m$ vary together. If we have $\mu_i = E(X_i)$, $\sigma_i^2 = V(X_i)$, and $\rho_{ij} = Cor(X_i, X_j)$, for $i = 1, \ldots, m$ and $j = 1, \ldots, m$, we may collect the variables in an m-dimensional *random vector* $X = (X_1, \ldots, X_m)^T$, and can likewise collect the means in a vector

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix}.$$

Similarly, we can collect the variances and covariances in a matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1m}\sigma_1\sigma_m \\ \rho_{21}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{2m}\sigma_1\sigma_m \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{m1}\sigma_1\sigma_m & \rho_{m2}\sigma_2\sigma_m & \cdots & \sigma_m^2 \end{pmatrix}.$$

Note that $\rho_{ij} = \rho_{ji}$ so that Σ is a symmetric matrix (the element in its ith row and jth column is equal to the element in its jth row and ith column, for every i and j). We write the *mean vector* $E(X) = \mu$ and the *variance matrix* $V(Y) = \Sigma$. The latter is also called the *covariance matrix.* Once again we wish to distinguish these from sample-based analogues. If we have m batches of numbers their collective *sample mean vector* is the vector of the m sample means, and their *sample variance matrix* is the matrix S having the form of Σ, above, but with each theoretical standard deviation σ_i being replaced by a corresponding sample standard deviation s_i, and each theoretical correlation ρ_{ij} replaced by a sample correlation $\hat{\rho}_{ij}$, i.e.,

$$S = \begin{pmatrix} s_1^2 & \hat{\rho}_{12} s_1 s_2 & \cdots & \hat{\rho}_{1m} s_1 s_m \\ \hat{\rho}_{21} s_1 s_2 & s_2^2 & \cdots & \hat{\rho}_{2m} s_1 s_m \\ \cdots & \cdots & \cdots & \cdots \\ \hat{\rho}_{m1} s_1 s_m & \hat{\rho}_{m2} s_2 s_m & \cdots & s_m^2 \end{pmatrix}. \tag{4.24}$$

Let w be an m-dimensional vector. By straightforward matrix manipulations we obtain the mean and variance of $w^T X$ as

$$E(w^T X) = w^T \mu \tag{4.25}$$
$$V(w^T X) = w^T \Sigma w. \tag{4.26}$$

Equations (4.25) and (4.26) generalize (4.1) and (4.6).

Let us now recall (see the Appendix, p. 617) that a symmetric $m \times m$ matrix A is positive semi-definite if for every m-dimensional vector v we have $v^T A v \geq 0$ and it is positive definite if for every nonzero m-dimensional vector v we have $v^T A v > 0$. From the definition of variance (involving the integral of a non-negative function), every variance is non-negative. Therefore, $V(w^T X) \geq 0$ so that the variance matrix Σ is necessarily positive semi-definite. However, a variance matrix may or may not be positive definite. The non-positive-definite case is the generalization of $\sigma_X = 0$ for a random variable X: in the non-positive-definite case the distribution of the random vector X "lives" on a subspace that has dimensionality less than m. For example, if X and Y are both normally distributed but $Y = X$ then their joint distribution "lives" on a one-dimensional subspace $y = x$ of the two-dimensional plane.

An important tool in analyzing a variance matrix is the spectral decomposition. As stated in Section A.8 of the Appendix (see p. 617), the spectral decomposition of a positive semi-definite matrix A is $A = PDP^T$ where D is a diagonal matrix with diagonal elements $\lambda_i = D_{ii}$ for $i = 1, \ldots, m$, and P is an orthogonal matrix, i.e., $P^T P = I$, where I is the m-dimensional identity matrix. Here, $\lambda_1, \ldots, \lambda_m$ are the eigenvalues of A and the columns of P are the corresponding eigenvectors.

Lemma If Σ is a symmetric positive definite matrix then there is a symmetric positive definite matrix $\Sigma^{\frac{1}{2}}$ such that

$$\Sigma = \Sigma^{\frac{1}{2}} \Sigma^{\frac{1}{2}}$$

and, furthermore, writing its inverse matrix as $\Sigma^{-\frac{1}{2}} = (\Sigma^{\frac{1}{2}})^{-1}$ we have

$$\Sigma^{-1} = \Sigma^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}}.$$

Proof: This follows from the spectral decomposition (Section A.8), which gives $\Sigma = PDP^T$, with D being diagonal. We define $D^{\frac{1}{2}}$ to be the diagonal matrix having elements $(\sqrt{D_{11}}, \ldots, \sqrt{D_{mm}})$ and take $\Sigma^{\frac{1}{2}} = PD^{\frac{1}{2}}P^T$. With $\Sigma^{-\frac{1}{2}} = PD^{-\frac{1}{2}}P^T$,

where $D^{-\frac{1}{2}}$ is the diagonal matrix having elements $(1/\sqrt{D_{11}}, \ldots, 1/\sqrt{D_{mm}})$, the stated results are easily checked. □

Theorem Suppose X is a random vector with mean μ and covariance matrix Σ. Define the random vector $Y = \Sigma^{-1/2}(X - \mu)$. Then $E(Y)$ is the zero vector and $V(Y)$ is the m-dimensional identity matrix.

Proof: This follows from the lemma. We omit the details. □

We will use this kind of standardization of a random vector in Chapter 6.

4.3.2 The dependence of two random vectors may be quantified by mutual information.

It often happens that the deviation of one distribution from another must be evaluated. Consider two continuous pdfs $f(x)$ and $g(x)$, both being positive on (A, B). The *Kullback-Leibler (KL) divergence* is the quantity

$$D_{KL}(f, g) = E_f\left(\log \frac{f(X)}{g(X)}\right)$$

where the subscript on the expectation E_f signifies that the random variable X has pdf $f(x)$. In other words, we have

$$D_{KL}(f, g) = \int_A^B f(x) \log \frac{f(x)}{g(x)} dx.$$

The KL divergence may also be defined, analogously, for discrete distributions. Note that $D_{KL}(f, g)$ may also be written in the difference form

$$D_{KL}(f, g) = E_f\left(\log f(X)\right) - E_f\left(\log g(X)\right) \tag{4.27}$$

and that, except for some special cases, $D(f, g) \neq D(g, f)$. In fact, the KL divergence is essentially unique (aside from linear rescaling) among all discrepancies $D(f, g)$ that satisfy

(i) $D(f, g) = E_f(\varphi(f(X))) - E_f(\varphi(g(X)))$ for some differentiable function φ, and
(ii) $D(f, g)$ is minimized over g by $g = f$.

Details: When there are finitely many outcomes (so that sums replace integrals in the definition of $D_{KL}(f, g)$) it may be shown that the form of φ must be logarithmic, i.e., φ must satisfy $\varphi(f(x)) = a + b \log f(x)$ for some a, b, with $b > 0$. See Konishi and Kitagawa (2007, [Section 3.1]) . □

In addition to having the special difference-of-averages property in (4.27), the KL divergence takes a simple and intuitive form when applied to normal distributions.

Illustration: Two normal distributions Suppose $f(x)$ and $g(x)$ are the $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ pdfs. Then, from the formula for the normal pdf we have

$$\log \frac{f(x)}{g(x)} = -\frac{(x-\mu_1)^2 - (x-\mu_2)^2}{2\sigma^2} = \frac{2x(\mu_1-\mu_2) - (\mu_1^2 - \mu_2^2)}{2\sigma^2}$$

and substituting X for x and taking the expectation (using $E_X(X) = \mu_1$), we get

$$D_{KL}(f, g) = \frac{2\mu_1^2 - 2\mu_1\mu_2 - \mu_1^2 + \mu_2^2}{2\sigma^2} = \left(\frac{\mu_1 - \mu_2}{\sigma}\right)^2.$$

That is, $D_{KL}(f, g)$ is simply the squared standardized difference between the means. This is a highly intuitive notion of how far apart these two normal distributions are. □

Example 4.4 Auditory-dependent vocal recovery in zebra finches Song learning among zebra finches has been heavily studied. When microlesions are made in the HVC region of an adult finch brain, songs become destabilized but the bird will recover its song within about 1 week. Thompson et al. (2007) ablated the output nucleus (LMAN) of the anterior forebrain pathway of zebra finches in order to investigate its role in song recovery. They recorded songs before and after the surgery. The multiple bouts of songs, across 24 h (hours), were represented as individual notes having a particular frequency composition and duration. The distribution of these notes post-surgery was then compared to the distribution pre-surgery. In one of their analyses, for instance, the authors examined the distributions of pitch and duration. Their method of comparing post-surgery and pre-surgery distributions was to compute the KL divergence. Thompson et al. found that deafening following song disruption produced a large KL divergence whereas LMAN ablation did not. This indicated that the anterior forebrain pathway is not the neural locus of the learning mechanism that uses auditory feedback to guide song recovery. □

The Kullback-Leibler divergence may be used to evaluate the association of two random vectors X and Y. We define the *mutual information* of X and Y as

$$I(X, Y) = D_{KL}(f_{(X,Y)}, f_X f_Y) = E_{(X,Y)} \log \frac{f_{(X,Y)}(X, Y)}{f_X(X) f_Y(Y)}. \tag{4.28}$$

In other words, the mutual information between X and Y is the Kullback-Leibler divergence between their joint distribution and the distribution they would have if they were independent. In this sense, the mutual information measures how far a joint distribution is from independence.

Illustration: Bivariate normal If X and Y are bivariate normal with correlation ρ some calculation following application of the definition of mutual information gives

$$I(X, Y) = -\frac{1}{2}\log(1 - \rho^2). \tag{4.29}$$

Thus, when X and Y are independent, $I(X, Y) = 0$ and as they become highly correlated (or negatively correlated) $I(X, Y)$ increases indefinitely. □

Theorem For random variables X and Y that are either discrete or jointly continuous having a positive joint pdf, mutual information satisfies (i) $I(X, Y) = I(Y, X)$, (ii) $I(X, Y) \geq 0$, (iii) $I(X, Y) = 0$ if and only if X and Y are independent, and (iv) for any one-to-one continuous transformations $f(x)$ and $g(y)$, $I(X, Y) = I(f(X), g(Y))$.

Proof: Omitted. See, e.g, Cover and Thomas (1991). □

Property (iv) makes mutual information quite different from correlation. For correlation, $Cor(X, Y)^2 = Cor(f(X), g(Y))^2$ when $f(x)$ and $g(y)$ are linear functions, but when they are nonlinear the value of the squared correlation can change.

The use here of the word "information" is important. For emphasis we say, in somewhat imprecise terms, what we think is meant by this word.

> Roughly speaking, information about a random variable Y is associated with the random variable X if the uncertainty in Y is larger than the uncertainty in $Y|X$.

For example, we might interpret "uncertainty" in terms of variance. If the regression of Y on X is linear, as in (4.18) (which it is if (X, Y) is bivariate normal), we have

$$\sigma^2_{Y|X} = (1 - \rho^2)\sigma^2_Y. \tag{4.30}$$

In this case, information about Y is associated with X whenever $|\rho| > 0$ so that $1 - \rho^2 < 1$. The reduction of uncertainty in Y provided by X becomes

$$\sigma^2_Y - \sigma^2_{Y|X} = \rho^2\sigma^2_Y,$$

which retains the multiplier σ^2_Y (coming from the multiplicative form of (4.31)). To remove the factor σ^2_Y we may consider the relative reduction of uncertainty,

$$\frac{\sigma^2_Y - \sigma^2_{Y|X}}{\sigma^2_Y} = \rho^2.$$

In this sense, ρ^2 becomes a measure of the information about Y supplied by X.

A different rewriting of (4.30) will help us connect it more strongly with mutual information. First, if we redefine "uncertainty" to be standard deviation rather than variance, (4.30) becomes

$$\sigma_{Y|X} = \sqrt{1 - \rho^2}\sigma_Y. \tag{4.31}$$

Like Equation (4.30), (4.31) describes a multiplicative (proportional) decrease in uncertainty in Y associated with X. An alternative is to redefine "uncertainty," and rewrite (4.31) in an *additive* form, so that the uncertainty in $Y|X$ is obtained by *subtracting* an appropriate quantity from the uncertainty in Y. To obtain an additive form we define "uncertainty" as the log standard deviation. Assuming $|\rho| < 1$, $\log\sqrt{1-\rho^2}$ is negative and, using $\log\sqrt{1-\rho^2} = \frac{1}{2}\log(1-\rho^2)$, we get

$$\log\sigma_{Y|X} = \log\sigma_Y - \left(-\frac{1}{2}\log(1-\rho^2)\right). \tag{4.32}$$

In words, Eq. (4.32) says that $-\frac{1}{2}\log(1-\rho^2)$ is the amount of information associated with X in reducing the uncertainty in Y to that of $Y|X$. If (X, Y) is bivariate normal then, according to (4.29), this amount of information associated with X is the mutual information.

Formula (4.32) may be generalized by quantifying "uncertainty" in terms of *entropy*, which leads to a popular interpretation of mutual information.

Details: We say that the *entropy* of a discrete random variable X is

$$H(X) = -\sum_x f_X(x)\log f_X(x) \tag{4.33}$$

We may also call this the entropy of the distribution of X. In the continuous case the sum is replaced by an integral (though there it is defined only up to a multiplicative constant, and is often called *differential entropy*). The entropy of a distribution was formalized analogously to Gibbs entropy in statistical mechanics by Claude Shannon in his development of communication theory. As in statistical mechanics, the entropy may be considered a measure of disorder in a distribution. For example, the distribution over a set of values $\{x_1, x_2, \ldots, x_m\}$ having maximal entropy is the uniform distribution (giving equal probability $\frac{1}{m}$ to each value) and, roughly speaking, as a distribution becomes concentrated near a point its entropy decreases.
For ease of interpretation the base of the logarithm is often taken to be 2 so that, in the discrete case,

$$H(X) = -\sum_x f_X(x)\log_2 f_X(x). \tag{4.34}$$

Suppose there are finitely many possible values of X, say $x_1, \ldots, x_m$, and someone picks one of these values with probabilities given by $f(x_i)$, then we try to guess which value has been picked by asking "yes" or "no" questions (e.g., "Is it greater than x_3?"). In this case the entropy (using $\log_2$, as above) may be interpreted as the minimum average number of yes/no questions that must be asked in order to

determine the number, the average being taken over replications of the game. When the outcomes $x_1, \ldots, x_m$ are equally likely we have $f(x_i) = 1/m$, for $i = 1, \ldots, m$, and (4.34) reduces to $H(X) = \log_2(m)$. Entropy may be used to characterize many important probability distributions. The distribution on the set of integers $0, 1, 2, \ldots, n$ that maximizes entropy subject to having mean μ is the binomial. The distribution on the set of all non-negative integers that maximizes entropy subject to having mean μ is the Poisson. In the continuous case, the distribution on the interval $(0, 1)$ having maximal entropy is the uniform distribution. The distribution on the positive real line that maximizes entropy subject to having mean μ is the exponential. The distribution on the positive real line that maximizes entropy subject to having mean μ and variance σ^2 is the gamma. The distribution on the whole real line that maximizes entropy subject to having mean μ and variance σ^2 is the normal.

Now, if Y is another discrete random variable then the entropy in the conditional distribution of $Y|X = x$ may be written

$$H(Y|X = x) = -\sum_y f_{Y|X}(y|x) \log f_{Y|X}(y|x)$$

and if we average this quantity over X, by taking its expectation with respect to $f_X(x)$, we get what is called the *conditional entropy* of Y given X:

$$H(Y|X) = \sum_x \left(-\sum_y f_{Y|X}(y|x) \log f_{Y|X}(y|x) \right) f_X(x).$$

Algebraic manipulation then shows that the mutual information may be written

$$I(X, Y) = H(Y) - H(Y|X).$$

This says that the mutual information is the average amount (over X) by which the entropy of Y decreases given the additional information $X = x$. In the discrete case, working directly from the definition we find that entropy is always non-negative and, furthermore, $H(Y|Y) = 0$. The expression for the mutual information, above, therefore also shows that in the discrete case $I(Y, Y) = H(Y)$. (In the continuous case we get $I(Y, Y) = \infty$.) For an extensive discussion of entropy, mutual information, and communication theory see Cover and Thomas (1991) or MacKay (2003).

Mutual information was used to define the *channel capacity* of a communication system that transmits a signal in the presence of noise: if X is a random variable representing a transmitted message and Y is a random variable representing the

received message after noise has been injected during the transmission process, then the channel capacity is

$$C = \max_X I(X, Y)$$

where the maximum is taken over all possible distributions of X. This concept, developed to characterize electronic communication channels, has also been applied to human behavior and neural activity. Because the mutual information in this context concerns discrete distributions for (X, Y), and $\log_2$ is used, the units are said to be in *bits* for "binary digits" (because, for a positive integer n, $\log_2(n)$ is the number of binary digits used to represent n in base 2). Thus, human and neural information processing capacity is usually reported in bits.

Example 4.5 The Magical Number Seven In a famous paper, George Miller reviewed several psychophysical studies that attempted to characterize the capacity of humans to process sensory input signals (Miller 1956). One study, for example, exposed subjects to audible tones of several different values of pitch (frequency) and asked them to identify the pitch (e.g., pitch 1, 2, or 3, corresponding to high, medium, or low). The question was, how many distinct values of pitch can humans reliably discriminate? It turned out that with five or more tones of different pitch, the human observers made frequent mistakes. The experimental design allowed calculation of the probability of responding with a particular answer Y based on a particular input tone X, and with this the mutual information could be calculated. By examining several different studies, of similar yet different types, Miller concluded that mutual information had an asymptotic maximum at about $C = 2.6 \pm .6$ bits, which could be interpreted as the channel capacity of a human observer. Transforming this back to numbers of discernible categories gives $2^{2.6-.6} = 4$ and $2^{2.6+.6} = 9.2$. After looking at other, related psychophysical data Miller summarized by saying there was a "magical number seven, plus or minus two," which characterized many aspects of human information processing in terms of channel capacity. □

Mutual information has also been used extensively to quantify the information about a stochastic stimulus Y associated with a neural response X. In that context the notation is often changed by setting $S = Y$ for "stimulus" and $R = X$ for neural "response," and the idea is to determine the amount of information about the stimulus that is associated with the neural response.

Example 4.6 Temporal coding in inferotemporal cortex In an influential paper, Optican and Richmond (1987) reported responses of single neurons in inferotemporal (IT) cortex of monkeys while the subjects were shown various checkerboard-style grating patterns as visual stimuli. Optican and Richmond computed the mutual information between the 64 randomly-chosen stimuli (the random variable Y here taking 64 equally-likely values) and the neural response (X), represented by a vector of time-varying firing rates across multiple time bins. They compared this with the mutual information between the stimuli and a single firing rate across a large time interval and concluded that there was considerably more mutual information in the time-varying firing rate vector. Put differently, more information about the stimulus was carried by the time-varying firing rate vector than by the overall spike count. □

In Examples 4.4 and 4.6 the calculations were based on pdfs that were *estimated* from the data. We discuss probability *density estimation* in Chapter 15.

4.3.3 Bayes' theorem for random vectors is analogous to Bayes' theorem for events.

Now suppose X and Y are random vectors with a joint density $f(x, y)$. Substituting $f(x, y) = f_{Y|X}(y|x)f(x)$ into (4.15), we have

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(x, y)}{f_Y(y)} = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)}. \tag{4.35}$$

This is a form of Bayes' Theorem (see Section 3.1.4).

Bayes' Theorem for Random Vectors If X and Y are continuous random vectors and $f_Y(y) > 0$ we have

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{\int f_{Y|X}(y|x)f_X(x)dx}. \tag{4.36}$$

If X is a discrete random vector, and Y is either discrete or continuous with $f_Y(y) > 0$, then we have

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{\sum_x f_{Y|X}(y|x)f_X(x)}. \tag{4.37}$$

Proof: These results follow[2] by using the definition of marginal pdf in the denominator of (4.35). □

The resemblance of this result to Bayes' Theorem for events may be seen by comparing (4.36) with (3.1), identifying X with A and Y with B. The theorem also holds, as a special case, if X and Y are random variables.

4.3.4 Bayes classifiers are optimal.

Suppose X is a random variable (or random vector) that may follow one of two possible distributions having pdf $f_1(x)$ or $f_2(x)$. If $X = x$ is observed, which distribution

[2] The result (4.37) when Y is continuous requires the notion of the joint distribution of (X, Y) when X is discrete and Y is continuous, which we have not discussed, but this case can be accommodated by an extension of the definitions we have given.

did it come from? This is the problem of binary *classification.* Typically, there is a random sample $X_1, \ldots, X_n$ and the problem is to classify (to one of the two distributions) each of the many observations. A *decision rule* or *classification rule* is a mapping that assigns to each possible x a classification (that is, a distribution). A classic scenario for binary classification is when patients having characteristics summarized in a vector x (for example, brain features found from PET imaging), are to be considered diseased (e.g., having Alzheimer-like amyloid deposits, see Vandenbergh et al. 2013) or not. The problem extends to m categories, where X follows one of many alternative distributions, with pdf $f_i(x)$, for $i = 1, \ldots, m$. A classification error is made if $X \sim f_k(x)$ and the observation $X = x$ is classified as coming from $f_i(x)$ with $i \neq k$. In this section we present a remarkable result: it is, in principle, possible to define a classifier that minimizes the probability of classification error.

Let C_i refer to the case $X \sim f_i(x)$. We use the letter C to stand for "class," so that the problem is to assign to each observed x a class C_i. We assume that X is selected from class C_i with probability $P(C = C_i) = \pi_i$, for $i = 1, \ldots, m$. Often the π_i probabilities are taken to be equal, i.e., $\pi_i = 1/m$, for $i = 1, \ldots, m$ (so that the classes are *a priori* equally likely), but the theory does not require this. The *Bayes classifier* assigns to each observed value x the class having the maximal *posterior probability*

$$P(C = C_k | X = x) = \frac{f_k(x)\pi_k}{\sum_{i=1}^{m} f_i(x)\pi_i} \tag{4.38}$$

among all the classes C_i. Writing $f_i(x) = f_{X|C}(x|C = C_i)$, Eq. (4.38) has the same form as (4.37). The following theorem says that Bayes classifiers minimize the probability of classification error.

Theorem on Optimality of Bayes Classifiers Suppose X is drawn from a distribution having pdf $f_i(x)$, where $f_i(x) > 0$ for all x, with probability π_i, for $i = 1, \ldots, m$, where $\pi_1 + \cdots + \pi_m = 1$, and let C_i be the class $X \sim f_i(x)$. Then the probability of committing a classification error is minimized if $X = x$ is classified as arising from the distribution having pdf $f_k(x)$ for which C_k has the maximum posterior probability given by (4.38).

The proof is somewhat lengthy and appears at the end of this section.

Corollary Suppose that with equal probabilities X is drawn either from a distribution having pdf $f_1(x)$, where $f_1(x) > 0$ for all x, or from a distribution having pdf $f_2(x)$, where $f_2(x) > 0$ for all x. Then the probability of committing a classification error is minimized if $X = x$ is classified to the distribution having the higher pdf at x.

Corollary Suppose n observations $X_1, \ldots, X_n$ are drawn, independently, from a distribution having pdf $f_i(x)$, where $f_i(x) > 0$ for all x, with probability π_i, for

$i = 1, \ldots, m$, where $\pi_1 + \cdots + \pi_m = 1$, and let C_i be the class $X \sim f_i(x)$. Then the expected number of misclassifications is minimized if each $X_j = x_j$ is classified as arising from the distribution having pdf $f_k(x_j)$ for which C_k has the maximum posterior probability

$$P(C_k|X_j = x_j) = \frac{f_k(x_j)\pi_k}{\sum_{i=1}^{m} f_i(x_j)\pi_i}$$

among all the classes C_i.

Proof: Let $Y_i = 1$ if X_i is misclassified, and 0 otherwise. The theorem says that $P(Y_i = 1) = P(Y_1 = 1)$ is minimized by the Bayes classifier, which maximizes (4.38). The expected number of misclassifications is then $E(\sum_i Y_i)$ and we have

$$\begin{aligned} E(\sum_i Y_i) &= \sum_i E(Y_i) \\ &= \sum_i P(Y_i = 1) \\ &= nP(Y_1 = 1). \end{aligned}$$

Therefore, the expected number of misclassifications is minimized by the Bayes classifier. □

One use of classifying neural data is to show that information about stimuli or behavior is contained in particular recorded signals. Here is an example.

Example 4.7 Decoding intended movement using MEG We introduced MEG neuroimaging in Example 1.2. One of its attractive features is that it is non-invasive while being potentially capable of supplying movement-related information with high temporal resolution, much like that obtained with highly invasive electrophysiological methods. Wang et al. (2010) studied MEG signals from subjects both during a wrist movement task and during imagined wrist movement. The idea was that there might be substantial information about intended wrist movement even when the wrist was not actually moving—this would be analogous to the situation in which a user was severely disabled. One purpose of this methodology would be to localize the movement-related information in order to help guide surgical implant of a more invasive device.

In the case of wrist movement, each subject had to move a joystick-controlled cursor, which was viewed on a projection of a computer screen. After one of 4 directional targets (up, down, left, right) was illuminated the subject then had to hit the target with the cursor. In the imagined movement case, each subject was told to imagine moving the wrist. For each of the experimental conditions, there were 120 recordings from 87 MEG sensors located above the sensorimotor areas during the movement and imagined movement tasks. A 1,500 ms time window was selected for analysis, and this window was partitioned into 150 time bins (each 10 ms in length, the signals being averaged within time bins), so that the data consisted vectors $X = x$ at each of 150 time points. It was assumed that each X was drawn from one of four

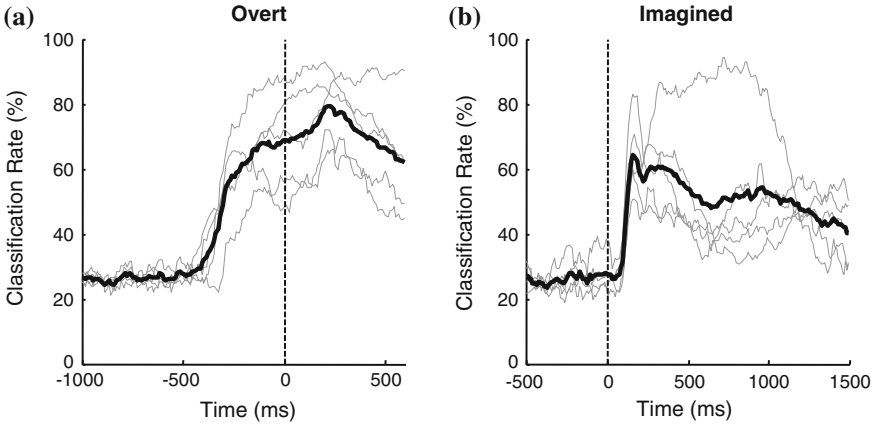

Fig. 4.4 Decoding accuracy from cursor controlled by overt wrist movement (**a**) and imagined wrist movement (**b**). Time $t = 0$ is onset of movement of the cursor. Thin gray lines show decoding accuracy using a Bayes classifier for each of 5 subjects across 150 time intervals, each interval being 10 ms in length. Thick black lines are accuracies averaged across subjects. Adapted from Wang et al. (2010).

multivariate normal distributions (see Chapter 5) having pdf $f_k(x)$, $k = 1, 2, 3, 4$, corresponding to the four experimental conditions (up, down, left, right). That is, for both hand movement and imagined movement, the four experimental conditions were assumed to produce multivariate normal data, but with four distinct sets of mean vectors μ_k and variance matrices Σ_k. A Bayes classifier was then used to try to recover from the data the experimental condition that had generated those data. If the classifier performed above chance levels of 25 % (1 out of 4), this would indicate the presence of directional movement information, or imagined movement information, in the MEG sensors located above the sensorimotor areas (to measure classification accuracy Wang et al. used leave-one-out cross-validation, discussed in Section 12.5.7). The results for 5 subjects are shown in Fig. 4.4. Chance classification accuracy would be 25 %. It may be seen that for every subject, during both movement and imagined movement, the classification accuracy rose sharply above chance. In the imagined movement case (panel b of Fig. 4.4) the peak classification accuracy ranged across subjects from about 50 % to about 90 %, with a mean of over 60 %. □

In our discussion of Example 4.7, above, we have omitted many details. Most importantly, in practice a Bayes classifier must learn (estimate) from some *training data* the distributions $f_i(x)$. In the multivariate normal case this requires estimating the mean vectors μ_i and variance matrices Σ_i, which is usually done by computing the sample mean vector and sample variance matrices defined in (4.24) (see Section 12.5.7 for further discussion of the use of training data and cross-validation). In many multidimensional problems the data are clearly non-normal and it is difficult to estimate $f_i(x)$ reliably. For such situations other, non-Bayesian classifiers are popular (see Section 17.4.2). Nonetheless, Bayes classifiers set the theoretical standard by achieving the smallest possible classification error rate.

The fundamental result given in the theorem also extends to the case in which different penalties result from the various incorrect classifications. This more general situation is treated by *decision theory*. Suppose $d(x)$ is a mapping that assigns to each x a class (a distribution). Such a mapping is called a *decision rule*. Let us denote the possible values of any such rule by $d(x) = a$ (for *action*), so that a may equal any of the C_i for $1, 2, \ldots, m$. The penalties associated with various classifications, or decisions, may be specified by a *loss function* $L(d(x), C_k) = L(a, C_k)$, where each $L(a, C_k)$ is the non-negative number representing the penalty for deciding to classify x as arising from class C_a when actually it arose from class C_k. We then consider the expected loss $E(L(d(X), C_i))$, i.e., the average behavior of the decision rule, which is also known as the *risk* of the decision rule for class C_i, and we may average these risks across classes by weighting them according to their probabilities π_i. The decision rule with the smallest average risk is called the *optimal decision rule*. Assuming that class C_i has probability π_i, for $i = 1, \ldots, m$, this optimal rule turns out to be the *Bayes rule*, which is found by minimizing the expected loss computed from the posterior distribution, i.e., minimizing $E_{C_i|x}(L(a, C_i))$ over possible actions a. The theorem above then becomes the special case in which $L(a, C_i) = 0$ if $a = C_i$ and $L(a, C_i) = 1$ otherwise, for then the risk is simply the probability of misclassification. The process of applying Bayes rules is often called *Bayesian decision-making*.

In many applications of decision theory one speaks not of losses but of gains, and then the loss function is replaced by a *utility function*. Typically one then writes the utility as $U(a, C_i)$ and the Bayes rule would maximize the expected utility based on the posterior distribution of C_i, for $i = 1, \ldots, m$.

Much has been written about the extent to which the nervous system implements Bayesian decision-making. A theoretical Bayesian decision-maker is often called an *ideal observer*. Thus, the issue is the extent to which a particular part of the nervous system performs a computation consistently with the way an ideal observer would use the available information.

Example 4.8 Vision as Bayesian decision-making Geisler (2011) reviews the benefits of using ideal observers to model visual perception (see also Yuille and Kersten 2006). In this case a typical task is to classify an object based on a noisy stimulus that reaches the eye. If there are biological constraints, these are implemented as costs that are incorporated into the loss function. There is prior information about the probability of each object, and for each object there is a probability distribution for the stimulus. These ingredients allow Bayesian decision-making to proceed. In applications there is considerable detail about each aspect of the formalism: the probability distributions for the data, those that represent the prior, and the loss function. The concept, however, is quite simple. □

Some additional references concerning ideal observer analysis, and Bayesian approaches to modeling neural systems more generally, appear at the beginning of Chapter 16. Here is a different setting in which utilities and Bayes rules have been invoked.

Example 4.9 ACT-R theory of procedural memory ACT-R is a theory of human problem-solving that is implemented in a computer program (Anderson 1993, 2007). A typical domain is elementary algebra problem-solving, involving equations such as $7x+3 = 38$. The many steps involved in solving algebra problems include actions such as "subtract," which require calls to memory (e.g., to retrieve $8-3 = 5$). These are encoded as *production rules* which are IF-THEN statements, and are often called *procedures.* At the completion of each step ACT-R must select from memory the next production rule to use. To do so it considers a utility function based on the value V of the goal, the probability P_i of achieving the goal if production rule i is selected, and the cost D_i of rule i. Each production rule is then assigned the utility

$$U_i = P_i V - D_i.$$

ACT-R picks the production rule with the highest utility. Because the probabilities are actually posterior probabilities based on previous experience, ACT-R may be considered to be using a Bayes rule for this situation. The acronym ACT stands for "adaptive character of thought" and the R is tacked on as a nod to "rational" in the sense of optimal decision-making. □

Proof of theorem on optimality of Bayes classifiers:
We consider the binary case where $m = 2$. We also assume the two distributions are discrete and, for simplicity, we take $\pi = \frac{1}{2}$. Here, the Bayes classifier assigns class C_1 to $X = x$ whenever $f_1(x) > f_2(x)$, and assigns class C_2 when $f_2(x) \geq f_1(x)$.
Let $R = \{x : f_1(x) \leq f_2(x)\}$. We want to show that the classification rule assigning $x \to f_2(x)$ whenever $x \in R$ has a smaller probability of error than the classification rule $x \to f_2(x)$ whenever $x \in A$ for any set A that is different than R. To do this we decompose R and its complement R^c as $R = (R \cap A) \cup (R \cap A^c)$ and $R^c = (R^c \cap A) \cup (R^c \cap A^c)$. We have

$$\sum_{x \in R} f_1(x) = \sum_{x \in R \cap A} f_1(x) + \sum_{x \in R \cap A^c} f_1(x) \tag{4.39}$$

and

$$\sum_{x \in R^c} f_2(x) = \sum_{x \in R^c \cap A} f_2(x) + \sum_{x \in R^c \cap A^c} f_2(x). \tag{4.40}$$

By the definition of R we have, for every $x \in R$, $f_1(x) \leq f_2(x)$ and, in particular, for every $x \in R \cap A^c$, $f_1(x) \leq f_2(x)$. Therefore, from (4.39) we have

$$\sum_{x\in R} f_1(x) \le \sum_{x\in R\cap A} f_1(x) + \sum_{x\in R\cap A^c} f_2(x). \tag{4.41}$$

Similarly, from (4.40) we have

$$\sum_{x\in R^c} f_2(x) < \sum_{x\in R^c\cap A} f_1(x) + \sum_{x\in R^c\cap A^c} f_2(x). \tag{4.42}$$

Strict inequality holds in (4.42) because A is distinct from R; if $A = R$ then $R^c \cap A = \emptyset$ and the first sums in both (4.40) and (4.42) become zero. Combining (4.41) and (4.42) we get

$$\sum_{x\in R} f_1(x) + \sum_{x\in R^c} f_2(x) < \sum_{x\in R\cap A} f_1(x) + \sum_{x\in R\cap A^c} f_2(x) + \sum_{x\in R^c\cap A} f_1(x) + \sum_{x\in R^c\cap A^c} f_2(x)$$

and the right-hand side reduces to $\sum_{x\in A} f_1(x) + \sum_{x\in A^c} f_2(x)$. In other words, we have

$$\sum_{x\in R} f_1(x) + \sum_{x\in R^c} f_2(x) < \sum_{x\in A} f_1(x) + \sum_{x\in A^c} f_2(x). \tag{4.43}$$

The left-hand side of (4.43) is the probability of an error using the rule $x \to f_2(x)$ whenever $x \in R$ while the right-hand side of (4.43) is the probability of an error using the rule $x \to f_2(x)$ whenever $x \in A$. Therefore the rule $x \to f_2(x)$ whenever $x \in R$ has the smallest probability of classification error.

The case for general π is essentially the same, and the continuous case replaces sums with integrals. When $m > 2$ the argument is similar. □

Chapter 5
Important Probability Distributions

In Chapter 1 we said that a measurement is determined in part by a "signal" of interest, and in part by unknown factors we may call "noise." Statistical models introduce probability distributions to describe the variation due to noise, and thereby achieve quantitative expressions of knowledge about the signal—a process we will describe more fully in Chapters 7 and 10. The essential ideas in statistical modeling are simple and very general, allowing modern methods to make reasonably realistic assumptions. Despite this wide-ranging generality, the models found in elementary statistics rely heavily on a small handful of probability distributions. For this reason alone, a beginning student must learn about the binomial model for binary observations, the Poisson model for counts, and the normal model for continuous measurements. But there are additional motivations for studying these and several other probability distributions. While it may be tempting to dismiss the ubiquity of these distributions as a historical quirk, a throwback to a pre-computer age in which simplicity was essential, a small number of distributions remain especially important in contemporary practice. This is partly because many methods of statistical inference, when applied carefully, are remarkably robust in the face of modest deviations from theoretical assumptions. In addition, the simplest distributions often serve as a starting point when building more general and elaborate models. Furthermore, these distributions continue to be important because they arise in theoretical calculations. In this chapter we discuss at greater length some of the probability distributions we mentioned in Chapters 3 and 4. We also introduce several others.

5.1 Bernoulli Random Variables and the Binomial Distribution

5.1.1 Bernoulli random variables take values 0 or 1.

A random variable X that takes the value 1 with probability p and 0 with probability $1-p$ is called a *Bernoulli random variable*. For example, patient P.S. in Example 1.4

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_5,

made repeated choices of the "burning" or "non-burning" house. Each such choice could be considered a Bernoulli random variable by coding "burning" as 0 and "non-burning" as 1 (or vice-versa).

5.1.2 *The binomial distribution results from a sum of independent and homogeneous Bernoulli random variables.*

In the case of the binomial distribution arising from two trials for patient P.S., discussed on p. 47, we made two probabilistic assumptions: (i) *independence*, the choices on the two trials were made independently, and (ii) *homogeneity*, the probability of choosing non-burning house remained the same across the two trials. With X being the number of times she chooses the non-burning house, and p being the probability that she chooses the non-burning house on any given trial, these assumptions lead to X having a binomial distribution over the possible values 0, 1, 2 with binary event probability p. We would write this by saying the distribution of X is $B(2, p)$, or $X \sim B(2, p)$.

The binomial distribution is easy to generalize: instead of counting the number of outcomes of a certain type out of a maximal possible value of 2, we allow the maximal value to be any positive integer n; under assumptions of independence and homogeneity we then would say X has distribution $B(n, p)$, or simply $X \sim B(n, p)$. For example, if we were considering 3 trials and again let X be the number of trials on which P.S. chooses the non-burning house, then X has a binomial distribution with $n = 3$ and binary event probability p, or $X \sim B(3, p)$. By a similar argument to that made for $n = 2$ on p. 47 we have $P(X = 3) = p^3$, $P(X = 2) = 3p^2(1 - p)$, $P(X = 1) = 3p(1 - p)^2$, $P(X = 0) = (1 - p)^3$. Similarly, for four trials we would have then $X \sim B(4, p)$ and $P(X = 4) = p^4$, $P(X = 3) = 4p^3(1 - p)$, $P(X = 2) = 6p^2(1 - p)^2$, $P(X = 1) = 4p(1 - p)^3$, $P(X = 0) = (1 - p)^4$.

The general formula for arbitrary n with $X \sim B(n, p)$, is

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{(n-x)} \tag{5.1}$$

for $x = 0, 1, 2, \ldots, n$, where $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ is the number of ways of choosing x objects from n without regard to ordering. Equation (5.1) is the binomial probability mass function (or pdf). If $X \sim B(n, p)$ then straightforward calculations produce

$$\begin{aligned} E(X) &= np \\ V(X) &= np(1 - p) \\ \sigma_X &= \sqrt{np(1 - p)}. \end{aligned} \tag{5.2}$$

The individual binary observations, such as the outcomes for the individual trials, are independent Bernoulli random variables all having the same probability of taking the value 1, i.e., the Bernoulli random variables are both independent and homogeneous. Such random variables are often called *Bernoulli trials*. The sum of n Bernoulli trials has a $B(n, p)$ distribution. That is, in general, if $Y_1, Y_2, \ldots, Y_n$ are independent Bernoulli random variables and $P(Y_i = 1) = p$ for all i, and we define $X = \sum_{i=1}^{n} Y_i$, then $X \sim B(n, p)$. Note that when $n = 1$ X is a Bernoulli random variable and we have

$$E(X) = P(X = 1) \tag{5.3}$$

which is a special case of (5.2) and is easy to check ($E(X) = 1 \cdot P(X = 1) + 0 \cdot P(X = 0) = P(X = 1)$).

Binomial distributions usually arise as the sum of Bernoulli trials. Thus, the binomial distribution is reasonable to assume if the Bernoulli random variables appear to be independent and homogeneous. It is important to consider both assumptions carefully. In particular, the assumptions of independence and homogeneity are frequently violated when the Bernoulli random variables are observed across time. Let us now state these assumptions again in the context of patient P.S.

Example 1.4 (continued from p. 9 and 47) In judging the 14 out of 17 occasions on which P.S. chose the non-burning house by statistical methods we would assume that the set of 17 forced choices were Bernoulli trials. The independence assumption would be violated if P.S. had a tendency, say, to repeat the same response she had just given regardless of her actual perception. The homogeneity assumption would be violated if there were a drift in her response probabilities (e.g., due to fatigue) over the time during which the experiment was carried out. □

The $B(2, p)$ arises as the Hardy-Weinberg distribution in genetics. There, if the probability that an allele A is inherited from a parent is p, and the probability that the other possible allele B is inherited is $1 - p$, then the number of A alleles is $B(n, p)$ under the assumptions of independence and homogeneity. In this case the assumption of independence would be violated if somehow the two parents were coupled at the molecular level, so that the processes of separating the alleles in the two parents were connected; in most studies this seems very unlikely and thus the assumption of independence is quite reasonable. The second assumption is that there is a single, stable value for the probability of the allele A. This clearly could be violated: for instance, the population might actually be a mixture of two or more types of individuals, each type having a different value of $P(A)$; or, when the population is not in equilibrium due to such things as non-random mating, or genetic drift, we would expect deviations from the binomial prediction of the Hardy-Weinberg model. Indeed, in population genetics, a check on the fit of the Hardy-Weinberg model to a set of data is used as a prelimary test before further analyses are carried out.

Example 5.1 Nicotinic acetylcholine receptor and ADHD Attention deficit hyperactivity disorder (ADHD), a major psychiatric disorder among children, has been the focus of much recent research. There is evidence of heritability of ADHD,

and
effective medications (such as Ritalin) involve inhibition of dopamine transport. There is also evidence of involvement of the nicotine system, possibly due to its effects on dopamine receptors. Kent et al. (2001) examined genotype frequencies for the nicotinic acetylcholine receptor subunit $\alpha 4$ gene among children with ADHD and their parents. At issue was the frequency of a $T \rightarrow C$ exchange in one base in the gene sequence. In order to carry out the standard analysis the authors first examined whether the population appeared to be in equilibrium. If so, the probabilities of the allele combinations TT, CT, CC would be given by $B(2, p)$ distribution, according to the Hardy-Weinberg model. The frequencies for the 136 parents in their study were as follows:

	TT	CT	CC
Number	48	71	17
Frequency	.35	.52	.13
Hardy-Weinberg Probability	.38	.47	.15

In this case, the probabilities determined from the Hardy-Weinberg model (how we obtain these will be discussed in Chapter 7) are close to the observed allele frequencies, and there is no evidence of disequilibrium in the population (we also discuss these details later). Kent et al. went on to find no evidence of an association between this genetic polymorphism and the diagnosis of ADHD. □

In some cases the probability p is not stable across repetitions. Indeed, sometimes the change in probability is the focus of the experiment, as when learning is being studied.

Example 5.2 Learning impairment following NMDA antagonist injection Experiments on learning often record responses of subjects as either correct or incorrect in sequences of trials during which the subjects are given feedback as to whether their responses are correct or not. The subjects typically begin with a probability of being correct that is much less than 1, perhaps near the guessing value of .5, but after some number of trials they get good at responding and have a high probability of being correct, i.e., a probability near 1. An illustration of this paradigm comes from Smith et al. (2005), who examined data from an experiment in rats by Stefani et al. (2003) demonstrating that learning is impaired following an injection of an NMDA antagonist into the frontal lobe. In a first set of trials, the rats learned to discriminate light from dark targets, then, in a second set of trials, which were the trials of interest, they needed to discriminate smooth versus rough textures of targets. In two groups of rats a buffered salt solution with the NMDA antagonist was injected prior to the second set of trials, and in two other groups of rats the buffered salt solution without the antagonist was injected. Figure 5.1 displays the responses across 80 learning trials for set 2. It appears from the plot of the data that the groups of rats without the NMDA antagonist did learn the second task more quickly than the second group of rats, as expected.

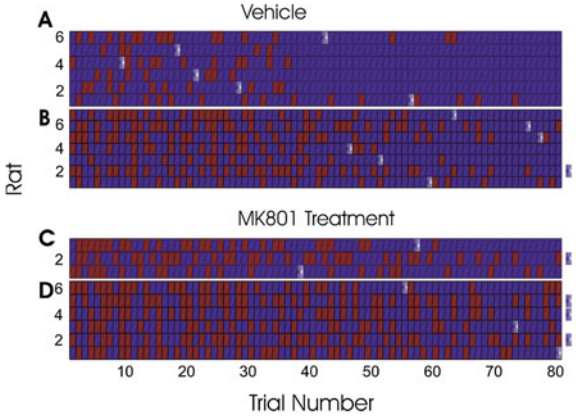

Fig. 5.1 Responses for 13 rats in the placebo group (labeled "Vehicle," in reference to the buffered solution vehicle) and 9 rats in the treatment group ("MK801 Treatment") for set 2. *Blue* and *red* indicate correct and incorrect responses, respectively. Each row displays responses for a particular rat across 80 trials. *Light blue triangles* indicate that the rat had 8 correct trials in a row. A *light blue triangle* appearing after the end of the trials, to the *right*, indicates that the rat did not achieve 8 correct trials in a row by the end of the 80 trials. Groups A and C were rewarded for dark arm on set 1 while groups B and D were rewarded for light arm on set 1. The rats in group A clearly learned the discrimination task relatively quickly. Modified and reprinted with permission from Smith et al. (2005).

The Smith et al. analysis was based on the method of maximum likelihood, which we will discuss in Chapter 7. For now, however, we may use the example to consider the possibility of aggregating the responses within groups of rats. Two possible ways to aggregate would be either across rats or across trials, the latter producing blocks of trials (e.g., 10 blocks of 8 trials). In each case, aggregation would produce a number X of correct responses out of a possible number n. We would then be able to plot the value of X across time in order to help examine the differences among the groups. If we were to assume $X \sim B(n, p)$, in each case, what would we be assuming about the trials themselves? If we were to aggregate across rats we would be assuming that the different rats' responses were independent, which is reasonable, and that the rats all had the same probability of responding correctly, which is dubious. Making this kind of dubious assumption is often a useful first step, and in fact can be innocuous for certain analyses, but it must be considered critically. After aggregating trials into blocks, the binomial assumption would be valid if the trials were independent and had the same probability of correct response, both of which would be dubious—though again potentially useful if its effects were examined carefully. In situations such as these it would be incumbent upon the investigator to show that aggregation would be unlikely to produce incorrect analytical results. □

Before leaving the binomial distribution, let us briefly examine one further application.

Example 5.3 Membrane conductance Anderson and Stevens (1973) were able to estimate single-channel membrane conductance by measuring total conductance at

a frog neuromuscular junction. Their method relied on properties of the binomial distribution. Suppose that there are n channels, each either open or closed, all acting independently, and all having probability p of being open. Let X be the number of channels that are open, and γ the single-channel conductance. Then the measured membrane conductance G satisfies $G = \gamma X$ where $X \sim B(n, p)$. From formulas (3.4) and (3.5) it follows that the mean and variance of G are given by

$$E(G) = \gamma np$$

and

$$V(G) = \gamma^2 np(1 - p).$$

Now, assuming that p is small, we have $1 - p \approx 1$ so that γ satisfies

$$\gamma = \frac{V(G)}{E(G)}.$$

Anderson and Stevens made multiple measurements of the membrane conductance at many different voltages, obtaining many estimates of $V(G)$ and $E(G)$. The slope of the line through the origin fitted to a plot of $V(G)$ against $E(G)$ thereby furnished an estimate of the single-channel conductance.[1] □

The Anderson and Stevens estimate of single-channel conductance is based on the approximate proportionality of the variance and mean across voltages. In the derivation above this was justified from the binomial, for small p. The small-p case of the binomial is very important and, in general, when p is small while n is large, the binomial distribution may be approximated by the Poisson distribution.

5.2 The Poisson Distribution

5.2.1 The Poisson distribution is often used to describe counts of binary events.

The Poisson distribution is the most widely-used distribution for *counts*. Strictly speaking, the Poisson distribution assigns a positive probability to every nonnegative integer 0, 1, 2, . . ., so that every nonnegative integer becomes a mathematical possibility. This may be contrasted with the binomial, which takes on numbers only up to some n, and leads to a proportion (out of n). The defining feature of the Poisson distribution, however, is that it arises as a small-p and large-n approximation to the binomial, which we discuss in Section 5.2.2. That mathematical charac-

[1] Additional comments on this method, and its use in analysis of synaptic plasticity, may be found in Faber and Korn (1991).

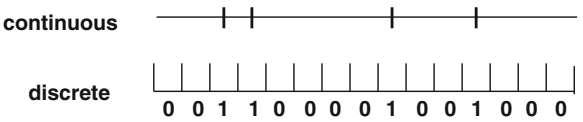

Fig. 5.2 Several event times displayed both in continuous time and in discrete time. In the discrete case time has been decomposed into bins and for each bin the presence or absence of an event is indicated by a 1 or 0.

terization portrays the count, approximately, as a sum of many binary variables, each indicating whether an event occurs (perhaps across time or across space), with each event occurrence having a small probability p. For example, neural spike counts are sometimes modeled as Poisson random variables. This results from a characterization of the spike train as a sequence of discrete event times, and if we decompose time into small bins (e.g., having 1 ms width) we may consider each time bin to define a binary variable that indicates whether a spike occurs within that bin, as depicted in Fig. 5.2. When we consider discrete events across time there is necessarily some time scale (corresponding to a small bin width) on which the events become rare, so that the probability p that any binary variable will take the value 1 becomes small. For a spiking neuron with a low or moderate firing rate (say 10 spikes per second or less), for example, a scale in milliseconds leaves large gaps (many milliseconds) between each spike and makes the probability of a spike within any 1 ms bin quite small (e.g., less than $10/1000 = .01$). For this reason the Poisson is often said to be a model for the variation in the number of occurrences of rare events.[2]

Counts of such "rare" events are common in neural data analysis, but it is important to recognize that many count distributions are discernibly *non-Poisson*. We begin our discussion with a classic data set from a situation where there are good reasons to think the Poisson distribution ought to provide an excellent description of the variation among counts. Although drawn from physics, this example helps to fix ideas about assumptions that generate Poisson variability. We then mention some situations in neural data analysis where Poisson distributions have been assumed. After that, we will elaborate on the motivation for the Poisson and then we will conclude with some discussion of frequently-occurring departures from Poisson variation among counts.

Example 5.4 Emission of α particles Rutherford et al. (1920) counted the number of α-particles emitted from a radioactive specimen during 2608 7.5 s (seconds) intervals.[3] The data are summarized in the table below. The first column gives the counts $0, 1, 2, \ldots, 9, \geq 10$, and the second column gives number of times the corresponding count occurred. For example, in 383 of the 2608 intervals there were 2 particles emitted. The third column provides the "expected" frequencies based on the Poisson distribution (obtained by maximum likelihood, defined in Section 7.2.2).

[2] The derivation of the Poisson distribution as an approximation to the binomial is credited to Siméon D. Poisson, having appeared in his book, published in 1837. Bortkiewicz (1898, *The Law of Small Numbers*) emphasized the importance of the Poisson distribution as a model of rare events.

[3] Rutherford et al. (1920, p. 172); cited in Feller (1968).

x	Observed	Expected
0	57	54.40
1	203	210.52
2	383	407.36
3	525	525.50
4	532	508.42
5	408	393.52
6	273	253.82
7	139	140.33
8	45	67.88
9	27	29.19
≥ 10	16	17.08

Here, the emission of any one particle is (on an atomic time scale) a "rare event" so that the number emitted during 7.5 s may be considered the number of rare events that occurred. □

The Poisson pdf is

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!} \tag{5.4}$$

and we write $X \sim P(\lambda)$. The mean, variance, and standard deviation of X are given by

$$\begin{aligned} E(X) &= \lambda \\ V(X) &= \lambda \\ \sigma_X &= \sqrt{\lambda}. \end{aligned}$$

The equality of variance and mean is highly restrictive and is often used to examine whether repeated series of observations depart from Poisson variation: a plot of variance versus mean should fall approximately on the line $y = x$.

Here is a physiological setting involving particle emissions where the Poisson distribution was used much as in Example 5.4.

Example 5.5 Human detection of light Hecht et al. (1942) investigated the sensitivity of the human visual system to very dim light, and calculated the number of light quanta required to drive perception. To do this, Hecht et al. constructed an apparatus that would emit very dim flashes of light, of 1 ms duration, in a darkened room; they presented these to several subjects and determined the proportion of times each subject would respond that he or she had seen a flash of light. In one part of their analysis, they assumed that the number of light quanta penetrating the retina would follow a Poisson distribution. If X is the number of quanta emitted, and if c is the number required for perception of the flash, then the probability of perception of flash is

$$P(X \geq c) = 1 - F(c - 1) \tag{5.5}$$

where $F(x)$ is the Poisson cumulative distribution function. (Note that the argument $c-1$ appears because $P(X \geq c) = P(X > c-1) = 1 - F(c-1)$.) Using the formula for the Poisson cdf (i.e., the summed pdf), Hecht et al. fit this to observed data and found that, roughly, a minimum of 6 quanta must be absorbed by the retina in order for a human to detect light. □

Not all applications of the Poisson distribution involve events across time. In the next example the events are distributed across space—on neural synaptic boutons.

Example 5.6 Quantal response in synaptic transmission The quantal response hypothesis is that a neurotransmitter is released from a large number of presynaptic vesicles in packets, or "quanta," each of which has a small probability of being released. To test this, del Castillo and Katz (1954) recorded postsynaptic potentials, or end-plate potentials (EPPs), at a frog neuromuscular junction. By assuming a Poisson distribution for the number of quanta released following an action potential, the authors obtained good experimental support for the quantal hypothesis. □

5.2.2 For large n and small p the binomial distribution is approximately the same as Poisson.

Example 5.6 (continued from Section 5.2.1) Let us go a step further in examining the argument of del Castillo and Katz. Under behavioral conditions the EPP would typically involve hundreds of quanta, but del Castillo and Katz used a magnesium bath to greatly decrease this number. In addition, they recorded spontaneous ("miniature") EPPs, which, according to the quantal hypothesis, should involve single quanta. They observed that this gave them two different ways to estimate the mean number of quanta released. The first method is to estimate the mean in terms of $P(X = 0)$ using the Poisson pdf formula $P(X = 0) = e^{-\lambda}$ or

$$\lambda = -\log P(X = 0). \tag{5.6}$$

To estimate $P(X = 0)$ they used the ratio D/C, where C was the total number of presynaptic action potentials and D was the number of times that the postsynaptic voltage failed to increase. Their second method used the ratio A/B, where A was the mean EPP voltage response following actional potentials and B was the mean spontaneous EPP voltage response. When the data from 10 experiments were plotted, the ten (x, y) pairs with $y = -\log D/C$ and $x = A/B$ were very close to the line $y = x$. □

A major motivation for the Poisson distribution is that it approximates the binomial distribution as p gets small and n gets large (with $\lambda = np$). One way to express this is given by the theorem below, but the argument used by del Castillo and Katz, described above, highlights both the key assumptions and the key mathematical result. Under the quantal hypothesis that vesicle release is binary *together with* the Bernoulli assumptions of independence and homogeneity, we have

$$P(X = 0) = (1 - p)^n$$

where p is the probability that any given vesicle will release and n is the number of vesicles. We define $\lambda = np$ and make the substitution $p = \lambda/n$, then take logs of both sides to get

$$\log P(X = 0) = n \log(1 - \frac{\lambda}{n}).$$

Now, for large n, an expansion of the log (see Section A.4 of the Appendix) gives $n \log(1 - \lambda/n) \approx -\lambda$. This says that Eq. (5.6) becomes a good approximation for small p and large n. The rest of the argument is given below.

Theorem: Poisson pdf approximation to binomial pdf For $\lambda > 0$, letting $p = \lambda/n$, as $n \to \infty$ we have

$$\binom{n}{k} p^k (1 - p)^{n-k} \to e^{-\lambda} \frac{\lambda^k}{k!}. \tag{5.7}$$

Proof: To derive Eq. (5.7), we use Eq. (A.6) from the Appendix, which we rewrite here by saying that as $n \to \infty$,

$$(1 - \frac{\lambda}{n})^n \to e^{-\lambda}. \tag{5.8}$$

Now let $\lambda = pn$, substitute $p = \lambda/n$ into the binomial pdf,

$$f(k) = \binom{n}{k} p^k (1 - p)^{n-k} = \frac{n!}{(n-k)!k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

and rearrange the terms to get

$$f(k) = A \cdot B$$

where

$$A = \left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)\cdots\left(\frac{n-k+1}{n}\right)$$

$$B = \underbrace{\left(\frac{\lambda^k}{k!}\right)}\ \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}\ \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-k}}.$$

As $n \to \infty$, the expression for A converges to 1; the expression over the first underbrace defining B remains constant (n does not appear there); by (5.8) the expression over the second underbrace defining B converges to $e^{-\lambda}$; and the expression over the third underbrace defining B converges to 1. This gives (5.7). □

5.2.3 *The Poisson distribution results when the binary events are independent.*

In thinking about the binomial assumption for a random variable X one generally ponders whether it is reasonable to conceptualize X as a sum of Bernoulli trials with the independence and homogeneity assumptions. Similarly, in the Poisson case, one typically asks whether the count variable X could be considered a sum of Bernoulli trials for small p and large n. The first requirement is that the counts really are sums of binary events. This means that X results from a string of 0s and 1s, as in Fig. 5.1, p. 109. In Example 5.4, p. 111, each emission event corresponds to a state transition in the nucleus of a particular atom. It is reasonable to assume that it is impossible for two nuclei to emit particles at precisely the same time and, furthermore, that each Geiger-counter "click" corresponds to exactly one particle emission. Independence, usually the crucial assumption, here refers to the independence of the many billions of nuclei residing within the specimen. This is an assumption, apparently well justified, within the quantum-mechanical conception of radioactive decay. It implies, for example, that any tendency for two particles to be emitted at nearly the same time would be due to chance alone: because there is no interaction among the nuclei, there is no physical "bursting" of multiple particles. Furthermore, the probability of an emission would be unlikely to change over the course of the experiment unless the specimen were so tiny that its mass changed appreciably. To summarize, the Poisson distribution for counts of events across time makes intuitive sense when we can conceptualize the events as Bernoulli trials, which are homogeneous and independent, where the success probability p is small.

The framework we have constructed above to discuss emission of α particles would apply equally well to quanta of light in the Hecht et al. experiment. What about the vesicles at the neuromuscular junction? Here, the quantal hypothesis is what generates the sequence of dichotomous events (release vs. no release). Is release at one vesicle independent of release at another vesicle? If neighboring vesicles tend to release in small clumps, then we would expect to see more variability in the counts than that predicted by the Poisson, while if release from one vesicle tended to inhibit release of neighbors we would expect to see more regularity, and less variability in the counts. It is reasonable to begin by assuming independence, but ultimately it is an empirical question whether this is justified. Homogeneity is suspect: the release probability at one vesicle may differ substantially from that at another vesicle. However, as del Castillo and Katz realized, homogeneity is actually not an essential assumption. We elaborate on this point when we return to the Poisson distribution, and its relationship to the Poisson process in Section 19.2.2.

Neuronal spike counts are sometimes assumed to be Poisson-distributed. Let us consider the underlying assumptions in this case. First, if measurements are made on a single neuron to a resolution of 1 ms or less, it is the case that a sequence of dichotomous firing events will be observed: in any given time bin (e.g., any given millisecond) the neuron either will or will not have an action potential, and it can not have two. But are these events independent? Immediately after a neuron has

fired, the membrane of a neuron undergoes changes that alter its propensity to fire again. In particular, there is a refractory period during which sodium channels are inactivated and the neuron can not fire again. This clearly violates the assumption of independence. In addition, there may be a build-up of ions, or activity in the local neural network, that makes a neuron more likely to fire if it has fired recently in the past (it may be "bursting"). This again would be a violation of independence. In many experiments such violations of independence produce markedly non-Poisson count distributions and turn out to have a substantial effect, but in others the effects are relatively minor and may be ignored. We indicated that, in the case of vesicle release of neurotransmitters, the homogeneity assumption is not needed in order to apply the Poisson approximation. The same is true for neuronal spike counts: the spike probabilities can vary across time and still lead to Poisson-distributed counts. The key assumption, requiring thought, is independence. On the other hand, the question of whether it is safe to assume Poisson variation remains an empirical matter, subject to statistical examination. As in nearly all statistical situations, judgment of the accuracy of the modeling assumptions—here, the accuracy of the Poisson distribution in describing spike count variation—will depend on the analysis to be performed.

5.3 The Normal Distribution

As we said in Chapter 3, the normal distribution (or Gaussian distribution) plays a dominant role in statistical theory because of the Central Limit Theorem, which we state in Chapter 6. In Section 5.3.1 we review a property of the normal distribution that leads to interpretation of standard errors and confidence intervals, and in Section 5.3.2 we note its relationship to the binomial and Poisson distributions.

5.3.1 Normal random variables are within 1 standard deviation of their mean with probability 2/3; they are within 2 standard deviations of their mean with probability .95.

We indicated on p. 60 that when X has a normal distribution probabilities of the form $P(a \leq X \leq b)$ can not be found directly by calculus and must, instead, be obtained numerically. Two such probabilities are so important in practice that they should be committed to memory. We will call these the "$\frac{2}{3}$ and 95% rule."

The $\frac{2}{3}$ and 95 % rule: For a normal random variable X with mean μ and standard deviation σ,

$$P(\mu - \sigma \leq X \leq \mu + \sigma) \approx \tfrac{2}{3}$$
$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx .95$$

We also have $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx .997$, but this is less important.

Example 5.7 Ebbinghaus on human memory A very early experiment on human memory was reported by Ebbinghaus (1885). Ebbinghaus used himself as the sole subject of his investigation, and he taught himself to learn lists of nonsense syllables made up of consonant-vowel-consonant trigrams such as DAX. Ebbinghaus memorized relatively long lists (e.g., 16 trigrams) to the point of being able to recite them without error, twice consecutively, and kept track of the time it took for him to achieve this success. He then repeated the task using the same lists after a delay period, that is, he re-learned the lists, and he examined the way his re-learning time increased with the length of the delay period. This was a way to quantify his rate of forgetting. (Compare the experiment of Kolers in Example 2.5 on p. 32.) The method Ebbinghaus used relied on the normal distribution. In one of his tabulations, he examined 84 memorization times, each obtained by averaging sets of 6 lists. He found the distribution of these 84 data values to be well approximated by the normal distribution, with mean 1,261 s and standard deviation 72 s.[4] This would mean that for about 2/3 of the sets of lists his learning time was between 1,189 s and 1,333 s. It also would mean that a set-averaged learning time less than 1,117 s or greater than 1,405 s would be rare: each of these would occur for only about 2.5 % of the sets of lists. □

It may seem odd that in examining the suitability of the normal distribution Ebbinghaus did not look at the distribution of learning times for lists, but rather chose to work with the distribution of *average* learning times across sets of 6 lists. The distribution of learning times was skewed. Only after averaging across several learning times did the distribution become approximately normal. This effect is due to the Central Limit Theorem, discussed in Section 6.3.1.

Normal distributions are often *standardized* so that $\mu = 0$ and $\sigma = 1$. In general, using Eq. (3.8) if $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$ then $Y \sim N(a\mu + b, a^2\sigma^2)$. As a special case, if $X \sim N(\mu, \sigma^2)$ and $Z = (X - \mu)/\sigma$ then $Z \sim N(0, 1)$. The $N(0, 1)$ distribution is called the *standard normal*. This is often used for calculation: if we know probabilities for the $N(0, 1)$ distribution then we can easily obtain them for any other normal distribution. For example, we also have

$$P(a \leq X \leq b) = P(\frac{a - \mu}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{b - \mu}{\sigma}) = P(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma}).$$

Thus, the right-hand side may be found in order to obtain the answer for the left-hand side. Standardized variables are often denoted by Z, sometimes with the terminology *Z-score*.

[4] He actually found the "probable error," which is $.6745\sigma$ to be 48.4 s. See Stigler (1986) for a discussion of these data.

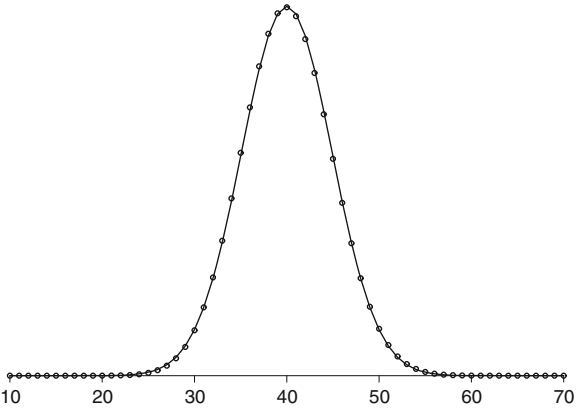

Fig. 5.3 The normal approximation to the binomial. *Black circles* are pdf values for a $B(100, .4)$ distribution; *curve* is pdf of a normal having the same mean and variance.

5.3.2 *Binomial and Poisson distributions are approximately normal, for large n or large λ.*

The normal distribution may be used to approximate a large variety of distributions for certain values of parameters. In the case of the binomial with parameters n and p, we take the normal mean and standard deviation to be $\mu = np$ and $\sigma = \sqrt{np(1-p)}$. An illustration is given in Fig. 5.3. The approximation is generally considered to be quite accurate for most calculations when n is large and p is not close to its boundary values of 0 and 1; a commonly-used rule of thumb (which is somewhat conservative, at least for $.2 < p < .8$) is that it will work well when $np \geq 5$ and $n(1-p) \geq 5$.

In the case of the Poisson with parameter λ we take the normal mean and standard deviation to be $\mu = \lambda$ and $\sigma = \sqrt{\lambda}$; the approximation is generally considered to be acceptably accurate for many calculations[5] when $\lambda \geq 15$.

These approximations are a great convenience, especially in conjunction with the "$\frac{2}{3} - 95\,\%$ rule."

[5] Actually, different authors give somewhat different advice. The acceptability of this or any other approximation must depend on the particular use to which it will be put. For computing the probability that a Poisson random variable will fall within 1 standard deviation of its mean, the normal approximation has an error of less than 10 % when $\lambda = 15$. However, it will not be suitable for calculations that go far out into the tails, or that require several digits of accuracy. In addition, a computational fine point is mentioned in many books. Suppose we wish to approximate a discrete cdf $F(x)$ by a normal, say $\tilde{F}(x)$. The the value $\tilde{F}(x + .5)$ is generally closer to $F(x)$ than is $\tilde{F}(x)$. This is sometimes called a continuity correction.

5.4 Some Other Common Distributions

5.4.1 *The multinomial distribution extends the binomial to multiple categories.*

In Example 5.1, on p. 107, we cited an application of the Hardy-Weinberg model in a study of genotype frequencies for the nicotinic acetylcholine receptor subunit $\alpha 4$ gene among children with ADHD and their parents. The three genotypes were labeled *TT*, *CT*, *CC*. This constitutes three distinct categories. For the ith individual in the study, let $Y_i = (1, 0, 0)$ if that individual has genotype *TT*, $Y_i = (0, 1, 0)$ if that individual has genotype *CT*, and $Y_i = (0, 0, 1)$ if that individual has genotype *CC*. The variable Y_i thus indicates the genotype of the ith individual, for $i = 1, 2, \ldots, n$. Let $p_1 = P(Y_i = (1, 0, 0))$, $p_2 = P(Y_i = (0, 1, 0))$, and $p_3 = P(Y_i = (0, 0, 1))$, where $p_1 + p_n + p_3 = 1$ and define $X = \sum_{i=1}^{n} Y_i$. Note that X gives the number of individuals, among a total of n, that have each of the three genotypes. In the Kent et al. data in Example 5.1 there were 136 individuals: 48 of genotype *TT*, 71 of genotype *CT*, and 17 of genotype *CC*, and we could write $X = (48, 71, 17)$. If we assume the $Y_1, Y_2, \ldots, Y_n$ are independent then X follows a *multinomial* distribution, written $X \sim M(n; p_1, p_2, p_3)$ with pdf

$$P(X = (x_1, x_2, x_3)) = \frac{n!}{x_1!x_2!x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3}. \tag{5.9}$$

According to the Hardy-Weinberg model the probabilities (p_1, p_2, p_3) would be restricted to satisfy the binomial pdf $p_1 = p^2$, $p_2 = 2p(1 - p)$ and $p_3 = (1 - p)^2$. However, (5.9) holds regardless of the validity of the Hardy-Weinberg model, as long as the genotypes are independent and homogeneous across individuals.

More generally, a random variable is distributed as $X \sim M(n; p_1, p_2, \ldots, p_k)$ if its pdf is given by

$$P(X = (x_1, x_2, \ldots, x_k)) = \frac{n!}{x_1!x_2! \cdots x_k!} \prod_{j=1}^{k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

where $p_1 + \cdots + p_k = 1$ and $x_1 + \cdots + x_k = n$. When $k = 2$ we obtain as a special case the binomial pdf of Eq. (5.1). (To see this, with x as in Eq. (5.1) define (x_1, x_2) in (5.1) to be $(x_1, x_2) = (x, n - x)$.) Thus, the multinomial is an extension of the binomial to multiple categories.

5.4.2 *The exponential distribution is used to describe waiting times without memory.*

We defined the exponential distribution in Eq. (3.12), p. 56, using it to illustrate calculations based on the pdf, and we showed how it may be applied to ion channel activation durations in Example 3.5. The exponential distribution is very special[6] because of its "memoryless" property. To understand this, let X be the length of time an ion channel is open, and let us consider the probability that the channel will remain open for the next time interval of length h. For example, h might be 5 ms. How do we write this? If we begin the moment the channel opens, i.e., at $x = 0$, the next interval of length h is $(0, h)$ and we want $P(X > h)$. On the other hand, if we begin at time $x = t$, for some positive t, such as 25 ms, the interval in question is $(t, t + h)$ and we are asking for a *conditional* probability: if the channel is open at time t we must have $X > t$, so we are asking for $P(X > t + h|X > t)$. We say that the channel opening duration is memoryless if

$$P(X > t + h|X > t) = P(X > h) \tag{5.10}$$

for all $t > 0$ and $h > 0$. That is, if $t = 25$ ms, the channel does not "remember" that it has been open for 25 ms already; it still has the same probability of remaining open for the next 5 ms that it had when it first opened; and this is true regardless of the time t we pick. The exponential distributions are the *only* distributions[7] that satisfy Eq. (5.10).

Contrast this memorylessness with, say, a uniform distribution on the interval [0, 10], measured in milliseconds. According to this uniform distribution, the event (e.g., the closing of the channel) must occur within 10 ms and initially every 5 ms interval has the same probability. In particular, the probability the event will occur in the first 5 ms, i.e., in the interval [0, 5], is the same as the probability it will occur in the last 5 ms, in [5, 10]. Both probabilities are equal to $\frac{1}{2}$. However, if at time $t = 5$ ms the event has not yet occurred then we are *certain* it will occur in the next half second [5, 10], i.e., this probability is 1, which is quite different than $\frac{1}{2}$. In anthropomorphic language we might say the random variable "remembers" that no event has yet occurred, so its conditional probability is adjusted. For the exponential distribution, the probability the event will occur in the next 5 ms, given that it has not already occurred, stays the same as time progresses.

Theorem A random variable X satisfies $X \sim Exp(\lambda)$ for some $\lambda > 0$ if and only if (5.10) is satisfied for all positive t and h, i.e., if X is memoryless.

Proof: Using Eq. (3.13) we have

---

[6] Another reason the exponential distribution is special is that among all distributions on $(0, \infty)$ with mean $\mu = 1/\lambda$, the $Exp(\lambda)$ distribution has the maximum entropy. See Eq. (4.33).

[7] The memoryless property can also be stated analogously for discrete distributions; in the discrete case only the *geometric* distributions are memoryless.

$$\begin{aligned} P(X > t+h | X > t) &= \frac{P(X > t+h, X > t)}{P(X > t)} \\ &= \frac{P(X > t+h)}{P(X > t)} \\ &= \frac{e^{-\lambda(t+h)}}{e^{-\lambda t}} \\ &= e^{-\lambda h} \\ &= P(X > h). \end{aligned}$$

Thus, every exponential distribution is memoryless. On the other hand, let $G(x) = 1 - F(x)$ where $F(x)$ is the distribution function of X. Memorylessness implies

$$P(X > t+h) = P(X > t)P(X > h)$$

i.e.,

$$G(t+h) = G(t)G(h)$$

for all positive t and h. But (as mentioned in Section A.4 of the Appendix), $G(x)$ can satisfy this equation for all positive t and h only if it has an exponential form $G(x) = ae^{bx}$. Because $F(x) = 1 - G(x)$ is a distribution function, it satisfies $F(x) \to 1$ as $x \to \infty$, which implies $b < 0$, and it satisfies $F(x) \to 0$ as $x \to 0$, which implies $a = 1$. Thus $F(x) = 1 - e^{-\lambda x}$ for some λ, i.e., $X \sim Exp(\lambda)$. □

An additional characterization of the exponential distribution is that it has a constant hazard function.

Theorem: A continuous random variable X satisfies $X \sim Exp(\lambda_0)$ if and only if its hazard function is $\lambda(x) = \lambda_0$.

Proof: First suppose $X \sim Exp(\lambda_0)$. The hazard function is easy to compute from the definition

$$\lambda(x) = \frac{f(x)}{1 - F(x)}.$$

Substituting $f(x) = \lambda_0 e^{-\lambda_0 x}$ and $F(x) = 1 - e^{-\lambda_0 x}$ we have

$$\begin{aligned} \lambda(x) &= \frac{\lambda_0 e^{-\lambda_0 x}}{e^{-\lambda_0 x}} \\ &= \lambda_0. \end{aligned}$$

On the other hand, if the hazard function is $\lambda(x) = \lambda_0$ we may rewrite the definition of $\lambda(x)$ and solve for $F(x)$,

$$F(x) = 1 - \lambda_0 f(x)$$

and then differentiate to get

$$f(x) = -\lambda_0 f'(x)$$

which implies that

$$f(x) = ce^{-\lambda_0 x}$$

for some constant c (see Section A.4) and because $f(x)$ must integrate to 1 we get $f(x) = \lambda_0 e^{-\lambda_0 x}$. □

The constant hazard of the exponential may be considered another way to view memorylessness: with constant hazard, given that the event has not already occurred at time t the probability that the event occurs in the next infinitesimal interval $(t, t + dt)$ is the same as it would be for any other infinitesimal interval $(t', t' + dt)$.

In Chapter 19 we will discuss the role played by the exponential distribution in *Poisson processes*, which are sometimes used to model spike trains. A technical result used there is a version of the probability integral transform derived in Section 3.2.5.

Theorem: Exponential Variables from the Probability Integral Transform Suppose X is a continuous random variable having pdf $f_X(x)$ and cdf $F_X(x)$, and suppose further that $f_X(x) > 0$ on an interval (A, B) and $f_X(x) = 0$ otherwise. Let $\lambda(x)$ be the associated hazard function of X. If we define a random variable Y by $Y = G(X)$ where

$$G(x) = \int_A^x \lambda(u)du \tag{5.11}$$

then $Y \sim Exp(1)$.

Proof: Let us write the cdf of the $Exp(1)$ distribution as F_{Exp}. From the corollary to the probability integral transform on 64, if we define Y by

$$Y = F_{Exp}^{-1}(F_X(X)) \tag{5.12}$$

then $Y \sim Exp(1)$. It remains to show that for $G(x)$ defined by (5.11) we get

$$G(x) = F_{Exp}^{-1} F_X(x). \tag{5.13}$$

We have (p. 56)

$$F_{Exp}(y) = 1 - e^{-y}.$$

The inverse of this function is

$$F_{Exp}^{-1}(w) = -\log(1 - w). \tag{5.14}$$

The hazard function of X (Section 3.2.4) is

$$\lambda(x) = \frac{f_X(x)}{1 - F_X(x)}$$

which gives

$$F_X(x) = 1 - \frac{f_X(x)}{\lambda(x)}$$

and, because $f_X(x) = \lambda(x)e^{-\int_{-\infty}^{x} \lambda(u)du}$, we get

$$F_X(x) = 1 - e^{-\int_{-\infty}^{x} \lambda(u)du}. \tag{5.15}$$

Putting $w = F_X(x)$ in (5.15) and applying (5.14) we have

$$\begin{aligned} F_{Exp}^{-1}(F_X(x)) &= F_{Exp}^{-1}(w) \\ &= \int_A^x \lambda(u)du \end{aligned}$$

which is (5.13). □

5.4.3 Gamma distributions are sums of exponentials.

In Example 3.5, on p. 58, we illustrated a basic property of a gamma distribution: if $X_1, X_2, \ldots, X_n$ are distributed as $Exp(\lambda)$, independently, and $Y = X_1 + \cdots + X_n$, then $Y \sim Gamma(n, \lambda)$. Note that a $Gamma(1, \lambda)$ distribution is the same as an $Exp(\lambda)$ distribution. More generally, a random variable X is said to have a $Gamma(\alpha, \beta)$ distribution when its pdf is

$$f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

for $x > 0$ and is 0 when $x \leq 0$. Here, the function $\Gamma(a)$ is the gamma function:

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx.$$

The gamma function is a variant of the factorial function; we have $\Gamma(n) = (n-1)!$ for any positive integer n. If $X \sim Gamma(\alpha, \beta)$ then

$$E(X) = \frac{\alpha}{\beta}$$

$$V(X) = \frac{\alpha}{\beta^2}$$
$$\sigma_X = \frac{\sqrt{\alpha}}{\beta}.$$

Plots of the gamma will be displayed for the special case of the chi-squared distribution, in the Section 5.4.4.

5.4.4 *Chi-squared distributions are special cases of gamma distributions.*

If $W \sim N(0, 1)$ then $X = W^2$ is said to have a *chi-squared distribution* on 1 degree of freedom, which is written $X \sim \chi_1^2$. If $W_i \sim \chi_1^2$ for all $i = 1, \ldots, n$, independently, and if $X = W_1 + W_2 + \cdots + W_n$, then X is said to have a chi-squared distribution on n degrees of freedom, written $X \sim \chi_n^2$. The most important way chi-squared distributions arise is as sums of squares of independent normal distributions. In general, a random variable X is said to have a chi-squared distribution with degrees of freedom ν, written χ_ν^2, if it has a $Gamma(\alpha, \beta)$ distribution with $\alpha = \frac{\nu}{2}$ and $\beta = \frac{1}{2}$.

If $X \sim \chi_\nu^2$ then

$$E(X) = \nu$$
$$V(X) = 2\nu$$
$$\sigma_X = \sqrt{2\nu}.$$

Figure 5.4 shows several chi-squared pdfs. Note that, for small degrees of freedom, the distribution is skewed toward high values (or skewed to the right). That is, it is not symmetrical, but rather large values distant from the middle (to the right) are more likely than small values distant from the middle (to the left). For the χ_4^2, the middle of the distribution is roughly between 1 and 6 but values less than 0 are impossible while values much greater than 7 have substantial probability. For large degrees of freedom ν the χ_ν^2 becomes approximately normal. For $\nu = 16$ in Fig. 5.4 there remains some slight skewness, but the distribution is already pretty close to normal over the plotted range.

5.4.5 *The beta distribution may be used to describe variation on a finite interval.*

A random variable X is said to have a beta distribution with parameters α and β if its pdf is

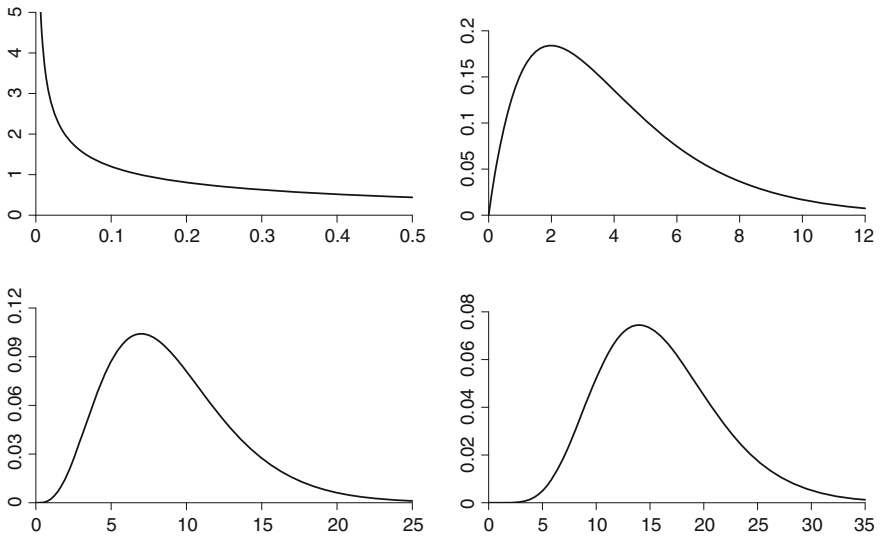

Fig. 5.4 Chi-squared pdfs for four values of the degrees of freedom: $\nu = 1$ (*Top left*), 4 (*top right*), 9 (*bottom left*), and 16 (*bottom right*).

$$f(x) = \frac{\gamma(\alpha+\beta)}{\gamma(\alpha)\gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{5.16}$$

for $0 < x < 1$ and is 0 otherwise. We then write $X \sim Beta(\alpha, \beta)$. Suppose $W_1 \sim Gamma(\alpha_1, \beta)$ and $W_2 \sim Gamma(\alpha_2, \beta)$, independently, and let $X = W_1/(W_1 + W_2)$. Then we have $X \sim Beta(\alpha, \beta)$.

If $X \sim Beta(\alpha, \beta)$ then $E(X) = \alpha/(\alpha+\beta)$ and $V(X) = \alpha + \beta + 1$. The beta distribution is sometimes written instead in terms of the parameters $\mu = E(X)$ and $\nu = V(X) - 1$, so that $\alpha = \mu\nu$ and $\beta = (1-\mu)\nu$. The beta distribution is commonly used to describe continuous variation that is confined to (0, 1). By rescaling it is easy to obtain a distribution confined to any finite interval (a, b). When $\alpha > 1$ and $\beta > 1$ the beta pdf is unimodal and $f(x) \to 0$ as $x \to 0$ or $x \to 1$, and if $\alpha = \beta$ the pdf is symmetric about $x = .5$. A unimodal symmetric beta pdf was plotted in Fig. 3.3.

The beta pdf arises in Bayesian analysis of binomial data, which is discussed in Section 7.3.9. There, the binomial parameter p must be in $(0, 1)$ and the beta distribution is used to represent knowledge about its value.

5.4.6 The inverse Gaussian distribution describes the waiting time for a threshold crossing by Brownian motion.

A random variable X is said to have an inverse Gaussian distribution if its pdf is

$$f(x) = \sqrt{\lambda/(2\pi x^3)}\exp(-\lambda(x-\mu)^2/(2\mu^2 x))$$

for $x > 0$. Here, $E(X) = \mu$ and $V(X) = \mu^3/\lambda$.

The inverse Gaussian arises as the theoretical interspike interval (ISI) distribution for integrate-and-fire neurons under simplifying assumptions. The essential idea is that excitatory and inhibitory post-synaptic potentials, EPSPs and IPSPs, are considered to arrive in a sequence of time steps of length δ, with each EPSP and IPSP contributing normalized voltages of $+1$ and -1, respectively, and with the probability of EPSP and IPSP being p and $1 - p$, where $p > 1 - p$ creates the upward "drift" toward positive voltages. Let X_t be the post-synaptic potential at time t with $t = 1, 2, \ldots$ and let $S_n = X_1 + X_2 + \cdots + X_n$. The variable S_n is said to follow a *random walk* (confer p. 530) and an action potential occurs when S_n exceeds a particular threshold value V_{thresh}. The process then resets to the resting potential V_{rest}. The behavior of a theoretical integrate-and-fire neuron based on such a random walk process is illustrated in Fig. 5.5. The continuous-time stochastic process known as *Brownian motion*, with drift, results from taking $\delta \to 0$ and $n \to \infty$, while also constraining the mean and variance in the form $E(S_n) \to m$ and $V(S_n) \to v$, for some m and v. The distribution of "first passage time," meaning the time it takes for the drifting Brownian motion to cross a boundary, is inverse Gaussian. (See Whitmore and Seshadri (1987). Also, Mudholkar and Tian (2002).) In particular, if we assume $p\delta \to \lambda_E$ and $(1 - p)\delta \to \lambda_I$, so that λ_E and λ_I are the limiting rates at which excitatory and inhibitory arrive, the drift toward the spiking threshold V_{thresh}, from the resting potential V_{rest}, becomes $\lambda_E - \lambda_I$ and the mean of the inverse Gaussian ISI distribution is

$$\mu = \frac{V_{thresh} - V_{rest}}{\lambda_E - \lambda_I}$$

and its coefficient of variation (defined in Eq. (3.11)) is

$$\sqrt{\frac{\mu}{\lambda}} = \sqrt{\frac{\lambda_E + \lambda_I}{2(V_{thresh} - V_{rest})(\lambda_E - \lambda_I)}}. \tag{5.17}$$

See Tuckwell (1988, Section 9.6). Thus, as the difference $V_{thresh} - V_{rest}$ increases the coefficient of variation of the ISIs decreases and the neuron fires more regularly. As excitation and inhibition become more nearly balanced, the coefficient of variation increases and the neuron fires more irregularly. Shadlen and Newsome (1998) used a closely-related analysis to argue the plausibility of roughly balanced excitation and inhibition in cortex. The random walk formulation was first given by Gerstein and Mandelbrot (1964).

Figure 5.6 gives an example of an inverse Gaussian pdf, with a Gamma pdf for comparison. Note in particular that when x is near 0 the inverse Gaussian pdf is very small. This gives it the ability to model, approximately, neuronal interspike intervals in the presence of a refractory period, i.e., a period at the beginning of the interspike interval (immediately following the previous spike) during which the neuron doesn't fire, or has a very small probability of firing.

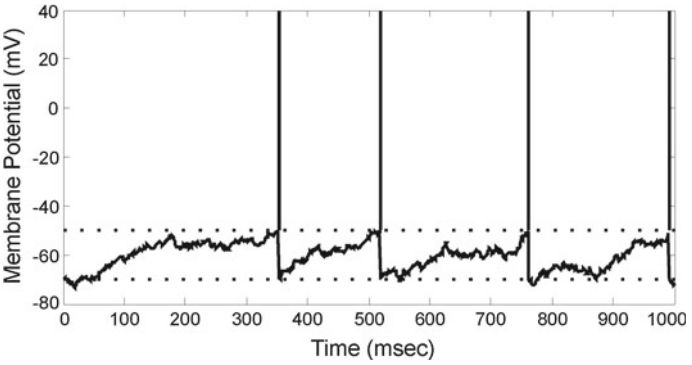

Fig. 5.5 Example of a computer-simulated integrate-and-fire neuron. At each time step there is either an EPSP or an IPSP, with probabilities p and $1-p$. For $p > 1-p$ this creates a stochastic upward "drift" of the voltage (as the inputs are summed or "integrated") until it crosses the threshold and the neuron fires. The neuron then resets to its baseline voltage. The resulting interspike interval (ISI) distribution is approximately inverse Gaussian.

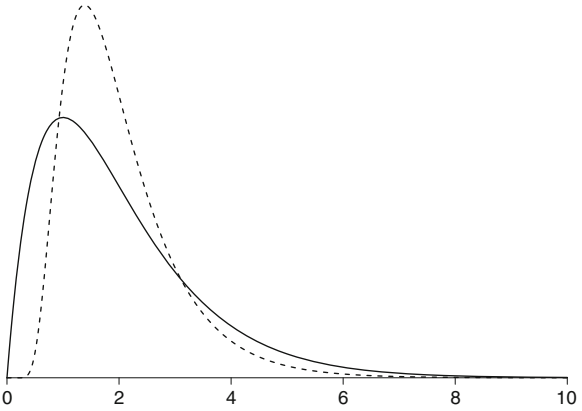

Fig. 5.6 Inverse Gaussian pdf plotted together with a *Gamma*$(2, 1)$ pdf. The inverse Gaussian (*dashed line*) has the same mean and variance as the gamma (*solid line*).

Example 5.8 Fit of integrate-and-fire model to cochlear neuron inter-spike intervals When they introduced the random walk integrate-and-fire model, and pointed out the inverse Gaussian would be the resulting approximate distribution for the inter-spike intervals, Gerstein and Mandelbrot (1964) provided illustrative fits to data. Figure 5.7 shows one such fit to a set of data from a cat cochlear neuron, under anesthesia. □

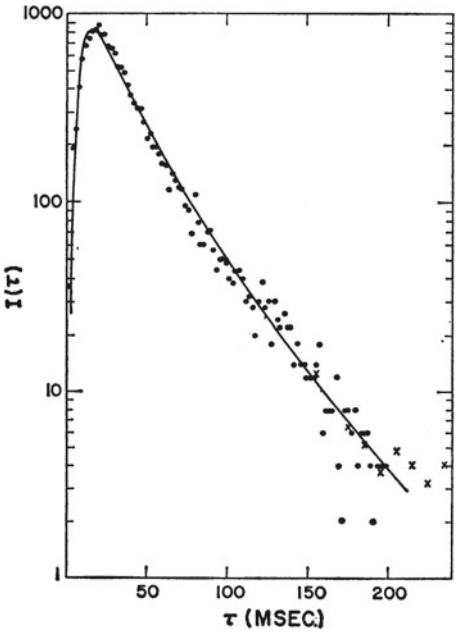

Fig. 5.7 Fitted inverse Gaussian pdf (*solid line*) together with dots indicating the heights of histogram bins based on inter-spike interval data from a cat cochlear neuron. The x-axis is time in milliseconds and the y-axis is the histogram height on a log scale. The data conform well to the inverse Gaussian distribution. Adapted from Gerstein and Mandelbrot (1964).

5.4.7 The t and F distributions are defined from normal and chi-squared distributions.

Two distributions are used very frequently in statistical hypothesis testing. The first is the t distribution.

If $X \sim N(0, 1)$ and $Y \sim \chi^2_\nu$, independently, then

$$T = \frac{X}{\sqrt{\frac{Y}{\nu}}}$$

is said to have a t distribution on ν degrees of freedom, which we write as $T \sim t_\nu$. This form of the T ratio arises in "t tests" and related procedures.

Note that T would be $N(0, 1)$ if the denominator were equal to 1. The denominator is actually very close to one when ν is large: if $Y \sim \chi^2_\nu$ we have $E(Y/\nu) = 1$ while $V(Y/\nu) = 2\nu/\nu^2$ which becomes very close to zero for large ν. That is, the random variable Y/ν has a very small standard deviation and thus takes values mostly very close to its expectation of 1. Therefore, for large ν, the t_ν distribution is very close to a $N(0, 1)$ distribution. One rule of thumb is that for $\nu > 12$, when computing

probabilities in the middle of the distribution, the t_ν distribution may be considered essentially the same as $N(0, 1)$. For small ν, however, the probability of large positive and negative values becomes much greater than that for the normal. For example, if $X \sim N(0, 1)$ then $P(X > 3) = .0014$ whereas if $T \sim t_3$ then $P(T > 3) = .029$, about 20 times the magnitude. To describe this phenomenon we say that the t_3 distribution has much *heavier tails* (or *thicker tails*) than the normal.

The t distribution was first derived by William Gosset under the pen name "A. Student." It is therefore often called *Student's t* distribution.

If $X \sim \chi^2_{\nu_1}$ and $Y \sim \chi^2_{\nu_2}$, independently, then

$$F = \frac{X/\nu_1}{Y/\nu_2}$$

is said to have an F distribution on ν_1 and ν_2 degrees of freedom, which are usually referred to as the numerator and denominator degrees of freedom. We may write this as $F \sim F_{\nu_1,\nu_2}$. This distribution arises in regression and analysis of variance, where ratios of sums of squares are computed and each sum of squares has (under suitable assumptions) a chi-squared distribution.

When $\nu_1 = 1$ the numerator is the square of a normal and $F = T^2$, where T is the ratio of a $N(0, 1)$ and the square-root of a $\chi^2_{\nu_2}$. That is, the square of a t_ν distributed random variable has an $F_{1,\nu}$ distribution. Also, analogously to the situation with the t_ν distribution, when ν_2 gets large the denominator Y/ν_2 is a random variable that takes values mostly very close to 1 and F_{ν_1,ν_2} becomes close to a $\chi^2_{\nu_1}$.

5.5 Multivariate Normal Distributions

5.5.1 A random vector is multivariate normal if linear combinations of its components are univariate normal.

We now generalize the bivariate normal distribution, which we discussed in Section 4.2.2. We say that an m-dimensional random vector X has an *m-dimensional multivariate normal distribution* if every nonzero linear combination of its components is normally distributed. If μ and Σ are the mean vector and variance matrix of X we write this as $X \sim N_m(\mu, \Sigma)$. Using (4.25) and (4.26) we thus characterize $X \sim N_m(\mu, \Sigma)$ by saying that for every nonzero m-dimensional vector w we have $w^T X \sim N(w^T \mu, w^T \Sigma w)$.

Notice that, just as the univariate normal distribution is completely characterized by its mean and variance, and the bivariate normal distribution is characterized by means, variances, and a correlation, the multivariate normal distribution is completely characterized by its mean vector and variance matrix. In many cases the components of a multivariate normal random vector are treated separately, with each diagonal element of the covariance matrix furnishing a variance, and the off-diagonal elements

being ignored. In some situations, however, the joint distribution, and thus all the elements of the variance matrix, are important.

If X has an m-dimensional multivariate normal distribution then each of its components has a univariate normal distribution. The following theorem extends this to the various components of X.

Theorem If X has an m-dimensional multivariate normal distribution and Y consists of the first k components of X, then Y has a k-dimensional multivariate normal distribution.

Proof: Let w be a non-zero k-dimensional vector. We must show that $w^T Y$ is univariate normal. Define $v(w)$ to be the m-dimensional vector consisting of the components of w followed by $m - k$ zeroes. Then $w^T Y = v(w)^T X$ and, by definition, $v(w)^T X$ is univariate normal; thus, $w^T Y$ is univariate normal. □

Example 4.1 (continued from p. 71) It is convenient to assume that the voltage amplitudes in Fig. 4.1 are 4-dimensional multivariate normal. According to the theorem above, this would imply that every pair of voltage amplitudes is bivariate normal. The $6 = \binom{4}{2}$ bivariate data plots in panel B of Fig. 4.1 indicate, very roughly, shapes consistent with bivariate normality, as indicated by the overlaid elliptical contours. Univariate histograms with normal pdfs and normal Q-Q plots are also given in that figure. The Q-Q plots clearly indicate some departure from normality, due to heavy tails in the first three channels. For many statistical analyses this degree of departure from normality would be unlikely to produce severe inferential problems, but the extent to which it is a cause for concern depends on the question being asked and the procedure used to answer it. □

The multivariate normal distribution is even more prominent in multivariate data analysis than the normal distribution is for univariate data analysis. The main reason is that specifying only the first two moments, mean vector and variance matrix, is a huge simplification. In addition, there is a generalization of the Central Limit Theorem, which we give in Section 6.3.2.

5.5.2 The multivariate normal pdf has elliptical contours, with probability density declining according to a χ^2 pdf.

The definition given above, in Section 5.5.1, does not require Σ to be positive definite (see p. 617 of the Appendix). In discussing the bivariate normal pdf for (X, Y) we had to assume $\sigma_X > 0$, $\sigma_Y > 0$, and $-1 < \rho < 1$. This is equivalent to saying that the variance matrix of the (X, Y) vector is positive definite. When we work with the multivariate normal distribution we usually assume the variance matrix is positive definite. If X is m-dimensional multivariate normal, having mean vector μ and positive definite covariance matrix Σ, then its pdf is given by

$$f(x) = \frac{1}{\sqrt{(2\pi)^m |\Sigma|}} e^{-\frac{1}{2}Q(x)} \tag{5.18}$$

where

$$Q(x) = (x - \mu_X)^T \Sigma^{-1} (x - \mu_X)$$

with $|\Sigma|$ being the determinant of Σ. We have labeled the exponent by $Q(x)$ to emphasize that it gives a quadratic in the components of x, so that Eq. (5.18) generalizes Eq. (4.14). The positive definiteness of Σ implies that $|\Sigma| > 0$, so that the pdf is well defined. It also implies that the contours of $Q(x)$ and, therefore, of $f(x)$ are multidimensional ellipses (see Section A.8 of the Appendix), generalizing remarks we made about the bivariate normal on p. 82.

Using simple matrix multiplication arguments, it is not hard to show that if $X \sim N_m(\mu, \Sigma)$, and Σ is positive definite, then $Q(X)$ has a chi-squared distribution with m degrees of freedom.

> *Details:* Let Z be m-dimensional multivariate normal with the zero vector as its mean vector and the m-dimensional identity matrix as its variance matrix. The components of Z follow $Z \sim N(0, 1)$, independently. Thus, from the definition of the chi-squared distribution in Section 5.4.4, $Z^T Z \sim \chi^2_m$. Now, if $X \sim N_n(\mu, \Sigma)$ then, by the theorem on p. 92, $Y = \Sigma^{-1/2}(X - \mu)$ satisfies $Y^T Y \sim \chi^2_m$. But $Y^T Y = Q(X)$. □

Taken together these results imply that, for $c > 0$, each contour $\{x : Q(x) = c\}$ of the multivariate normal pdf is elliptical and encloses a region $\{x : Q(x) \leq c\}$ having probability determined from the χ^2_m distribution function.

The remarks we have just made about elliptical contours apply when Σ is positive definite, so that we may write the pdf in (5.18). Occasionally, however, one must deal with the non-positive definite case. This arises, for example, when one wants to model the joint variation of m variables by assuming it is concentrated in fewer than m dimensions (analogously to the bivariate case with $\rho = 1$). If $X \sim N_m(\mu, \Sigma)$ and Σ is not positive definite but instead has rank k where $k < m$, we may use the spectral decomposition to find a k-dimensional subspace in which the distribution may be represented by a pdf with elliptical contours. This arises in some applications of multivariate analysis. See Chapter 17.

> *Details:* If there are k positive eigenvalues of Σ we may write
>
> $$\Sigma = PDP^T$$
>
> where the first k diagonal elements of D are the positive eigenvalues. Let P_1 be the $m \times k$ matrix consisting of the first k columns of P, which are the eigenvectors corresponding to the positive eigenvalues. These k eigenvectors span a k-dimensional subspace V. Let $v_j = \text{col}_j(P)$ for $j = 1, \ldots, k$, so that every vector $x \in V$ may be written in the

form

$$x = \sum_{j=1}^{k} u_j(x) v_j$$

and the n-dimensional vector x may instead be represented as a k-dimensional vector $u(x) = (u_1(x), \ldots, u_k(x)) = P_1^T x$. The distribution of X then lies in V in the sense that (i) $P(X \in V) = 1$ and (ii) for all non-zero $x \in V$,

$$x^T \Sigma x = u(x)^T D_\lambda u(x) > 0,$$

where D_λ is the $k \times k$ diagonal matrix with (i, i) element equal to the positive eigenvalue D_{ii}; in other words, D_λ is the $k \times k$ matrix formed by eliminating all the zero column and row vectors of D. Furthermore, setting $U = u(X)$ it may be shown that $U \sim N_k(\mu_U, D_\lambda)$, where $\mu_U = P_1 \mu$, and U has pdf

$$f_U(u) = \frac{1}{\sqrt{(2\pi)^k |D_\lambda|}} e^{-\frac{1}{2}(u-\mu_U)^T D_\lambda^{-1}(u-\mu_U)}.$$

□

We illustrate this kind of dimensionality reduction in Example 17.2 on p. 500.

5.5.3 *If X and Y are jointly multivariate normal then the conditional distribution of Y given X is multivariate normal.*

In Section 4.2.2 we introduced the bivariate normal distribution for a pair of random variables X and Y and in Section 4.2.4 we discussed the conditional expectation $E(Y|X = x)$, which is the regression function. We now generalize this to the case in which X and Y are random vectors. Let us suppose X and Y are, respectively, m_1-dimensional and m_2-dimensional; they are $m_1 \times 1$ and $m_2 \times 1$ vectors. Let us define U to be the concatenation of these two vectors,

$$U = \begin{pmatrix} X \\ Y \end{pmatrix}$$

with mean $\mu = E(U)$. Let us partition the components of μ so that they correspond to $E(X)$ and $E(Y)$, and let us use subscripts a and b to indicate this partitioning:

$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}$$

so that $\mu_a = E(X)$ and $\mu_b = E(Y)$. In this subsection we will partition matrices in the same way, separating the first m_1 rows and columns from the last m_2 rows and columns based on these subscripts. Thus, we write the variance matrix $\Sigma = V(U)$ as

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \tag{5.19}$$

so that $V(X) = \Sigma_{aa}$ and $V(Y) = \Sigma_{bb}$.

The generalization of the normal regression results in Section 4.2.4 is the following.

Theorem With the definitions above, if X and Y are jointly m-dimensional multivariate normal, then $Y|X = x$ is m_2-dimensional multivariate normal with mean vector and variance matrix given by

$$\mu_{b|a} = \mu_b + \Sigma_{ba}\Sigma_{aa}^{-1}(x - \mu_a) \tag{5.20}$$

$$\Sigma_{b|a} = \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}. \tag{5.21}$$

Outline of Proof: The theorem is proved by writing the quadratic exponent in the multivariate normal pdf of U, breaking it into pieces corresponding to the a and b components in the partitioning above, using the definition of conditional density, and then simplifying while applying the following matrix identity:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} E & -A^{-1}BF^{-1} \\ -F^{-1}CA^{-1} & F \end{pmatrix}$$

where

$$E = (A - BD^{-1}C)^{-1}$$

and

$$F = (D - CA^{-1}B)^{-1}$$ □

In carrying out calculations such as those used in proving the theorem above it is helpful to define the *precision matrix*,

$$\Gamma = \Sigma^{-1},$$

which is partitioned as

$$\Gamma = \begin{pmatrix} \Gamma_{aa} & \Gamma_{ab} \\ \Gamma_{ba} & \Gamma_{bb} \end{pmatrix}.$$

It is *not* generally true that $\Gamma_{bb} = \Sigma_{bb}^{-1}$. Instead we have

$$\Gamma_{bb} = \left(\Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}\right)^{-1}$$
$$\Gamma_{ba} = -\left(\Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}\right)^{-1}\Sigma_{ba}\Sigma_{aa}^{-1}$$

and by reversing the subscripts a and b we get the corresponding expressions for Γ_{aa} and Γ_{ab}.

Now suppose X and Y are random variables, U is a random vector, and (U, X, Y) is multivariate normal. Then, putting $V(U) = \Sigma_{aa}$ and $V(X, Y) = \Sigma_{bb}$ and applying the theorem, we write the components of the 2×2 matrix $\Sigma_{b|a}$ as

$$\sigma_{XX|U} = \Sigma_{b|a,11}$$
$$\sigma_{YY|U} = \Sigma_{b|a,22}$$
$$\sigma_{XY|U} = \Sigma_{b|a,12}.$$

We may then define the *partial correlation* of X and Y given U to be

$$\rho_{XY|U} = \frac{\sigma_{XY|U}}{\sqrt{\sigma_{XX|U} \cdot \sigma_{YY|U}}}. \tag{5.22}$$

The partial correlation $\rho_{XY|U}$ measures the remaining linear dependence of X and Y after conditioning on U. The *sample partial correlation* is the analogous quantity based on the sample covariance matrix S. That is, if we define the sample covariance matrix S as in (4.24) based on samples $x_1, \ldots, x_n$, $y_1, \ldots, y_n$ and $u_1, \ldots, u_n$ (where u is the vector sample analogue of U), and we then partition S as we partitioned Σ in (5.19), we write

$$s_{XX|U} = S_{b|a,11}$$
$$s_{YY|U} = S_{b|a,22}$$
$$s_{XY|U} = S_{b|a,12}$$

and then the sample partial correlation of x and y given u is[8]

$$\hat{\rho}_{XY|U} = \frac{s_{XY|U}}{\sqrt{s_{XX|U} \cdot s_{YY|U}}}. \tag{5.23}$$

The sample partial correlation in (5.23) is an estimate of the partial correlation[9] in (5.22).

[8] It may be shown that $\hat{\rho}_{XY|U}$ is equal to the correlation between the pair of residual vectors found from the multiple regressions (see Chapter 12) of x on u and y on u.

[9] In fact, $\hat{\rho}_{XY|U}$ is the maximum likelihood estimate; maximum likelihood estimation is discussed in Chapter 7.

Example 5.9 Network models from fMRI Many investigations have sought to describe large-scale network activity across the brain based on fMRI, particularly during a task-free "resting state." Suppose many regions of interest (ROIs) are defined, and let x_t be the sum of the fMRI signals across all voxels in one particular ROI at time t, for $t = 1, \ldots, T$. Let us call this ROI1. Similarly, let y_t be the sum of the fMRI signals across all voxels in another ROI at time t, and let us call this ROI2. Then the sample correlation $\hat{\rho}_{XY}$ of the vectors $(x_1, \ldots, x_T)$ and $(y_1, \ldots, y_T)$ may be used to define a "network connection" between ROI1 and ROI2. However, this measure suffers from the defect that any association between activity at these ROIs, represented by random variables X_t and Y_t, could be due to their correlated activity with other ROIs, which could be represented by a random vector U_t. That is, the other ROIs could be connected to both ROI1 and ROI2, and then X_t and Y_t would be correlated even if there were no connection between ROI1 and ROI2. An alternative is to use the sample partial correlations $\hat{\rho}_{XY|U}$ to define each network connection. Smith et al. (2011) conducted a large simulation study of fMRI network activity and found that partial correlation could be effective at identifying connected network nodes defined by ROIs. □

Chapter 6
Sequences of Random Variables

One of the great ideas in data analysis is to base probability statements on large-sample approximations, which are often easy to obtain either analytically or numerically. This short chapter contains the two fundamental results that produce most of the methodology, the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT). Both concern the behavior of the sample mean $\bar{X} = \sum_{i=1}^{n} X_i$. These theorems form a foundation for much data analytic theory because many statistical functions may be either rewritten or approximated in terms of sample means.

While sample means are important, the power of the LLN and CLT reaches far beyond means themselves to other summaries of the data. In general a numerical summary of the data is called a *statistic*. That is, a statistic is scalar or vector-valued function defined on the set of possible data values. For example, a regression coefficient, i.e., the slope of a least-squares fitted line, is a statistic. Many statistics may be written, at least approximately, as some function of a sample mean. This often produces approximate normality of the statistic which, as we will see in Chapters 7 and 8, becomes the basis for statistical inferences, such as confidence intervals and significance tests.

6.1 Random Sequences and the Sample Mean

We need a crucial piece of preliminary terminology: if $X_1, X_2, \ldots, X_n$ are drawn independently from the same distribution, then $X_1, X_2, \ldots, X_n$ is said to form a *random sample* from that distribution, and the random variables X_i are said to be *independent and identically distributed (i.i.d.)*. This section is about means computed from random samples (sets of i.i.d. random variables). Let $\mu = E(X_i)$. The LLN says that $\bar{X}$ gets arbitrarily close to μ as n increases indefinitely. The CLT says that the distribution of $\bar{X}$ becomes arbitrarily close to a normal distribution as n increases indefinitely. Similar results hold for many other data summaries, as well (because they may be written in terms of sample means). They are extremely important because

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_6,

they allow calculations based on normality, such as those in Section 5.3.1, to be applied, producing simple and useful probability statements.

In analyzing the behavior of the sample mean, the first point to recognize is that drawing a new sample would produce a new value of the sample mean, so that if we were to repeat the process of drawing a new sample many times, we would observe variability in the sample mean. In Example 5.7, p. 117, for example, we described some data on re-learning time from Ebbinghaus (1885), and noted that he examined 84 means, each of which was obtained by averaging the re-learning time across 6 lists of trigrams. Each mean was slightly different: they exhibited variation. The second point is that, typically,[1] the variation in the sample mean is smaller than that in the original data, and it decreases with increasing sample size.

Example 3.4 (continued from p. 46) Figure 3.2 displays a histogram of 60 spike counts from a motor cortical neuron during a reaching task. The mean among these 60 counts is 13.6 spikes. (The time interval was 600 ms, so this neuron's mean firing rate was 22 spikes per second.) Imagine drawing one spike count at random from among the 60, and doing this repeatedly. The histogram gives a sense of the variability we would see in these repeated random draws. Now suppose instead we were to draw 4 spike counts at random, and compute their mean, and then repeat this process many times. Because it would be likely that some of the 4 values would be bigger than 13.6, and some would be less, a mean of these 4 values would tend to be closer to 13.6 than any single random value would be—in other words, the mean of 4 observations would tend to exhibit less variability than did the original observations themselves. We can see this by considering the first 12 of the spike counts:

16 12 14 9 9 4 12 14 13 13 17 16

The mean count among these 12 is 12.4 spikes and the standard deviation is 3.7 spikes. Now consider the remaining 48 spike counts:

21	16	16	10	12	15	11	11	8	26	12	12
18	13	13	12	8	16	14	12	7	13	12	14
14	16	10	11	7	17	15	14	16	10	13	13
14	10	14	15	16	17	12	18	32	11	19	13

The data have been arranged in 12 columns of length 4 in order to consider the column means. In this case, the mean of these 12 means is 13.9 spikes and the standard deviation is 1.8 spikes: we find that the variation among the 12 means (the standard deviation of 1.8) is smaller than the variation among the 12 raw counts (the standard deviation of 3.7). □

The points illustrated by these motor cortical spike counts in Example 3.4 are (i) if we calculate the mean of a set of observations (a set of 4 trials) repeatedly for

[1] There are exceptions to this rule if the expectation does not exist, which can occur when the tails of the pdf fall to zero very slowly. An example is the Cauchy distribution, which is the t distribution on 1 degree of freedom.

new data (12 repetitions of the sets of 4) we observe variation among the means, and (ii) the variation among the means (the standard deviation of 1.8) is smaller than the variation we would typically see among the raw spike counts (the standard deviation of 3.7). However, this illustration was intended only to set the stage for an entirely theoretical discussion. In this section we consider the *random variable* $\bar{X}$. Its variation may be quantified by its standard deviation $\sigma_{\bar{X}}$. Notice that this is not the same thing as the standard deviation σ_X of the original data. In fact, $\sigma_{\bar{X}}$ decreases as the sample size increases; qualitatively, the larger the sample size, the less variation in the sample mean. Specifically, we have $\sigma_{\bar{X}} = \sigma_X/\sqrt{n}$. After giving this result in Section 6.1.1 we present the law of large numbers in Section 6.2.1 and the Central Limit Theorem in Section 6.3.1. These theorems require the use of some mathematics for dealing with sequences of random variables, which is the topic of Section 6.1.2.

6.1.1 The standard deviation of the sample mean decreases as $1/\sqrt{n}$.

If we repeatedly draw random samples $X_1, \ldots, X_n$, and from them repeatedly compute $\bar{X}$, the value of $\bar{X}$ will fluctuate: it will be a random variable. The dominant features of the distribution of $\bar{X}$ are captured by its mean and variance, which may be computed easily from the formulas (4.1) and (4.5).

Theorem If $X_1, X_2, \ldots, X_n$ form a random sample from a distribution having mean μ_X and standard deviation σ_X then the expectation and standard deviation of the mean $\bar{X}$ are

(i) $E(\bar{X}) = \mu_X$, and
(ii) $\sigma_{\bar{X}} = \sigma_X/\sqrt{n}$.

Proof: The expectation $E(\bar{X})$ is immediate from (4.1). For the variance, in formula (4.5) plug in $V(X_i) = \sigma_X^2$ to get $V(X_1 + X_2 + \cdots + X_n) = n\sigma_X^2$. Then take square-roots and, remembering that $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$, apply (3.6). □

The statement that $E(\bar{X}) = \mu_X$ says that the average amount by which $\bar{X}$ exceeds μ is equal to the average amount by which μ exceeds $\bar{X}$. The statement $\sigma_{\bar{X}} = \sigma_X/\sqrt{n}$ quantifies how rapidly the fluctuations in $\bar{X}$ diminish as a function of sample size. It is sometimes called "the square-root of n law." A consequence of diminishing fluctuations is that $\bar{X}$ must tend to get closer and closer to μ_X. This is the LLN, given in Section 6.2.1.

These results may be illustrated in the case of Bernoulli trials, where X_i is either 0 or 1. If $p = P(X_i = 1) = .4$ and $n = 4$ the sum $\sum_{i=1}^{n} X_i$ takes possible values of 0, 1, 2, 3, 4, with binomial probabilities .0625, .25, .375, .25, .0625. Thus, the mean $\bar{X}$

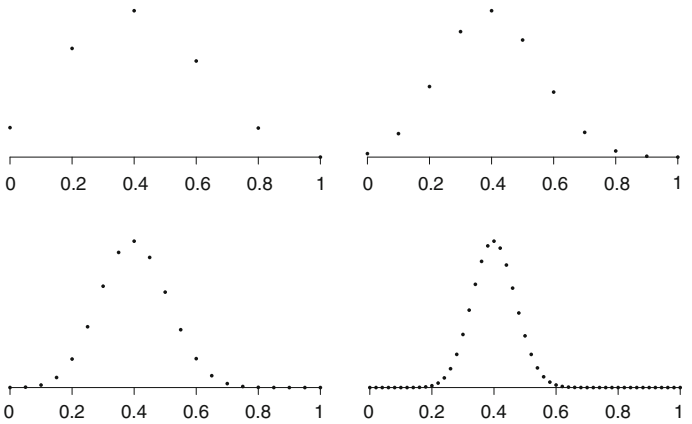

Fig. 6.1 The pdf of the binomial mean $\bar{X}$ *when* $p = .4$ for four different values of n. *As* n increases the distribution becomes concentrated ($\sigma_{\bar{X}}$ becomes small), with the center of the distribution getting close to $\mu_X = .4$ (the LLN). In addition, the distribution becomes approximately normal (the CLT).

takes possible values of 0, .25, .5, .75, 1, also with probabilities .0625, .25, .375, .25, .0625. The pdf is plotted in Fig. 6.1. The pdfs when $n = 10$, 25 and 100 are also shown there. For $n = 4$ the distribution is relatively wide, but as n increases it gets more concentrated. Note that in the case of the binomial we may write $Y = \sum_{i=1}^{n} X_i$, so that $Y \sim B(n, p)$ and then $\bar{X} = Y/n$. Using the binomial formula $V(Y) = np(1-p)$ (see p. 107) together with the general formula $V(aY) = a^2 V(Y)$ (see Eq. (3.9)) we get $\sigma_{\bar{X}} = \sqrt{p(1-p)/n}$.

For the square-root of n law to hold, the assumption of independence among the random variables $X_1, \ldots, X_n$ is crucial. Suppose instead that $Cor(X_i, X_j) = \rho$, with $\rho > 0$, for $i \neq j$ and let $\sigma^2 = V(X_i)$ for all i. A straightforward calculation shows that

$$V(\bar{X}) = \frac{\sigma^2}{n} + \frac{n-1}{n}\rho\sigma^2 \tag{6.1}$$

so that the variance does not vanish but instead reaches an asymptote: as $n \to \infty$ we have

$$V(\bar{X}) \to \rho\sigma^2. \tag{6.2}$$

Thus, even a small positive correlation among the variables destroys the result.

Details: For $i \neq j$ we have $Cov(X_i, X_j) = \rho\sigma^2$ and then

$$V(\bar{X}) = \frac{1}{n^2}\left[\sum_{i=1}^{n} V(X_i) + 2\sum_{i<j} Cov(X_i, X_j)\right]$$
$$= \frac{\sigma^2}{n} + \frac{n-1}{n}\rho\sigma^2. \qquad \square$$

Example 6.1 Neural spike count correlation could limit fidelity Shadlen and Newsome (1998) noted that common input to neurons can produce small, positive correlations in spike counts, and that this has been observed in recordings from primate cortex. As a consequence, they suggested, the information transmitted by groups of neurons acting together may be severely limited. The idea is that, according to the conception of integrate-and-fire neural transmission, an ensemble of neurons might transmit information to a downstream neuron based on their average spike count over small time intervals. In recordings from the MT area of visual cortex, correlations were estimated to be, on average, approximately $\rho = .12$. Shadlen and Newsome used the formula (6.2), stating that the asymptote in mean spike counts would be reached, approximately, by about 50–100 neurons. They concluded that "50–100 neurons might constitute a minimal signaling unit in cortex."

Details: Let $R = V(\bar{X})/\sigma^2$ and suppose we want to have the variance $V(\bar{X})$ be within 10 % of its asymptotic value. Letting $\epsilon = 1/10$ we set $R = \rho(1+\epsilon)$ and solve for n. From (6.1) we have

$$R = \frac{1-\rho}{n} + \rho$$

and solving for n we get

$$n = \frac{1-\rho}{R-\rho}.$$

We now insert $R - \rho = \rho\epsilon$ to get

$$n = \frac{1-\rho}{\rho}\frac{1}{\epsilon}.$$

With $\rho = .12$ and $\epsilon = .1$ this gives $n \approx 73$, supporting the observation made by Shadlen and Newsome.

$\square$

Various rebuttals to the argument in Example 6.1 have appeared in the literature, the most convincing being simply that neural computations could be more complicated than simple summation (averaging of spike counts), and more complicated combinations of inputs need not suffer from this difficulty. In any case, it is important to recognize the fundamental fact that small correlations can severely limit the information in a mean.

6.1.2 *Random sequences may converge according to several distinct criteria.*

In discussing the large-n behavior of a sequence of random variables $X_1, X_2, \ldots, X_n, \ldots$ we need a formalism for two kinds of statements. First, we want to be able to say that the distribution of X_n is approximately of a particular form. We do this by examining the cdfs. Suppose that the variables $X_1, X_2, \ldots, X_n, \ldots$ have corresponding cdfs $F_1(x), F_2(x), \ldots, F_n(x), \ldots$ and suppose further that the particular distribution that we want to consider an approximating distribution has cdf $F(x)$. We may then formalize the approximation by giving a precise meaning to the expression $F_n(x) \approx F(x)$ for n large, meaning that $F_n(x)$ is approximately equal to $F(x)$ for n large. We make this precise using limits. Recall that a sequence of numbers x_n, for $n = 1, 2, \ldots$, converges to x if for every $\epsilon > 0$ we have $|x_n - x| < \epsilon$ for all sufficiently large n. This is written $\lim_{n\to\infty} x_n = x$.

Definition Suppose $X_1, X_2, \ldots,$ is a sequence of random variables and F_n is the cdf of X_n. We say that X_n *converges in distribution* to a continuous random variable X with cdf F if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for all x. More generally, X_n converges in distribution to a random variable X with cdf F (which may or may not be continuous) if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for all x at which F is continuous. We often write this as

$$X_n \xrightarrow{D} X.$$

In cases in which X follows a particular well-known distribution we put the distribution on the right-hand side; e.g., if $X \sim N(0, 1)$ we write

$$X_n \xrightarrow{D} N(0, 1).$$

The second kind of statement we want to make has to do with the case in which the sequence of random variables $X_1, X_2, \ldots, X_n, \ldots$ gets progressively closer to a number, i.e., a fixed constant c rather than having some probability distribution. This is needed for the LLN. We may think of the constant as a probability distribution that has collapsed down to a point: we say that a random variable X is *degenerate*, meaning that it is identically equal to a constant c, when $P(Y = c) = 1$. In this situation the cdf of X is $F(x) = 0$ for $x < c$ and $F(x) = 1$ for $x \geq c$.

Definition Suppose $X_1, X_2, \ldots,$ is a sequence of random variables and F_n is the cdf of X_n. We say that X_n *converges in probability* to c if X_n converges in distribution

to the degenerate random variable X for which $P(X = c) = 1$. We often write this as

$$X_n \xrightarrow{P} c.$$

The notion of convergence in probability is more general than the definition above indicates, but we do not need the general definition. There are also two stronger notions of convergence, convergence in quadratic mean and convergence with probability one—but again we do not need these here.

Details: In applying convergence in probability, the criterion that is used is the following.

Theorem A sequence $X_1, X_2, \ldots$ converges in probability to c if and only if for every $\epsilon > 0$, $P(|X_n - c| > \epsilon) \to 0$ as $n \to \infty$.
Proof: This involves straightforward manipulations using the definition. The details are omitted. □

6.2 The Law of Large Numbers

6.2.1 As the sample size n increases, the sample mean converges to the theoretical mean.

The LLN is an accessible result, in the sense that its statement may be understood without advanced mathematics. The proof is not especially difficult, and we include it here, but we will regard it as an inessential detail.

Theorem: The Law of Large Numbers If $X_1, X_2, \ldots$ is a sequence of i.i.d. random variables having a distribution with mean μ_X and standard deviation σ_X, then $\bar{X}$ converges in probability to μ_X, i.e.,

$$X_n \xrightarrow{P} \mu_X.$$

The form of the LLN given here is sometimes called the "weak" law of large numbers. The strong law instead says that convergence occurs with probability 1. However, considerably more machinery is needed in order to say this in precise mathematical terms. Intuitively, "with probability 1" means that the convergence is certain to occur.

Details: The proof will require the following lemma.

Lemma (Markov's Inequality) Let Y be a positive random variable on (A, B) with $\mu_Y = E(Y) < \infty$. Then for any positive α,

$$P(Y > \alpha) < \frac{\mu_Y}{\alpha}.$$

Proof of Lemma: Let us assume that Y is continuous. We have

$$P(Y > \alpha) = \int_{\alpha}^{B} f_Y(y)dy$$

and

$$\alpha \int_{\alpha}^{B} f_Y(y)dy \leq \int_{\alpha}^{B} y f_Y(y)dy.$$

Combining these, and continuing, we then have

$$\begin{aligned} \alpha P(Y > \alpha) &\leq \int_{\alpha}^{B} y f_Y(y)dy \\ &\leq \int_{A}^{\alpha} y f_Y(y)dy + \int_{\alpha}^{B} y f_Y(y)dy \\ &= \int_{A}^{B} y f_Y(y)dy = E(Y). \end{aligned}$$

The case in which Y is not continuous may be handled by an analogous argument. □

Proof of Theorem: We need to show that for any positive ϵ we may find n sufficiently large that $P(|\bar{X} - \mu_X| > \epsilon)$ becomes arbitrarily close to 0. We have $P(|\bar{X} - \mu_X| > \epsilon) = P((\bar{X} - \mu_X)^2 > \epsilon^2)$. Let $Y = (\bar{X} - \mu_X)^2$, note that $E(Y) = \sigma_X^2/n$, and apply the Lemma to get

$$P(|\bar{X} - \mu_X| > \epsilon) < \frac{\sigma^2}{\epsilon^2 n}.$$

This shows that for sufficiently large n, $P(|\bar{X} - \mu_X| > \epsilon)$ becomes arbitrarily close to 0. □

6.2.2 The empirical cdf converges to the theoretical cdf.

We introduced the empirical cdf $\hat{F}_n(x)$ in Section 3.3 and noted there that, for large n, it approximates the cdf $F_X(x)$ and illustrated the phenomenon in Fig. 3.9. We now relate this behavior to the LLN.

In the proof we need the following definition: for a random variable X, we let the *indicator variable* $I_{\{X \leq x\}}$ be 1 if $X \leq x$ and 0 otherwise.

Theorem If $X_1, X_2, \ldots$ is a sequence of i.i.d. random variables then, for every x, $\hat{F}_n(x)$ converges in probability to $F(x)$.

Proof: Another way to think about $\hat{F}_n(x)$ is that it counts the number of random variables X_i in the random sample $X_1, \ldots, X_n$ for which $X_i \leq x$, and then divides by n. This is the same thing as adding $1/n$ for each of the X_i variables that are less than x. Mathematically, we express this counting operation using indicator variables. Consider a sequence $X_1, X_2, \ldots$ of i.i.d. random variables with cdf $F(x)$. We may write the empirical cdf in the form

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq x\}}.$$

We now use

$$\begin{aligned} E(I_{\{X_i \leq x\}}) &= 1 \cdot P(X_i \leq x) + 0 \cdot P(X_i > x) \\ &= P(X_i \leq x) = F(x) \end{aligned}$$

and apply the LLN. □

In addition to supplying the theoretical foundation for P-P and Q-Q plots, as discussed in Chapter 3, this result is also the starting point for the *bootstrap* method of statistical inference, which we cover in Chapter 9.

6.3 The Central Limit Theorem

6.3.1 For large n, the sample mean is approximately normally distributed.

The LLN concerns only the large-sample tendency of $\bar{X}$ to get arbitrarily close to μ_X. The CLT describes the large-sample probability distribution of $\bar{X}$. Actually, we are speaking a bit loosely here: the LLN says that the distribution of $\bar{X}$ becomes degenerate at μ_X; to get fluctuations that are described, approximately, by a normal distribution we have to introduce rescaling. Instead of $\bar{X}$, the CLT describes the behavior of the random sequence of variables Z_n, in which $\bar{X}$ is standardized by subtracting its mean and dividing by its standard deviation (the standard deviation of $\bar{X}$ being $\sigma_X/\sqrt{n}$).

The Central Limit Theorem: Suppose $X_1, X_2, \ldots$ is an i.i.d. sequence of random variables having mean μ_X and standard deviation σ_X, and let $Z_n = \sqrt{n}(\bar{X} - \mu_X)/\sigma_X$. Then Z_n converges in distribution to a normal random variable having mean 0 and variance 1, i.e.,

$$Z_n \xrightarrow{D} N(0, 1).$$

Proof Outline: The CLT may be proved using the Fourier transform. The Fourier transform of a pdf is called the *characteristic function* of the distribution. If $X_1, X_2, \ldots, X_n, \ldots$ is a sequence of random variables with characteristic functions $\phi_n(t)$, for $n = 1, 2, \ldots$, and $\phi_n(t) \to \phi(t)$ for all t with $\phi(t)$ being a characteristic function of the distribution of a random variable X, then X_n converges in distribution to X; this basic result is a version of the *continuity theorem*. Let us take $\phi_n(t)$ to be the sequence of characteristic functions of the distributions of the normalized sample means Z_n. Calculations show that $\phi_n(t)$ converges to the characteristic function of a $N(0, 1)$ distribution; therefore, by the continuity theorem, Z_n converges in distribution to a $N(0, 1)$ random variable. □

The effects of the LLN and CLT are illustrated in Fig. 6.1. For $n = 4$ the distribution of $\bar{X}$ does not look very close to normal. However, as n increases the distribution of $\bar{X}$ gets more tightly concentrated near the mean $\mu_X = .4$ (a consequence of the LLN) and it looks more and more normal (the CLT).

What we have just done is looked at the distribution of $\bar{X}$ for Bernoulli trials for several values of n with $p = .4$. The distribution of $n\bar{X}$ is binomial and the picture of its distribution would look just like the pictures we had for the distribution of $\bar{X}$ except that the x-axis would be multiplied by n. In particular, as n gets large we see that the distribution looks normal. This effect of the CLT may be considered an explanation for the normal approximation to the binomial.

In fact, there are much more general versions of the CLT. We do not want to build up the machinery needed for a general theorem, but it is worth stating one result in an imprecise form.

Roughly speaking, if $X_1, X_2, \ldots, X_n$ are independent random variables, possibly having different distributions but with no individual X_i making a dominant contribution to the mean $\bar{X}$, then for n sufficiently large, the distribution of $\bar{X}$ is approximately normal with mean $E(\bar{X})$ and standard deviation $\sqrt{V(\bar{X})}$.

The "no dominant contribution" phrase may be made precise as the *Lindeberg condition*, and the CLT then follows (see Billingsley 1995, Section 27). This version of the CLT helps to explain why the normal distribution arises so often in statistical

theory, and also why it seems to fit, at least crudely, so many observed phenomena. It says that whenever we average a large number of small independent effects, the result will be approximately normally distributed.

> *A detail*: Another way to interpret the CLT uses entropy, as defined in Eq. (4.33). Among all distributions having mean μ and standard deviation σ, the $N(\mu, \sigma^2)$ distribution is the most disorderly possible, in the sense of having maximal entropy. The CLT says that as the sample size gets very large the distribution of the sample mean becomes as disorderly as possible. This characterization provides an alternative way to understand and prove the CLT. See Madiman and Barron (2007).

There are also versions of the CLT for non-independent variables, though they are considerably more complicated. Those results typically require the sequence to be *stationary*, as defined on p. 515 of Chapter 18, and further limit the dependence among the random variables X_i and X_j within the sequence as $j - i$ increases. See Billingsley (1995, Theorem 27.4) and also Francq and Zakoian (2005).

6.3.2 For large n, the multivariate sample mean is approximately multivariate normal.

The multivariate version of the CLT is analogous to the univariate CLT. We begin with a set of multidimensional samples of size n: on the first variable we have a sample $X_{11}, X_{12}, \ldots, X_{1n}$, on the second, $X_{21}, X_{22}, \ldots, X_{2n}$, and so on. In this notation, X_{ij} is the jth observation on the ith variable. Suppose there are m variables in all, and suppose further that $E(X_{ij}) = \mu_i$, $V(X_{ij}) = \sigma_i^2$, and $Cor(X_{ij}, X_{kj}) = \rho_{ik}$ for all $i = 1, \ldots, m$, $j = 1, \ldots, n$, and $k = 1, \ldots, m$. As before, let us collect the means into a vector μ and the variances and covariances into a matrix Σ. We assume, as usual, that the variables across different samples are independent. Here this means X_{ij} and X_{hk} are independent whenever $i \neq h$. The sample means

$$\begin{aligned}
\bar{X}_1 &= \frac{1}{n}\sum_{j=1}^{n} X_{1j} \\
\bar{X}_2 &= \frac{1}{n}\sum_{j=1}^{n} X_{2j} \\
&\vdots \\
\bar{X}_m &= \frac{1}{n}\sum_{j=1}^{n} X_{mj}
\end{aligned}$$

may be collected in a vector

$$\bar{X} = \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_m \end{pmatrix}.$$

Multivariate Central Limit Theorem: Suppose $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_m$ are means from a set of m random samples of size n, as defined above, with the covariance matrix Σ being positive definite. For any m-dimensional vector w define

$$Z_n(w) = \sqrt{n} w^T \Sigma^{-\frac{1}{2}} (\bar{X} - \mu). \tag{6.3}$$

Then for every nonzero m-dimensional vector w, $Z_n(w)$ converges in distribution to a normal random variable having mean 0 and variance 1.

More loosely, the multivariate CLT says that $\bar{X}$ is approximately multivariate normal with mean μ and variance matrix $\frac{1}{n}\Sigma$. As in the univariate case, there are much more general versions of the multivariate CLT.

Chapter 7
Estimation and Uncertainty

7.1 Fitting Statistical Models

The examples in previous chapters, involving experimental settings ranging from human and animal behavior, to neuroimaging, EEG and EMG, neural spike trains, and in vitro recording, have illustrated the way statistical models describe regularity and variability of neural data. All of these models involve free parameters. In Example 1.5, on p. 11, we reviewed the use of least squares in demonstrating an approximately linear relationship between conduction velocity and nerve diameter. Least squares is easy to understand and often works well for models of the form

$$Y_i = f(x_i) + \epsilon_i.$$

But what about other situations? In Fig. 3.8 of Example 3.5, on p. 60, we displayed fits of $Gamma(\alpha, \beta)$ distributions to histograms of ion-channel opening durations, but we did not say how the parameters α and β were chosen. A näive approach to the problem of using the data to determine suitable values of parameters might propose a particular method and argue for it on intuitive grounds. According to the doctrine of statistics, however, principles may be introduced and used in analyzing the performance of alternative methods. By demonstrating the properties of solutions under general conditions, statistical theory brings coherence to an otherwise bewildering array of disparate problems. In this chapter, together with Chapters 8 and 9, we present the key ideas.

We start with a traditional, though somewhat artificial, separation of two aspects of the fitting problem that are intimately connected in practice: estimation of parameters and assessment of uncertainty. In Section 7.2 we formalize the process of estimation and then give two alternative methods, the *method of moments* and *maximum likelihood* (*ML*). In the 1920s Ronald Fisher proposed maximum likelihood and demonstrated that it is optimal quite generally for large sample sizes. Fisher also showed how uncertainty about the answer can be assessed, and an alternative perspective was provided at about the same time by Harold Jeffreys using Bayes' Theorem.

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_7,

It took roughly 50 more years to refine the early concepts to its full-fledged modern incarnation and, in fact, new variants of algorithms continue to be developed so that it may be applied to ever more complicated situations. In contexts where finitely-many parameter values completely specify[1] the statistical model, implementation of ML estimation is conceptually straightforward while, from a theoretical perspective, ML estimation is also provably unbeatable—no other method offers better performance, for large samples. ML estimation has, therefore, become the dominant approach to parameter estimation. We will review basic properties and uses of ML estimation in Chapter 8.

In Section 7.3 we discuss confidence intervals. In Chapter 1, on p. 13, we described the use of a confidence interval to assess the uncertainty associated with responses of patient P.S. when forced repeatedly to choose between pictures of burning and non-burning houses; we noted that an approximate 95 % confidence interval for her propensity to choose the non-burning house was (.64, 1.0) and we concluded it was not very likely that she was choosing them with equal probabilities (a propensity of .5); instead, she apparently saw the two complete pictures without conscious awareness of processing their left ends, which is where the fire appeared. As a data-analytic tool, confidence intervals have become straightforward to use in many, varied situations. We treat several simple yet important problems in Section 7.3 and supplement with more general methods in Chapters 8 and 9. As one thinks harder about interpretation, the subject gets somewhat more subtle. We review the issues in Sections 7.3.8 and 7.3.9. On the other hand, confidence intervals are fundamental to statistical practice and, from a contemporary standpoint, they seem very natural. Seen in historical context, the introduction of confidence intervals by Jerzy Neyman in the 1930s was quite ingenious, and a giant leap forward.

One of the ways confidence intervals are found in conjunction with maximum likelihood is to apply the *bootstrap*, which is discussed in Chapter 9. As additional motivation for the discussion in this and subsequent chapters, here is a concrete example where these methods have been used in fitting a statistical model of mental processes.

Example 7.1 A Model of Visual Attention Experiments on visual attention often study the ability of subjects to see and remember multiple objects that are exposed to them for a very short time. Following Sperling (1967), Bundesen and colleagues developed a quantitative theory of visual attention (Bundesen, 1998) according to which, objects in the visual field are compared with representations in visual memory, and if the comparison is completed prior to the end of visual exposure, the object is recognized. In this theory the time taken to process and store an object identity is a random variable. For object i call this random variable X_i. The processing is considered to begin after a latency of length t_0, so that if t is the total time an object is displayed then the ith object is recognized if $X_i \leq t - t_0$. Bundesen assumed $X_i \sim Exp(\lambda_i)$. Letting $f_i(x)$ and $F_i(x)$ be the $Exp(\lambda_i)$ pdf and cdf, for exposure of length $x = t - t_0$, $F_i(t - t_0)$ is the probability of object recognition success

[1] From the point of view of the mathematical theory, a nonparametric method does not eliminate the parameters but rather makes them infinite dimensional.

and $1 - F_i(t - t_0)$ is the probability of object recognition failure. Suppose S is the stimulus set and let R denote some particular subset of objects that are recognized. If the subject's memory capacity is not exceeded, and if recognition of object i is independent of recognition of all other objects (and this is true for every i), then the probability that the subject will recognize all objects in R, and fail to recognize all objects not in R (i.e., fail to recognize those in the complement, which may be written $S - R$), is given by

$$P_S(R) = \prod_{i \in R} F_i(t - t_0) \prod_{j \in S-R} (1 - F_j(t - t_0)). \tag{7.1}$$

This model has several unknown parameters (the encoding rates λ_i, the latency t_0, and the memory capacity) which must be determined in order to compute the probabilities and compare them to data. Figure 7.1 displays fits of the model to data from three subjects. The model fitting was performed by the method of maximum likelihood, and uncertainties associated with each of the parameters of interest may be obtained by bootstrap methods. See Kullingsbaek (2006). □

7.2 The Problem of Estimation

In order to fit a model to data, a parameter or set of parameters needs to be determined. Following a convention in the statistical literature, we use θ to denote a generic parameter. In much of our initial discussion we will focus on the case of a single, scalar parameter, but in most real-world problems θ becomes a vector. For example, in fitting a $Gamma(\alpha, \beta)$ model we would be taking $\theta = (\alpha, \beta)$ and we would speak of "the parameter" θ in place of "the parameters" α and β. The problem of estimation is to determine a method of estimating θ from the data. To constitute a well-defined method we must have an explicit procedure, that is, a formula or a rule by which a set of data values $x_1, x_2, \ldots, x_n$ produces an estimate. We consider an *estimator* to have the form $T = T(X_1, X_2, \ldots, X_n)$, i.e., the estimator is a random variable derived from the random sample. The properties of an estimator may be described in terms of its probabilistic behavior.

Before presenting the method of moments and maximum likelihood, we need to make two comments on notation. First, when we write $T = T(X_1, \ldots, X_n)$ we are using capital letters to indicate clearly that we are considering the estimator to be a random variable, and the terminology distinguishes the random "estimator" from an "estimate," the latter being a value the estimator takes. Nonetheless, neither we nor others in the literature are systematically careful in making this distinction; it is important conceptually, but some sloppiness is tolerable. Second, we often write θ^* or $\hat{\theta}$ for the value of an estimator, so we would have, say, $T = \hat{\theta}$. The latter notation, using $\hat{\theta}$ to denote an estimate, or an estimator, is very common in

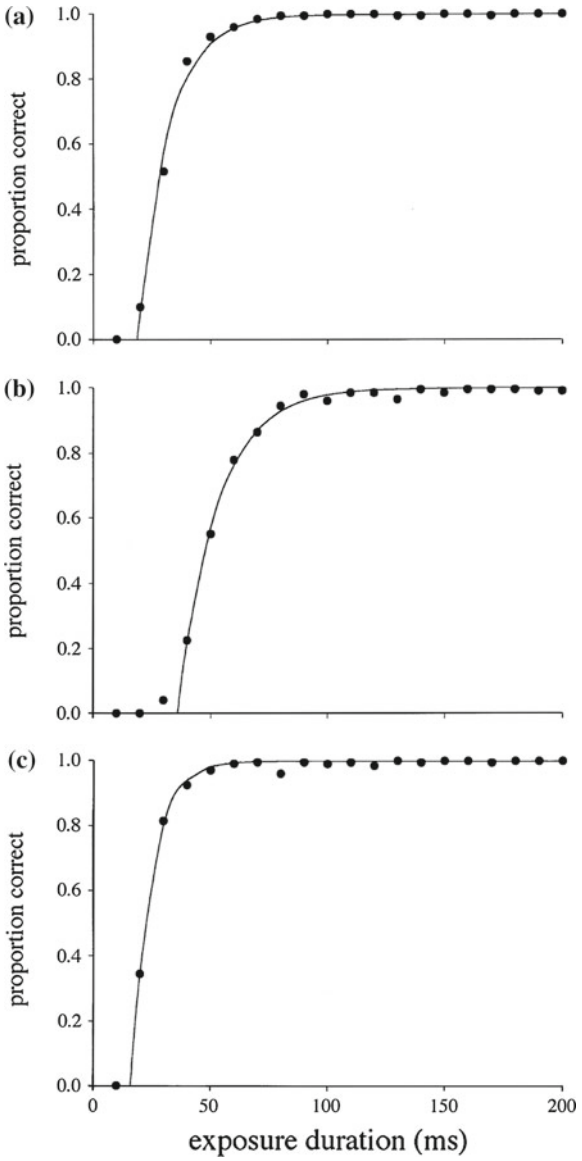

Fig. 7.1 Data from three subjects, together with fits of a model for probability of letter identification as a function of exposure duration. Adapted from Bundesen (1998).

the statistical literature. Sometimes, however, $\hat{\theta}$ refers specifically to the maximum likelihood estimator (MLE). This is another potential source of confusion, which the context should clarify.

7.2.1 The method of moments uses the sample mean and variance to estimate the theoretical mean and variance.

We have already indicated that ML is the dominant approach to estimating a parameter vector θ. For various reasons, however, other methods are sometimes used. In this section we present one of these other methods, the *method of moments*, which preceded the development of ML and is still used for some purposes. The idea is simple: to fit a probability distribution to a set of data we equate the theoretical mean and variance to the sample mean and variance and then solve for the unknown parameters.

Illustration: Fitting a gamma distribution On p. 124 we noted that the mean and variance of a $Gamma(\alpha, \beta)$ random variable are

$$\mu = \frac{\alpha}{\beta}$$
$$\sigma^2 = \frac{\alpha}{\beta^2}.$$

We may solve these for β and α: dividing the first equation by the second we get

$$\beta = \frac{\mu}{\sigma^2};$$

squaring the first and dividing by the second we get

$$\alpha = \frac{\mu^2}{\sigma^2}.$$

We then substitute $\bar{x}$ and s^2 for μ and σ^2 to obtain the method of moments estimator:

$$\beta^* = \frac{\bar{x}}{s^2}$$
$$\alpha^* = \frac{\bar{x}^2}{s^2}.$$

□

The method of moments is, in some cases, like the gamma, quite easy to apply. In principle, higher-order moments could be used (e.g., $E(\sum(X_i - \mu)^3)$ could be equated to the sample analogue).

7.2.2 *The method of maximum likelihood maximizes the likelihood function, which is defined up to a multiplicative constant.*

To introduce maximum likelihood estimation, let us begin by framing the estimation problem concretely, using the binomial, and let us write the binomial pdf in the form

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

which was previously denoted by $f(x) = P(X = x)$, with p replacing θ. Here the notation $f(x|\theta)$ is used to imply that we are examining the pdf of X given the value of θ. The binomial pdf describes the probabilities to be attached to varying possible values $X = x$ for a given fixed value of θ. That is, once we plug in a value of θ we have completely determined the pdf for all values of x. The problem of estimation, however, attempts to find a sensible guess at θ given that $X = x$ has been observed. It thus reverses the situation: instead of assuming a value for θ and finding values of x, we must assume a value of $X = x$ and come up with a value of θ. In this sense, it involves an *inverse* or *inductive* form of reasoning. The method of maximum likelihood chooses the value $\hat{\theta}$ of θ that assigns to the observed data x the highest possible probability:

$$f(x|\hat{\theta}) = \max_{\theta} f(x|\theta).$$

In the binomial problem we will, below, show that $\hat{\theta} = x/n$. In other words, maximum likelihood estimates the theoretical proportion (or propensity) θ by the observed proportion x/n.

> *A detail:* Why do we call θ a theoretical proportion? We have that X/n is the mean of n Bernoulli trials, each having probability θ of being 1. By the law of large numbers
>
> $$\frac{X}{n} \xrightarrow{P} \theta$$
>
> so that θ is, roughly speaking, the proportion of 1s observed in infinitely many trials. In this sense we can say that θ is a theoretical proportion. □

To understand the maximum likelihood idea better we consider what the pdf $f(x|\theta)$ tells us about the various possible values of θ. To do this we *invert* its functionality by thinking of $f(x|\theta)$ as a function of θ rather than of x. That is, having observed $X = x$, we fix x in the pdf $f(x|\theta)$ and then consider how each different choice of θ produces a different probability $f(x|\theta)$. We do not regard this as an intuitively obvious thing to do. It becomes much more intuitive from a Bayesian point of view, as we mention in

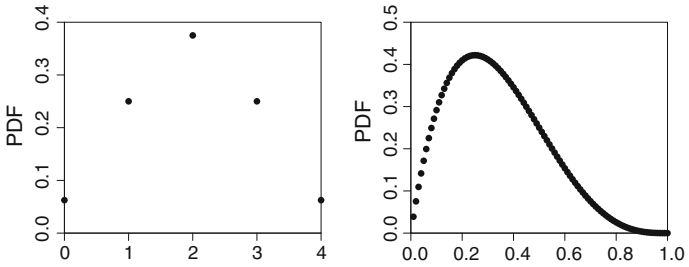

Fig. 7.2 Comparison of pdf $f(x|\theta)$ when viewed as a function of x with θ fixed at $\theta = .5$ (on *left*) or of θ with x fixed at $x = 1$ (on *right*). On the right-hand side, the pdf is evaluated for 99 equally-spaced values of θ from .01 to .99 .

Section 7.3.8. For now we ask the reader to bear with us and make sure to understand what we mean.

The distinction we are trying to draw here, between $f(x|\theta)$ as a function of x and $f(x|\theta)$ as a function of θ is illustrated in Fig. 7.2, which displays the binomial pdf viewed both ways when $n = 4$: first (on the left) as a function of x when $\theta = .5$ and then (on the right) as a function of θ when $x = 1$. First, when $\theta = .5$, the pdf is evaluated for five possible values of x: 0, 1, 2, 3, 4. These are all the possible values of x. (When $n = 4$, these are all the possible values of x regardless of the value of θ, as long as it is a permissible value, i.e., it is between 0 and 1, which is often written $\theta \in (0, 1)$.) When $x = 1$ and the pdf is regarded as a function of θ there is a whole continuum of possible values of θ in (0, 1). In the second part of the figure we set $x = 1$ and the pdf is evaluated for 99 values of θ, among all the possibilities for $\theta \in (0, 1)$. There is nothing of interest about the contrast between the picture on the left and the picture on the right *except* that the two representations are conceptually different.

When the pdf is considered as a function of the parameter θ rather than the values x of the random variable, it is called *the likelihood function*. We will denote it by $L(\theta)$. (Other notations are variations on this; all authors use some form of the letter "L.") The *maximum likelihood estimator* (*MLE*) is the value of θ that maximizes $L(\theta)$. We will denote it[2] by $\hat{\theta}$.

So far, we have discussed the pdf and likelihood based on a single (scalar) random variable. The concept generalizes immediately to vectors. In fact, one would typically have a vector of observed data $x = (x_1, \ldots, x_n)$ that has a joint pdf $f(x|\theta) = f(x_1, \ldots, x_n|\theta)$. In the subsequent parts of this chapter we will take x to be a vector, often corresponding to a sample of data, and regard as a special case any application when it becomes a scalar.

Note that the value of θ maximizing $L(\theta)$ is the same as the value of θ maximizing $c \cdot L(\theta)$ for any positive constant c. We therefore always understand the likelihood function to be defined only up to a positive constant. Thus, we may write $L(\theta)$ in

[2] There is some potential for confusion because, as we said on p. 152, in the literature the "hat" sometimes denotes a generic estimator and sometimes specifies the MLE.

proportionality form, using the proportionality symbol ($\propto$), as

$$L(\theta) \propto f(x|\theta)$$

and choose the constant for arithmetic convenience.

Illustration: Binomial likelihood We may write the binomial likelihood function as

$$L(\theta) = \theta^x (1-\theta)^{n-x}.$$

Here, in going from the pdf to the likelihood function we have omitted the factor $\binom{n}{x}$ because it does not involve θ. □

From the second part of Fig. 7.2 it is apparent that when $x = 1$ the MLE is $\hat{\theta} = .25$, which is an instance of the formula $\hat{\theta} = x/n$. To find the maximum, more generally, some combination of analytic (calculus-based) and numerical methods may be used. In the simplest problems, analytic methods suffice. In either case, however, it is easiest to begin by taking logs, because the value maximizing $\log L(\theta)$ is the same as the value maximizing $L(\theta)$, and because the pdf typically has a product form which is thereby converted to a sum. Suitably enough, the log of the likelihood function is called the *loglikelihood function*. We denote it here by $\ell(\theta)$:

$$\ell(\theta) = \log L(\theta).$$

Note that in writing a formula for $\ell(\theta)$ we may omit any additive terms that do not involve θ, because these become multiplicative constants in $L(\theta)$ and do not affect the maximization.

Illustration: Binomial MLE. To derive the general form $\hat{\theta} = x/n$ for the MLE we begin with the loglikelihood function

$$\ell(\theta) = x \log \theta + (n-x) \log(1-\theta)$$

where we have omitted the term $\log \binom{n}{x}$ because it does not involve θ. To maximize this function we set its derivative equal to zero and solve:

$$0 = \ell'(\theta) = \frac{x}{\theta} - \frac{n-x}{1-\theta}$$

so that

$$x(1-\theta) = (n-x)\theta$$

which gives the solution

$$\hat{\theta} = \frac{x}{n}.$$

It is also easy to check that $\ell''(\hat{\theta}) < 0$, which verifies that $\hat{\theta}$ is a maximum. □

Illustration: Normal MLE. Suppose we have a sample $x_1, \ldots, x_n$ from a $N(\theta, \sigma^2)$ distribution, where σ is known and the problem is to estimate θ. The ith normal density has pdf

$$f(x_i|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \theta)^2}{2\sigma^2}\right)$$

and the random variables $X_1, \ldots, X_n$ are independent, so the joint pdf is

$$\begin{aligned} f(x_1, \ldots, x_n|\theta) &= \prod_{i=1}^{n} f(x_i|\theta) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \theta)^2}{2\sigma^2}\right). \end{aligned}$$

From this, the loglikelihood function is

$$\begin{aligned} \ell(\theta) &= -\sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2\sigma^2} \\ &= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 - 2x_i\theta + \theta^2 \\ &= -\frac{n}{2\sigma^2}(\theta^2 - 2\bar{x}\theta) + R \end{aligned}$$

where R is a term that does not involve θ. Because the loglikelihood function is defined only up to an additive constant, we have

$$\ell(\theta) = -\frac{n}{2\sigma^2}(\theta^2 - 2\bar{x}\theta). \tag{7.2}$$

Setting its derivative equal to 0 we obtain

$$0 = \frac{n}{2\sigma^2}(\theta - \bar{x})$$

so that $\hat{\theta} = \bar{x}$. □

7.3 Confidence Intervals

7.3.1 For scientific inference, estimates are useless without some notion of precision.

In Example 1.4 P.S. preferred the non-burning house about 80 % of the time. However, this information by itself is not enough to say anything useful about her preferences: four out of five trials would also provide a preference for the non-burning house 80 % of the time, as would 80 out of 100 trials. But four out of five is far different than 80 out of 100. With 100 trials we could say pretty accurately what her preference rate is, while with four out of five it would not be clear that this is different than guessing. In scientific contexts, an estimate is useless unless we have some idea how accurate it is. One need not always drag around a standard error or confidence interval, and it is common to speak in terms of estimates without stating uncertainty; however, this convention assumes the uncertainty to be small relative to the size of the effects under discussion. It is important to include a statement of uncertainty whenever the uncertainty is non-negligible. In our judgment, inclusion of uncertainty should be considered the rule rather than the exception. We keep returning to Example 1.4 precisely because 14/17 is intermediate between the obvious situations where one doesn't need uncertainty (80/100) and where the estimate is hopelessly uncertain (4/5). Even a trained statistician might have some trouble saying correctly where 14/17 falls in this continuum without doing some calculations. So let us look at $14/17 = .82$ and ask, "How much error is there in this estimate"?

At first glance it appears impossible to answer this question: if we knew θ then the error in estimating it with $\hat{\theta}$ would be $\hat{\theta} - \theta$; but we *don't know* θ, which is why we are trying to estimate it. Nonetheless, even though we can not say precisely how big the error is, we can use probability and say something about *the likely magnitude of error*. This is usually quantified with the *standard error*. The idea begins with the recognition that every estimator $T = T(X_1, X_2, \ldots, X_n)$ exhibits variation. That is, if we were to examine T across many different samples we would get many different values. Because $X_1, \ldots, X_n$ are random variables having some probability distribution, T is a random variable. A simple summary of the magnitude of the variation of T is its standard deviation

$$\sigma_T = \sqrt{V(T)}. \tag{7.3}$$

This is almost, but not quite, the standard error of T. The problem with formula (7.3) is that $V(T)$ is typically not known and so itself must be estimated from the data. We illustrate in the context of Example 1.4.

Example 1.4 (Continued, see p. 13) Let $Y \sim B(n, p)$ and note that the usual estimator of p is the sample proportion $T = \hat{p} = Y/n$. Because $V(Y) = np(1-p)$ we have $V(T) = p(1-p)/n$. Thus, we have the formula

$$\sigma_T = \sqrt{\frac{p(1-p)}{n}}. \tag{7.4}$$

The formula in Eq. (7.4) quantifies the variation we can associate with the observed proportion $\hat{p} = 14/17 = .824$. However, we can not compute a numerical value for σ_T from Eq. (7.4) because we do not know what value of p to use. The obvious solution is to substitute $\hat{p}$ for p in Eq. (7.4). When we do this we obtain the *standard error* for the binomial proportion

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}. \tag{7.5}$$

Applying this to the data from P.S. we get

$$SE = \sqrt{\frac{\frac{14}{17}(1-\frac{14}{17})}{17}} = .092.$$

We then typically write the estimate in the form $.824 \pm .092$, with the $\pm$ indicating that the likely variability in the estimate is .092. When, instead, we write $\hat{p} \pm 2SE$ we get the confidence interval (.64, 1.0), reported on p. 13. □

The general procedure for computing the standard error is, in essence, the same as in the binomial case. To emphasize the substitution of the estimated parameter for the unknown parameter we define the *standard error* of an estimator T to be of the form

$$SE(T) = \sqrt{\hat{V}(T)} \tag{7.6}$$

with the hat on V indicating that we have estimated the variance. In fact, definition (7.6) is very general in the sense that it does not specify *how* we estimate the variance. As we will see in Chapters 8 and 9, several different methods are used to obtain variance estimates. We have used T in (7.6) to emphasize that it is a random variable, but in an alternative notation we use more often we may rewrite (7.6) as

$$SE(\hat{\theta}) = \sqrt{\hat{V}(\hat{\theta})}.$$

One note on terminology: the term "standard error" is sometimes used to refer to the standard error of the mean, as in Eq. (7.17), which is a special case of (7.6).

It is very common practice to report an estimate together with its standard error in the form

$$\hat{\theta} \pm SE(\hat{\theta}).$$

This gives a simple, rough sense of how accurate the estimate is. A more refined statement, made in terms of probability, comes from the use of a confidence interval:

a 95 % *confidence interval* (*CI*) for a parameter θ is an interval of the form (L, U) (L for lower, U for upper), where $L = L(X_1, \ldots, X_n)$ and $U = U(X_1, \ldots, X_n)$ are random variables derived from the data and

$$P(L < \theta < U) = .95. \tag{7.7}$$

This rather abstract definition becomes clear by examining particular problems, as we do below. In words, Eq. (7.7) says that if θ were the value of the unknown parameter, the probability that the interval would include this unknown value is 95 %. The probability .95 is the *level of confidence* associated with the interval (L, U).

In many applications an estimator $\hat{\theta}$ follows an approximately normal distribution (because estimators may often, at least approximately, be written in the form of the mean of some random variables). This is a tremendous simplification because it gives a simple method for finding L and U in (7.7). According to the 2/3–95 % rule (p. 117), from the approximate normality of $\hat{\theta}$ we may get an *approximate* 95 % confidence interval (L, U) by taking $L = \hat{\theta} - 2SE(\hat{\theta})$ and $U = \hat{\theta} + 2SE(\hat{\theta})$, that is,

$$\text{approx. } 95\,\%\ \text{CI} = (\hat{\theta} - 2SE(\hat{\theta}), \hat{\theta} + 2SE(\hat{\theta})). \tag{7.8}$$

The ingeniously simple construction that drives confidence intervals is most easily understood in the case of estimating the mean of a normal distribution, which we consider in Section 7.3.2. We then give some justification for the more general form in (7.8) on p. 166.

7.3.2 Estimation of a normal mean is a paradigm case.

Suppose $X_1, \ldots, X_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution with the value of σ known. Here, for notational ease, we drop the subscript X from μ and σ. Note that μ may be estimated by the sample mean $\bar{X}$ and in this special case $V(\bar{X}) = \sigma^2/n$ so that the standard error is

$$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}. \tag{7.9}$$

Theorem If $X_1, \ldots, X_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution, with the value of σ known, then

$$\bar{X} \sim N(\mu, (SE(\bar{X}))^2) \tag{7.10}$$

where $SE(\bar{X})$ is given by (7.9).

Proof: Let 1_{vec} be the n-dimensional vector with all components equal to 1. According to the definition of a random sample, the random variables in the sample are independent. Because $X_1, \ldots, X_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution, the vector $X = (X_1, \ldots, X_n)$ is, therefore, multivariate normal with mean $\mu 1_{vec}$ and

variance matrix $\sigma^2 I_n$ where I_n is the $n \times n$ identity matrix. Note that

$$\bar{X} = \frac{1}{n} 1_{vec}^T X. \tag{7.11}$$

From the definition of multivariate normality on p. 129 (which used Eqs. (4.25) and (4.26)) we have that $1_{vec}^T X$ is normally distributed with mean $1_{vec}^T \mu 1_{vec} = n\mu$ and variance $\sigma^2 1_{vec}^T I_n 1_{vec} = n\sigma^2$. Multiplying by $1/n$ and using (3.8) and (3.9), with $a = 1/n$ and $b = 0$, we have

$$\frac{1}{n} 1_{vec}^T X \sim N(\mu, \frac{\sigma^2}{n}). \tag{7.12}$$

Combining (7.12) with (7.11) gives

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \tag{7.13}$$

which is (7.9). □

Theorem If $X_1, \ldots, X_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution, with the value of σ known, then the interval $(\bar{X} - 2 \cdot SE(\bar{X}), \bar{X} + 2 \cdot SE(\bar{X}))$ is a 95 % CI for μ, where $SE(\bar{X})$ is given by (7.9).

Proof: We must show that

$$P(\bar{X} - 2 \cdot SE(\bar{X}) \leq \mu \leq \bar{X} + 2 \cdot SE(\bar{X})) = .95. \tag{7.14}$$

From (7.13) we have

$$P(\mu - 2\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + 2\frac{\sigma}{\sqrt{n}}) = .95. \tag{7.15}$$

We observe

$$\begin{aligned} \mu - 2\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + 2\frac{\sigma}{\sqrt{n}} &\iff |\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}| \leq 2 \\ &\iff \bar{X} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 2\frac{\sigma}{\sqrt{n}}. \end{aligned}$$

Therefore, (7.15) gives (7.14). □

The beauty of confidence lies in the simple manipulations, given above, that allow us to reason from (7.15) to (7.14). We take the description of variation given in (7.13) and convert it to a quantitative inference about the value of the unknown parameter μ.

7.3.3 For non-normal observations the central limit theorem may be invoked.

Now suppose $X_1, \ldots, X_n$ form a sample from a distribution with mean μ and standard deviation σ, with the distribution not necessarily normal. For simplicity, suppose again that σ is known.

By the CLT we have

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{D} N(0, 1).$$

We now apply the same manipulations used in deriving (7.15). We have

$$P\left(|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}| \leq 2\right) \approx .95$$

and, in turn, this is equivalent to

$$P\left(\bar{X} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 2\frac{\sigma}{\sqrt{n}}\right) \approx .95. \tag{7.16}$$

Therefore, for n sufficiently large, Eq. (7.16) provides an approximate 95 % CI. Written slightly differently, an approximate 95 % CI is given by $\bar{X} \pm 2 \cdot SE(\bar{X})$, where $SE(\bar{X}) = \sigma/\sqrt{n}$. The important point here is that we do not require the distribution of the data to be normal, yet we still get a quantitative inference based on asymptotic normality of the mean because of the CLT.

7.3.4 A large-sample confidence interval for μ is obtained using the standard error $s/\sqrt{n}$.

In Sections 7.3.2 and 7.3.3 we assumed σ was known. This was for purely pedagogical purposes. In practice, σ is almost always unknown and, as a consequence, we don't have a value to plug in when we want to calculate $SE = \sigma/\sqrt{n}$. The way to proceed, however, is pretty clear. As in the binomial standard error formula (7.5), we simply replace σ with an estimate, the obvious estimate being the sample standard deviation s. In the scenario envisioned in Section 7.3.3, with σ unknown we replace it with s in $\sigma/\sqrt{n}$ to get the *standard error of the mean*,

$$SE(\bar{x}) = \frac{s}{\sqrt{n}} \tag{7.17}$$

and from this we obtain a more practical version of (7.16) for our approximate 95 % CI. Because we state the result in terms of probability, we replace the observed value s with its random-variable counterpart S.

Result If $X_1, \ldots, X_n$ is a random sample from a distribution having mean μ and standard deviation σ, and n is sufficiently large, then an approximate 95 % CI for μ is given by $\bar{x} \pm 2 \cdot SE(\bar{x})$, where $SE(\bar{x})$ is given by (7.17), i.e., for n sufficiently large,

$$P(\bar{X} - 2\frac{S}{\sqrt{n}} < \mu < \bar{X} + 2\frac{S}{\sqrt{n}}) \approx .95. \tag{7.18}$$

This result follows from manipulations similar to those used in deriving (7.14) and (7.16). In establishing (7.16) we applied the CLT. The following theorem modifies the CLT used in Section 7.3.3 by replacing σ with S.

Theorem Suppose $X_1, \ldots, X_n$ is a random sample from a distribution having mean μ and standard deviation σ. Assume $E\left((X_i - \mu)^4\right) < \infty$, let S_n be the sample standard deviation calculated from $X_1, \ldots, X_n$, and let $Y_n = \sqrt{n}(\bar{X} - \mu)/S_n$. Then, as $n \to \infty$, we have

$$Y_n \xrightarrow{D} N(0, 1).$$

Details: In order to prove the theorem we first need two lemmas.

Lemma 1 Let $X_1, \ldots, X_n, \ldots$ be i.i.d. sequence for which $E((X_i - \mu)^4) < \infty$ and let S_n be the sample standard deviation calculated from $X_1, \ldots, X_n$. Then we have

$$S_n \xrightarrow{P} \sigma. \tag{7.19}$$

Proof: Let $Y_i = (X_i - \mu)^2$, so that $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Note that $E(Y_i) = \sigma_X^2$ $E(Y_i) = \sigma_x^2 = \sigma^2$ and, from (3.10), $V(Y_i) = E\left((X_i - \mu)^4\right) - \sigma^4$ which shows that $V(Y_i) < \infty$ so that the law of large numbers may be applied. By the law of large numbers we have that $\bar{Y}$ converges to σ^2. Because $n/(n-1) \to 1$, we also have that $\frac{n}{n-1}\bar{Y}$ converges to σ in probability. But $S_n = \frac{n}{n-1}\bar{Y}$. □

Lemma 2 (Slutsky's Theorem) If U_n converges to c in probability and V_n converges to Y in distribution, then $U_n V_n$ converges to cY in distribution.

Proof: The proof of this result, while straightforward, involves quite a bit of detailed manipulation. We omit it. (See Bickel and Doksum (2001, Theorem A.14.9).) □

Proof of Theorem: By the CLT $Z_n = \sqrt{n}(\bar{X} - \mu)/\sigma$ converges in distribution to $N(0, 1)$. Applying Lemma 1 we have that S_n converges to σ in probability or, equivalently, σ/S_n converges to 1 in probability. Writing $U_n = \sigma/S_n$ and $V_n = Z_n$, and noting that Y_n defined in the statement of the theorem satisfies $Y_n = U_n V_n$, we may apply Lemma 2 to obtain the desired convergence in distribution. □

Example 3.4 (continued from p. 138) On p. 138 we considered spike counts from a motor cortical neuron across 60 trials, each spike count being recorded during a 600 millisecond interval. The mean spike count across the 60 trials was 13.63 spikes. Converting the counts to firing rates (by dividing by .6 s (seconds)), we get a mean of 22.72 spikes per second and a standard deviation of 7.17 spikes per second. This gives a standard error of

$$SE = \frac{7.17}{\sqrt{60}} = .93.$$

We might then report the firing rate of this neuron, under the particular experimental condition, to be 22.72 (±.93) spikes per second. An approximate 95 % confidence interval for the firing rate is then (20.8, 24.6) spikes per second. □

The result is tremendously important in practice. However, it leaves open the question of how large the sample must be in order for the approximation to be good, i.e., for the probability of coverage (the probability the interval will cover μ) to be nearly .95. There is no universal answer to this question. Because we have the exact result in (7.14), this approximation tends to be good for moderate-size samples when the data are nearly normal. It may not be very good in moderate-size samples with strongly non-normal data. This is why it is important to check normality. The small-sample case is more problematic. We return to it in Section 7.3.10.

7.3.5 Standard errors lead immediately to confidence intervals.

We now return to the general form for an approximate 95 % CI given by (7.8) and derive it. First we consider the special case of the binomial probability p. Recall that if $X_1, \ldots, X_n$ are Bernoulli trials with probability p, and if $Y = \sum_{i=1}^{n} X_i$, then $Y \sim B(n, p)$. We have $Y/n = \bar{X}$, $E(X_i) = p$ and $V(X_i) = p(1 - p)$ so the CLT gives

$$\frac{\sqrt{n}(\bar{X} - p)}{\sqrt{p(1-p)}} \xrightarrow{D} N(0, 1). \tag{7.20}$$

By the $\frac{2}{3}$–95 % rule (p. 117) this implies

$$P(-2 \leq \frac{\sqrt{n}(\bar{X} - p)}{\sqrt{p(1-p)}} \leq 2) \approx .95$$

and, multiplying through the inequalities by $\sqrt{\frac{p(1-p)}{n}}$, we have

$$P(\bar{X} - 2 \cdot \sqrt{\frac{p(1-p)}{n}} \leq p \leq \bar{X} + 2 \cdot \sqrt{\frac{p(1-p)}{n}}) \approx .95.$$

Here p is unknown. Using $\bar{X}$ as an estimator of p we replace p by $\bar{X}$ and get

$$P(\bar{X} - 2 \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \leq p \leq \bar{X} + 2 \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}) \approx .95 \tag{7.21}$$

which is (7.8) for the binomial case, where the standard error is given by (7.5). The replacement of p with $\hat{p}$ in the standard error formula is analogous to the replacement of σ with s in Section 7.3.4. The binomial case is sufficiently important that we state it formally, rewriting (7.21) in terms of $\hat{p}$, where $\hat{p} = \bar{X}$ so that the standard error is $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Result If $Y \sim B(n, p)$ then p may be estimated by $\hat{p} = Y/n$ with standard error $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. For large n, an approximate 95 % CI is given by

$$\hat{p} \pm 2 \cdot SE(\hat{p}),$$

meaning that for n sufficiently large we have

$$P(\hat{p} - 2 \cdot SE(\hat{p}) \leq p \leq \hat{p} + 2 \cdot SE(\hat{p})) \approx .95. \tag{7.22}$$

Details: To justify the replacement of p with $\hat{p}$ we first note that the LLN gives us

$$\bar{X} \xrightarrow{P} p.$$

Then, by Slutsky's Theorem (p. 163), $\bar{X}(1-\bar{X})$ converges to $p(1-p)$ in probability and, from (7.20), we have

$$\frac{\sqrt{n}(\bar{X} - p)}{\sqrt{\bar{X}(1-\bar{X})}} \xrightarrow{D} N(0, 1)$$

which gives (7.21).

□

To generalize this argument we consider the problem of estimating a parameter vector θ in some statistical model using an estimator $T_n = T(X_1, \ldots, X_n)$. We

have written the subscript n on T to indicate that we are examining its behavior as $n \to \infty$. Two things drove the derivation of (7.22) above. First, the CLT was invoked to produce the approximate normality of $\bar{X}$ according to (7.20) and, second, in the standard deviation $\sqrt{\frac{p(1-p)}{n}}$, p was replaced by $\hat{p}$ (which was justified by the convergence of $\bar{X}$ to p in probability). If we assume these two phenomena apply, then we obtain (7.8) according to the following theorem.

Theorem If T_n is an asymptotically normal estimator of θ satisfying

$$\frac{T_n - \theta}{\sigma_{T_n}} \xrightarrow{D} N(0, 1)$$

and $\hat{\sigma}_{T_n}$ satisfies

$$\frac{\hat{\sigma}_{T_n}}{\sigma_{T_n}} \xrightarrow{P} 1$$

then we have

$$\frac{T_n - \theta}{\hat{\sigma}_{T_n}} \xrightarrow{D} N(0, 1).$$

Proof: This follows by Slutsky's theorem (p. 163), as in the binomial case. □

We now re-state the theorem as a "result", by putting it in a form that is less precise mathematically but more useful in practice.

Result If T_n is an asymptotically normal estimator of θ satisfying

$$\frac{T_n - \theta}{\sigma_{T_n}} \xrightarrow{D} N(0, 1) \tag{7.23}$$

and $\hat{\sigma}_{T_n}$ provides the standard error of T_n in the sense that

$$\frac{\hat{\sigma}_{T_n}}{\sigma_{T_n}} \xrightarrow{P} 1$$

then

$$\text{approx. } 95\,\% \text{ CI} = (T_n - 2\hat{\sigma}_{T_n}, T_n + 2\hat{\sigma}_{T_n})$$

which may also be written, equivalently, in the form (7.8), i.e.,

$$\text{approx. } 95\,\% \text{ CI } = (\hat{\theta} - 2SE(\hat{\theta}), \hat{\theta} + 2SE(\hat{\theta})).$$

The method given by (7.8) is widely applicable because (i) lots of estimators are approximately normally distributed, as in the first assumption of the theorem, and (ii) there are good ways to get standard errors, as in the second assumption of the theorem. The useful "result" is imprecise because of the approximation. The precise statement is in the theorem. This degree of imprecision, and the unclear relevance of arguments that treat the sample size n as sufficiently large, or essentially infinite, are core components of the bond between theory and practice in data analysis.

> *A Detail:* An additional consequence of (7.23) returns us to the characterization, on p. 158 of the standard error. After saying that the standard error represents the likely magnitude of error $T - \theta$ we then discussed standard error as estimating the standard deviation of T, which is not the same thing. It is in principle possible for the estimator T to be systematically wrong (being close to, say, $\theta + 10$ instead of θ) and yet have a small variance; in this case the standard error would not represent the likely magnitude of error. When (7.23) holds all is well: it says that $T - \theta$ is approximately normally distributed with mean 0 and approximate standard deviation σ_{T_n}, so that σ_{T_n} is indeed the likely magnitude of error. This notion of standard error is justified because (7.23) holds in a variety of commonly-found cases.
>
> □

An important kind of application of (7.8) arises when we have two parameters ϕ_1 and ϕ_2 and we are interested in the magnitude of their difference $\theta = \phi_1 - \phi_2$. If we have two independent estimators T_1 and T_2 (we could write T_{1,n_1} and T_{2,n_2} but are suppressing the dependence on the sample sizes n_1 and n_2) with standard errors SE_1 and SE_2 then

$$V(T_j) = SE_j^2$$

for $j = 1, 2$ and, by independence (see Eq. (4.4)),

$$V(T_1 - T_2) = SE_1^2 + SE_2^2,$$

and we get

$$SE(T_1 - T_2) = \sqrt{SE_1^2 + SE_2^2}. \tag{7.24}$$

This expression provides the standard error needed to produce a confidence interval for the difference $\theta = \phi_1 - \phi_2$, according to (7.8).

Example 7.2 Test-enhanced learning Tests are used to assess whether students have learned subject-matter material. A line of research has emphasized the additional value of testing as a way to *enhance* learning (Karpicke and Roediger 2008). The idea is that when students are tested, they recall information and thereby reinforce memory of it. In one study, Roediger and Karpicke (2006) had subjects read a short passage and then get tested on it after a delay period during which they would forget

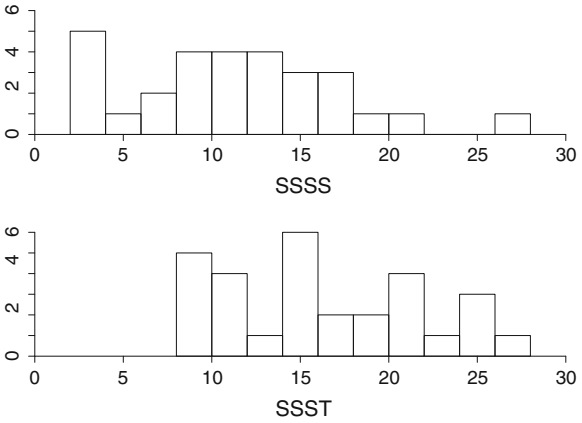

Fig. 7.3 Histograms of test-enhanced learning data. Data are assessment scores (number of recalled idea units, out of a maximum of 30) for 30 subjects under the SSSS condition (*top*) and the SSST condition (*bottom*). Data courtesy of J.D. Karpicke.

some of the material. Let us call this test the assessment test. After reading but before the assessment test there was an experimental manipulation: some subjects were asked to restudy the text, while other subjects were instead given a learning test, identical to the assessment test. These tests simply asked the subjects to write down everything they could remember about the passages. The tests were scored according to the number of "idea units" correctly recalled. A key part of the study focused on retention of the material following a delay period of 1 week, asking whether the learning-test group retained the material better than in the restudying group.

After finding strong evidence of a benefit from testing, the authors did a second experiment, using four study or testing sessions. In one condition, labelled SSSS, there were four study sessions, and in another, labelled SSST, there were three study sessions followed by a testing session. The assessment administered following a delay of 1 week had a maximal score of 30 idea units. Data from 60 subjects, 30 in each of the SSSS and SSST groups are displayed in Fig. 7.3.

For the data displayed in Fig. 7.3 the means were 11.9 and 16.7 idea units, with medians 12 and 16 idea units, and lower and upper quartiles (8.25, 15) and (11.25, 21) idea units. It appears that the SSST scores tend to be higher than the SSSS scores. To formalize the comparison, we consider the population mean scores under these two conditions. If we let X_{1i} be the score of the ith subject in the SSSS condition and X_{2i} be the score of the ith subject in the SSST condition and if μ_1 and μ_2 are the mean scores within these two conditions, we may estimate the difference $\theta = \mu_1 - \mu_2$. Applying (7.8) with (7.24) we first used (7.17) to obtain SE_1 and SE_2, and then (7.24) gave

$$SE(\bar{X}_1 - \bar{X}_2) = 1.5$$

idea units. We then found the approximate 95 % confidence interval to be

$$11.9 - 16.7 \pm 2(1.5) = -4.8 \pm 3.0$$

which produced the interval (1.8, 7.8) for the estimated mean number of additional idea units recalled in the SSST condition, compared with the SSSS condition. □

7.3.6 Estimates and standard errors should be reported to two digits in the standard error.

We recommend rounding standard errors to two leading (nonzero) digits, and then rounding the estimate to match the standard error. For example, if we found an estimate to be 5.582 and the standard error to be .207 we would report the result as $5.58 \pm .21$. Our reasoning is as follows. On the one hand, it is generally good to avoid too many digits both because numbers with many digits become hard to read, and also because extra digits may imply more accuracy than is present in the results. In this illustration, because the standard error is .21, the second digit in the estimate is already very uncertain: the 95 % CI is (5.2, 6.0) so we really don't know much about that second digit. We could report only a single digit in the standard error, but we prefer to report two because a standard error of .249 is quite a bit larger than a standard error of .151, yet to single-digit accuracy both would be rounded to .2. No rule is perfect, but it seems to us that reporting standard errors to two digits, but not more, is a good idea. Thus, in Example 1.4 on p. 159 we reported the estimate $\hat{p}$ of the propensity p to be $.824 \pm .092$, and in Example 3.4 on p. 164 we reported the firing rate of the M1 neuron to be $22.72 \pm .93$ spikes per second.

7.3.7 Appropriate sample sizes may be determined from desired size of standard error.

In Example 1.4, based on the confidence interval reported on p. 13, the results seemed conclusive but, in some situations, we would like even stronger evidence. A natural question is then, How much data would we need to achieve a decisive result? By assuming preliminary data give us a good idea of what to expect, we can answer this question. In the case of Example 1.4, we found $\hat{p} = .824$ with $SE = .092$. If we assume p is, in fact, somewhere around $\hat{p}$, the way we would obtain stronger evidence is by decreasing the standard error. In general terms we proceed in two steps. First, we determine how small we want the standard error to be. Writing our current standard error as SE_1 and our desired standard error as SE_2, we then write an expression that tells us how big a sample size we would need in order to reduce SE_1 to SE_2.

The key extra assumption is that the standard error tends to decrease as $\sqrt{n}$. This holds for many estimators, including MLEs (which follows from the discussion in Section 8.4.3). Let us suppose that SE_1 is based on a sample of size n_1 and we wish to determine the sample size n_2 that would give us SE_2. Because we want the standard error SE_1 to decrease by a factor SE_1/SE_2 (e.g., if we want SE_2 to be half the size of SE_1 we want to decrease SE_1 by a factor of 2), we write

$$\frac{SE_1}{SE_2} = \sqrt{\frac{n_2}{n_1}}$$

and solve for n_2, which gives

$$n_2 = n_1 \left(\frac{SE_1}{SE_2}\right)^2. \tag{7.25}$$

If, for instance, we wanted to decrease the standard error by a factor of 2 we would have to multiply our current sample size by a factor of 4. This is just a restatement of the $\sqrt{n}$ decrease in the standard error, with (7.25) providing the explicit formula we would use to compute n_2 in practice.

Using confidence intervals, the simple rule[3] in Eq. (7.25) is about as far as we can go. An investigator may wonder about step one, the choice of the "desired" SE_2. The selection of SE_2 must be determined by careful thinking about the scientific issues involved in the particular case at hand. The desired size of the standard error in Example 3.4, p. 164, for instance, depends on the way the information about spike counts will be used as part of the overall project. In Example 3.4 a relatively large number of trials were collected because the experiment was part of a comparative study in which relatively small differences across conditions appeared possible—yet still would have been of interest. According to the standard error on p. 164, the firing rate was determined within about ± 1 spike per second. If 15 trials had been used instead of 60, according to the $\sqrt{n}$ law and (7.25) we would expect an accuracy of about ± 2 spikes per second, which may or may not have seemed adequate.

7.3.8 Confidence assigns probability indirectly, making its interpretation subtle.

Here are two interpretations of the confidence interval found for the propensity p of P.S. to choose the non-burning house:

[3] More complicated formulas exist; however, the uncertainties involved in replicating results when collecting more data are often much larger than any extra precision one might gain from a more detailed calculation.

Interpretation A: If p were the true value, then the probability that the interval given by (7.22) would contain p is approximately 95 %. Based on the data from P.S., the approximate 95 % CI is (.64, 1.0).

Interpretation B: Based on the data from P.S., the probability that (.64, 1.0) contains p is approximately 95 %.

It may seem that interpretation B is an immediate consequence of interpretation A. After all, once we apply interpretation A to all values of p, then, regardless of the data we observe, the CI will cover p with approximately 95 % probability; we need only apply this to the data we actually did observe to get interpretation B. Unfortunately, to the shock and dismay of many students of statistical inference, this simple logic is fallacious. Interpretation B is a famously incorrect interpretation of a confidence interval. The correct interpretation of confidence, in interpretation A, can *not* be translated into interpretation B because interpretation A involves the *random variables* L and U that specify the lower and upper endpoints of the CI; probability concerns random variables, not constants; and in interpretation B, .64 and 1.0 are constants, they are not random variables. Once the data have been observed, the probability formalism at the foundation of (7.22) no longer speaks. So it is *incorrect* to think that the confidence interval (.64, 1.0) tells us the probability that p is in the range (.64, 1.0) is approximately 95 %. The math involved in deriving confidence intervals is clear, neat and clean. If we want to provide a linguistic interpretation of the confidence interval, however, we must revert to the somewhat clumsy and indirect interpretation A. On p. 175 we give a more careful re-statement of interpretations A and B.

To highlight the meaning of CIs let us consider the blindsight example further.

Example 1.4 (continued, see p. 13) The first three columns of the table below gives possible CIs using (7.22) when $X \sim B(17, p)$. For example, when $X = 11$ we find $L = .415$ and $U = .879$ so that the CI becomes (.42, .88).

x	L	U	Cover
7	.17	.65	N
8	.23	.71	N
9	.29	.77	N
10	.35	.83	Y
11	.42	.88	Y
12	.49	.93	Y
13	.56	.97	Y
14	.64	1.01	Y
15	.73	1.04	Y
16	.83	1.06	N
17	1	1	N

Now suppose the true value of p were .8. We would find that the CI would contain or "cover" p for some of the values of x but not others, as indicated in the fourth column of the table ("Y" for yes, the interval (L, U) covers .8, "N" for no it does

not). The table shows that (L, U) covers .8 when $10 \leq x \leq 15$. To find the level of confidence associated with (L, U) we may compute $P(10 \leq X \leq 15)$ when $X \sim B(17, .8)$. We find $P(10 \leq X \leq 15) = .871$, which says that the approximate 95 % CI found from (7.22) has probability .87 of containing the true value .8. The value .87 is a little smaller than the probability of .95 the interval would have if it were an exact, as opposed to an approximate CI. We discuss this further on p. 175. Here, the point is that the probability attached to the CI refers to the theoretical calculation based on drawing an observation X from a $B(17, p)$ distribution, as in interpretation A, rather than referring to the probability that p lies in the specific CI that was found from the data $x = 14$, as in interpretation B. □

There is another way to look at confidence intervals. Suppose we draw N random samples, independently, and compute CIs (L, U) for each. Let $Y_i = 1$ if (L, U) contains p for the ith random sample and $Y_i = 0$ if not, so that $P(L \leq p \leq U) = P(Y_i = 1)$. Then $\bar{Y}$ is the fraction of random samples for which (L, U) contains p. By the LLN,

$$\bar{Y} \xrightarrow{P} P(Y_i = 1)$$

that is,

$$\bar{Y} \xrightarrow{P} P(L \leq p \leq U).$$

We may therefore consider the confidence level $P(L \leq p \leq U)$ to be the long-run limit of the fraction of confidence intervals that contain p.

> *Interpretation C:* If we were to obtain CIs using (7.22) repeatedly, indefinitely many times, then, in the long run, approximately 95 % of those CIs would contain p. Based on the data from P.S., the CI is (.64, 1.0).

More generally, the level of confidence is usually considered to be the long-run frequency with which the CI covers the true value. For this reason, level of confidence is often called a *frequentist* property of a CI.

The big achievement of confidence intervals is the use of probability as a description of variation (the distribution $X \sim B(n, p)$) to suggest values of a parameter that are plausible in light of the data. However, this achievement comes at a cost: the formal statement is very weak, as it only calibrates the variability of interval (interpretations A and C). We might prefer interpretation B, which is analogous to saying "I am 90 % sure the capital of Louisiana is Baton Rouge", but, strictly speaking, confidence intervals do not allow such a statement. At best we might regard a confidence interval as a heuristic suggestion of uncertain knowledge. An alternative approach, based on Bayes' Theorem, *does* allow the more direct interpretation B. As we will see in Section 7.3.9, it has its own cost.

7.3.9 *Bayes' theorem may be used to assess uncertainty.*

Recall Bayes' Theorem for random variables and vectors, given in Section 4.3.3. From Eq. (4.36), for continuous random variables or vectors U and V we have

$$f_{U|V}(u|v) = \frac{f_{V|U}(v|u)f_U(u)}{\int f_{V|U}(v|u)f_U(u)du}. \tag{7.26}$$

Let us apply this to the problem of estimating the binomial parameter p. In this section we replace p by θ, so we suppose $X \sim B(n, \theta)$. To apply (7.26) we take $U = \theta$ and $V = X$ to get

$$f_{\theta|x}(\theta|x) = \frac{f_{X|\theta}(x|\theta)f_\theta(\theta)}{\int f_{X|\theta}(x|\theta)f_\theta(\theta)d\theta}. \tag{7.27}$$

(We use θ for both capital and lower case theta.) Ordinarily we would take θ as a known constant. Here, however, we acknowledge that θ is uncertain by considering it to be a *random variable* and assigning it a probability distribution. We take $f_\theta(\theta)$ to be the pdf representing our knowledge before seeing the data. It is the pdf corresponding to the *prior distribution*. When we treat θ as a known constant it is implicitly part of the binomial pdf, so we write the binomial pdf as $f_X(x)$. Here, however, the binomial pdf must be determined *conditionally* on a value of θ, so it is written $f_{X|\theta}(x|\theta)$. The pdf that summarizes our knowledge *after observing the data* $X = x$ is $f_{\theta|X}(\theta|x)$. This is the pdf corresponding to the *posterior distribution*. It is common to write the prior pdf as $\pi(\theta) = f_\theta(\theta)$ (this special notation makes it clear where the prior appears in various equations) and, because the likelihood function is $L(\theta) \propto f_{X|\theta}(x|\theta)$, the posterior pdf may be written

$$f_{\theta|x}(\theta|x) = \frac{L(\theta)\pi(\theta)}{\int L(\theta)\pi(\theta)d\theta}. \tag{7.28}$$

In order to do computations we must assign a specific probability distribution as the prior distribution. Assuming we know very little about the value of θ *a priori*, a natural choice is to use the uniform distribution, $\theta \sim U(0, 1)$, i.e., $f_\theta(\theta) = 1$. With this prior pdf we obtain

$$f(\theta|x) = \frac{\binom{n}{x}\theta^x(1-\theta)^{n-x} \cdot 1}{\int \binom{n}{x}\theta^x(1-\theta)^{n-x} \cdot 1 d\theta}$$

which reduces to

$$f(\theta|x) = \frac{\theta^x(1-\theta)^{n-x}}{\int \theta^x(1-\theta)^{n-x}d\theta}. \tag{7.29}$$

This formula is a special case of a *beta* distribution introduced briefly in Chapter 5: from Eq. (5.15), the $Beta(\alpha, \beta)$ pdf is

$$f(w) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} w^{\alpha-1}(1-w)^{\beta-1}. \tag{7.30}$$

Therefore, the posterior distribution of θ is $Beta(x+1, n-x+1)$ which has mean and standard deviation

$$\mu_{\theta|x} = \frac{x+1}{n+2}$$
$$\sigma_{\theta|x} = \sqrt{\frac{(x+1)(n-x+1)}{(n+2)^2(n+3)}}.$$

Example 1.4 (continued from p. 171) Let us apply this to the data from patient P.S. From (7.29) and (7.30) we have just found the posterior distribution to be $Beta(x+1, n-x+1)$. Here $n = 17$ and $x = 14$ so the posterior distribution is $Beta(15, 4)$ and the posterior mean and standard deviation are $\mu_{\theta|x} = .79$ and $\sigma_{\theta|x} = .091$. Thus, roughly speaking, these data lead us to conclude that the frequency with which P.S. will prefer the non-burning house is approximately .79 and our uncertainty may be summarized by saying that the average amount by which this guess misses the truth is approximately .091. These numbers are similar to those obtained in earlier analyses of the data from patient P.S., but here they have a different interpretation. Before giving this interpretation let us press on. We may obtain an interval having 95 % posterior probability from the .025 and .975 quantiles (the 2.5 and 97.5 percentiles) of the $Beta(15, 4)$ distribution, which gives (.59, .94). That is, $P(\theta < .59|y) = P(\theta > .94|x) = .025$ so that $P(.59 < \theta < .94|x) = .95$. The posterior interval (.59, .94) is sometimes called a *credible interval* to distinguish it from a confidence interval. Credible intervals are based on posterior distributions, whereas confidence intervals may be obtained from other arguments. The interval (.59, .94) is a succinct summary of what we know about θ based on the data. It is close to, but a little different than, the approximate 95 % CI of (.64, 1.0), which was obtained from (7.22). □

It is now legitimate to say what the posterior interval means, using words that are in essence just like interpretation B of Section 7.3.8.

Bayesian interpretation: Based on the data from P.S., together with the uniform prior, the probability that (.59, .94) contains θ is 95 %.

The use of Bayes' Theorem has thus bought us a highly intuitive interpretation of the credible interval. Like confidence intervals, credible intervals convert probability as a description of variation (the distribution $X \sim B(n, p)$) into a statement of knowledge. In this case, unlike the indirect situation with confidence intervals, the Bayesian statement is very much analogous to saying "I am 90 % sure the capital of Louisiana is Baton Rouge."

The straightforward Bayesian interpretation is very appealing. We issue two notes of caution. First, as we said at the end of Section 7.3.8, Bayes' Theorem requires the additional assumption of a particular form for the prior distribution. For the binomial problem it makes a good deal of sense to use the $U(0, 1)$ distribution for θ *a priori*.

In many settings, however, it is not clear what prior distribution should be used. Secondly, while confidence is undeniably less direct than posterior probability, we must keep in mind the fundamental distinction between the theoretical world of random variables and formal inferences, and the real world of data. There remains a degree of indirectness in the Bayesian statements as well, because they always say it is *as if* the data *were* to arise as random variables following the probability model (e.g., the binomial distribution). There is an inescapable divide between theoretical inferences and real-world conclusions; they are not quite the same thing, no matter what approach we take. Thus, the following elaborations to interpretations *A* and *B* on p. 171 would be more complete:

> *Interpretation A:* If we were to draw a random sample of $n = 17$ Bernoulli trials with parameter p, then the probability that the interval given by (7.22) would contain p is approximately 95 %. This is a theoretical statement. Assuming the theoretical and real worlds are aligned well, "the approximate 95 % CI is (.64, 1.0)" is a useful statement of knowledge.
>
> *Interpretation B:* If we were to draw a random sample of $n = 17$ Bernoulli trials with parameter p, and if we were to obtain $\hat{p} = 14/17$, then the probability that (.64, 1.0) contained p would be approximately 95 %. This is a theoretical statement. Assuming the theoretical and real worlds are aligned well, "the probability that (.64, 1.0) contains p is approximately 95 %" is a useful statement of knowledge.

Statistical methods that apply Bayes' theorem are usually called *Bayesian* and those that do not are usually called "frequentist", because of the frequency interpretation given on p. 172. Both Bayesian and frequentist methods have been applied in a wide range of data analysis problems. The form of the problem and the predilections of the practitioner dictate which approach is taken and, sometimes, both approaches appear within a single scientific article. It is widely recognized that Bayesian procedures should have good frequentist properties; for example, Bayesian 95 % credible intervals should have close to 95 % frequentist coverage probability, as they often do. We return to Example 1.4 to illustrate this.

Example 1.4 (continued from p. 171) We calculate the frequentist coverage probability of the posterior credible intervals obtained by the method on p. 174. To do this we apply the same reasoning used previously on p. 171, where we computed the coverage probability of the approximate CI based on (7.22). Note first that if we were to observe a value x from $X \sim B(17, p)$ we would obtain a $Beta(x + 1, 17 - x + 1)$ posterior distribution (according to (7.29) and (7.30)). The second and third columns of the table below give the resulting possible credible intervals using .025 and .975 quantiles of the $Beta(x + 1, 17 - x + 1)$ distribution, labeled $q_{.025}$ and $q_{.975}$.

We again suppose $p = .8$. From this table we find that the Bayesian credible intervals would cover the true value of $p = .8$ when $11 \leq x \leq 16$ (again indicated by "Y" for "yes" in the last column). To find the level of confidence associated with the credible intervals we compute $P(11 \leq X \leq 16)$ when $X \sim B(17, .8)$. We find

x	$q_{.025}$	$q_{.975}$	Cover
7	.22	.64	N
8	.26	.69	N
9	.31	.74	N
10	.36	.78	N
11	.41	.83	Y
12	.47	.87	Y
13	.52	.90	Y
14	.59	.94	Y
15	.65	.96	Y
16	.73	.99	Y
17	.81	1	N

$P(11 \leq X \leq 16) = .94$, which says that these credible intervals have probability .94 of containing the true value .8. This is very nearly equal to the desired value of .95, and is much closer to .95 than the value of .87 obtained on p. 171 for the approximate CI. The discrepancy between the putative value .95 and the correct coverage probability .87 for the approximate CI is due to the small sample size ($n = 17$). As the sample size gets large, the approximate 95 % CI found from (7.22) will have very nearly probability .95 of covering the true value of p. The Bayesian method performs better in this small-sample setting. When sample sizes are relatively small it is often possible to study coverage probabilities numerically in order to determine whether they are likely to be performing according to specifications, at least approximately. □

There are many important theoretical results concerning posterior distributions. In particular, the approximate CIs given by (7.22) have a Bayesian justification for large samples (see Section 8.3.3), making valid interpretation B of Section 7.3.8, which is re-phrased above. We return to Bayesian methods in Chapter 16.

7.3.10 For small samples it is customary to use the t distribution instead of the normal.

When the sample size is small, the approximation (7.18) may not be accurate. An alternative is to derive an "exact" confidence interval analogous to (7.14) that corrects for the substitution of s for σ. This leads to an adjustment of the multiplier put in front of the standard error. The adjustment to the small-sample CI uses the t distribution. Recall from Chapter 5 that if $U \sim N(0, 1)$ and $V \sim \chi^2_\nu$ independently then

$$W = \frac{U}{\sqrt{\frac{V}{n}}}$$

has a t distribution on ν degrees of freedom. In the context of a single batch of numbers, $\nu = n - 1$.

Note first that if $Z \sim N(0, 1)$ then $P(Z \leq 2) = .975$. In other words, 2 is the .975 quantile of the $N(0, 1)$ distribution. We now replace Z with W, which has a t distribution on ν degrees of freedom, and we write the .975 quantile of the t distribution on ν degrees of freedom as $t_{.975,\nu}$, i.e., $P(W \leq t_{.975,\nu}) = .975$. We then replace the value 2 in (7.18) with the somewhat larger value $t_{.975,\nu}$, so that $t_{.975,\nu}$ multiplies the standard error. The distributional result that makes this work is the following.

Theorem If $X_1, \ldots, X_n$ is a sample from a $N(\mu, \sigma^2)$ distribution, then $\bar{X}$ and S^2 are independent random variables with

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1)$$

and

$$\frac{S^2}{\sigma^2} \sim \chi^2_\nu$$

with $\nu = n - 1$.

Proof: We omit the proof of this theorem (which follows, with some effort, by manipulation of the joint pdf). □

Theorem If $X_1, \ldots, X_n$ is a sample from a $N(\mu, \sigma^2)$ distribution, then a 95 % CI is given by $\bar{x} \pm t_{.975,\nu} \cdot SE(\bar{x})$, where $\nu = n - 1$ and $SE(\bar{x})$ is given by (7.17), meaning

$$P(\bar{X} - t_{.975,n-1} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{.975,n-1} \cdot \frac{S}{\sqrt{n}}) = .95. \tag{7.31}$$

Proof: Let us write

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S} = \frac{\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}}{\sqrt{\frac{S^2}{\sigma^2}}}.$$

The previous theorem then gives the required t distribution of $\frac{\sqrt{n}(\bar{X} - \mu)}{S}$. □

Formula (7.31) is the standard method used by most statistical software to provide a confidence interval for an unknown mean μ. When the sample size is large, say, $n \geq 12$, then $t_{.975,\nu} \approx 2$ and (7.31) agrees with (7.16). Customary terminology refers to the CI in (7.31) as based on t (because the t distribution is used) while the CI in (7.16) is based on z (because the standard normal distribution is used). One would

not need to bother with the distinction between these two formulas unless $n < 12$, except that as a matter of convention (found in many journals, for example), there tends to be a preference for procedures based on a t, such as (7.31). In other words, it is worth being aware that many people say they are reporting t-based intervals as in (7.31) even when n is large and they might just as well say they are reporting (7.16)—there is in that case no practical distinction between the two.

Example 3.4 (continued from p. 164) Let us now consider the first 12 trials of counts from the motor cortical neuron, examined on p. 164. We get a mean firing rate of 24.31 spikes per second, and a standard deviation of 5.20 spikes per second, giving a standard error of

$$SE = \frac{5.20}{\sqrt{12}} = 1.50$$

spikes per second. The t_ν-based CI uses $\nu = 12 - 1 = 11$ and we find $t_{.975,11} = 2.20$. For the 95 % CI we take $L = 24.31 - 2.20(1.5) = 21.0$ and $U = 24.31 + 2.20(1.5) = 27.6$, giving us the CI (21.0, 27.6) spikes per second. □

It is also worth emphasizing a fundamental difficulty with this approach. The cases in which (7.31) differs from (7.16) are those in which n is small. But in such situations it is quite hard to tell whether the sample is really close to being normal. Application of (7.31) based on small samples should be considered only rough guides to evaluation of uncertainty.

Chapter 8
Estimation in Theory and Practice

In Section 7.2.1 we showed how the method of moments may be used to estimate the parameters of a *Gamma*(α, β) distribution, and we immediately stated that the method of maximum likelihood provides a better solution. How do we know this? In general, how should alternative methods of estimation be compared? In this chapter we lay out a series of principles that serve as guides to practice. The main ideas came from Ronald Fisher (1922); they were modified and made more precise by Jerzy Neyman (1937), and have been refined and incorporated into textbooks on statistical theory ever since, beginning notably with Cramér (1946).

Suppose we have a family of probability distributions that depends on a parameter θ, which must be estimated, and we have an estimator T. For now let us assume that θ is a scalar. If we were to say that T is a good estimator of θ, what might we mean? In particular, what might we mean when we say that maximum likelihood produces a good estimator? Clearly, for T to be a good estimator it must be "close" to θ, but because T is a random variable the notion of closeness must be stated probabilistically. For example, if we consider the mean $\bar{X}$ of a random sample $X_1, \ldots, X_n$ from a $N(\theta, 1)$ distribution, we might want to say that the mean $\bar{X}$ is close to θ when $|\bar{X}-\theta| < .1$. Because $\bar{X} \sim N(\theta, 1/n)$, even if n is large it is *possible* that $|\bar{X}-\theta| > .1$. We can not say that $|\bar{X} - \theta| < .1$. All we can say is the probability that $|\bar{X} - \theta| < .1$ is large, meaning close to one or, equivalently, the probability that $|\bar{X} - \theta| > .1$ is small, meaning close to zero.

For a general estimator T we can use the same approach and say that T is a good estimator of θ when it is *highly probable* that T is close to θ. Specifically, we introduce a tolerance ϵ, understanding that ϵ will be some small positive number, and then we require that $P(|\bar{X} - \theta| < \epsilon)$ is close to one or, equivalently, $P(|\bar{X} - \theta| > \epsilon)$ is close to zero. It is, in general, rather difficult to provide guarantees on the size of $P(|\bar{X} - \theta| > \epsilon)$ for fixed sample sizes. In most realistically complicated problems computer simulation studies must be used (as in Section 8.1.2) and they are based on specific cases so they do not provide general assurances. On the other hand, general results may be obtained asymptotically, letting the sample size grow indefinitely large. To take a concrete case, because the mean $\bar{X}$ of a random sample from a

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_8,

$N(\theta, 1)$ distribution follows a $N(\theta, 1/n)$ distribution, if we take n=10,000, from the normal cdf we find $P(|\bar{X} - \theta| > .1) = 1.5 \cdot 10^{-23}$. Indeed, no matter how small we take ϵ we have $P(|\bar{X} - \theta| > \epsilon) \to 0$ as $n \to \infty$. This is simply a restatement of the law of large numbers (p. 143)

$$\bar{X} \xrightarrow{P} \theta.$$

We discuss asymptotic results in Sections 8.2.1–8.3.1.

When we examine what happens as $n \to \infty$ it is helpful to write the generic estimator in the form $T_n = T(X_1, \ldots, X_n)$ to emphasize its dependence on n as we did in Section 7.3.5. One of the most important of the large-sample findings considers estimators that are *asymptotically normal*, as in Eq. (7.23),

$$\frac{T_n - \theta}{\sigma_{T_n}} \xrightarrow{D} N(0, 1). \tag{8.1}$$

For such estimators, in large samples, the probabilistic closeness of T_n to θ depends entirely on σ_{T_n} and we seek estimators that make σ_{T_n} as small as possible. In Sections 8.2.2–8.3.1 we go over the remarkable discovery by Fisher that σ_{T_n} can be minimized, and the minimum is obtained by the MLE. There has been a lot of theoretical work on the general subject of large-sample optimality, all of which leads to the conclusion that in well-behaved parametric problems, the method of maximum likelihood is essentially unbeatable. This, coupled with its very wide applicability (which began to be appreciated with the development of generalized linear models, see Section 14.1.6), has made maximum likelihood an essential tool in data analysis.

Fisher's theoretical insight seems to have been based on geometrical intuitions, which were elaborated in a mathematically rigorous framework by Bradley Efron in the 1970s and early 1980s. For details and references on the asymptotic arguments and their geometrical origins see Kass and Vos (1997). For a rigorous treatment in a more general context see van der Vaart (1998).

While asymptotic results are important, they have an inherent weakness: they apply when the sample size is large, but they do not say what "large" means in practice. In some cases $n = 20$ is more than adequate while in others n =20,000 is not large enough. One approach to coping with this problem is to evaluate a measure of likely deviation for specific cases, with specified sample sizes. The most common assessment of deviation of T from θ is the *mean squared error* (*MSE*) defined by

$$MSE(T) = E((T - \theta)^2). \tag{8.2}$$

In Chapter 4, p. 80, and 89, we considered the mean squared error in predicting one random variable from another. We discuss mean squared error in estimation in Section 8.1. In Section 8.4 we describe some of the practical considerations in applying ML estimation.

The most important points about ML estimation are the following:

1. ML estimation is applicable when the statistical model depends on an unknown parameter vector.[1] See Sections 7.2.2 and 8.4.1.
2. Together with ML estimates it is possible to get large-sample confidence intervals (Sections 8.2.2, 8.3.2, and 8.4.3).
3. In large samples, ML estimation is optimal (Section 8.3.1).
4. In large samples ML estimation agrees with Bayesian estimation (Section 8.3.3).

8.1 Mean Squared Error

The mean squared error criterion defined in (8.2) uses the squared magnitude of the deviation $T - \theta$ rather than its absolute value $|T - \theta|$ because it is easier to work with mathematically, and because it has a very nice decomposition given in Section 8.1.1. Intuitively, because $MSE(T)$ is an average of the values $(T - \theta)^2$, when $MSE(T)$ is small, large values of $(T - \theta)^2$ (and thus also large values of $|T - \theta|$) must be highly improbable. In fact, even more is true: we have

$$P(|T - \theta| > \epsilon) < \frac{E((T - \theta)^2)}{\epsilon^2}. \tag{8.3}$$

Thus, we can make sure it is highly probable for T to be close to θ by instead making sure that $MSE(T)$ is small.

> *Details:* We can use Markov's inequality, which appeared as a lemma in Section 6.2.1, to guarantee that $P(|T - \theta| > \epsilon)$ will be small if $MSE(T)$ is small. First, we have
>
> $$P(|T - \theta| > \epsilon) = P((T - \theta)^2 > \epsilon^2).$$
>
> Now, assuming $E((T - \theta)^2) < \infty$, Markov's inequality gives (8.3). □

In some cases $MSE(T)$ may be evaluated by analytical calculation, but in most practical situations computer simulation studies are used. We give two examples of such studies in Section 8.1.2.

8.1.1 Mean squared error is bias squared plus variance.

Two ways an estimator can perform poorly need to be distinguished. The first involves the systematic tendency for the estimator T to miss its target value θ. An estimator's

[1] The parameter must be finite-dimensional; in nonparametric inference the parameter is, instead, infinite-dimensional. Also, there are regularity conditions that make ML estimation work properly. See Bickel and Doksum (2001).

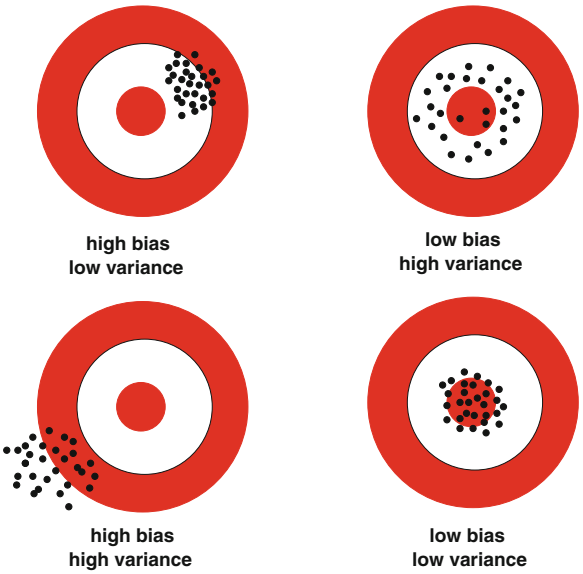

Fig. 8.1 Drawing of *shots* aimed at a target to illustrate the way estimates can miss their "target." They may be systematically biased, or they may have high variability, or both. The best situation, of course, is when there is little systematic bias and little variability.

bias is Bias(T) $=$ E(T) $- \theta$. When the bias is large, T will not be close to θ *on average*. The second is the variance $V(T)$. If $V(T)$ is large then T will rarely be close to θ. Figure 8.1 illustrates, by analogy with shooting at a bullseye target, the situations in which only the bias is large, only the variance is large, both are large (the worst case) and, finally, both are small (the best case). Part of the appeal of mean squared error is that it combines bias and variance in a beautifully simple way.

Theorem Suppose $E((T-\theta)^2) < \infty$. Then

$$E((T-\theta)^2) = (E(T-\theta))^2 + V(T).$$

That is,

$$MSE(T) = \text{Bias(T)}^2 + \text{Variance(T)}.$$

Proof: Let us write $\mu_T = E(T)$ and $T - \theta = (T - \mu_T) + (\mu_T - \theta)$, and then square both sides to get

$$(T-\theta)^2 = (T-\mu_T)^2 + 2(T-\mu_T)(\mu_T - \theta) + (\mu_T - \theta)^2.$$

Now consider taking the expectation of the cross-product term on the right-hand side. The quantity $\mu_T - \theta$ is a constant (it is not a random variable), while because $E(T) = \mu_T$, we have $E(T - \mu_T) = 0$ and,

therefore, $E(2(T-\mu_T)(\mu_T-\theta))=0$. Thus, we have

$$E((T-\theta)^2)=E((T-\mu_T)^2)+(E(\mu_T-\theta))^2$$

and, since $V(T)=E((T-\mu_T)^2)$, we have proven the theorem. □

This decomposition of MSE into squared bias and variance terms is used in various contexts to "tune" estimators in an attempt to decrease MSE. This typically involves some increase in one term, either the squared bias term or the variance term, in order to gain a larger decrease in the other term. Thus, reduction of MSE is often said to involve a *bias variance trade-off*. For an example, see p. 434.

Before we present an illustration of a *MSE* calculation, let us mention a property of the sample mean and sample variance. Assuming they are computed from a random sample $X_1,\ldots,X_n$, we have $E(\bar{X})=\mu_X$ which may be written

$$E(\bar{X})-\mu_X=0.$$

This says that, as an estimator of the theoretical mean, the sample mean has zero bias. When an estimator has zero bias it is called *unbiased*. If an estimator T is unbiased we have $MSE(T)=V(T)$ so that consideration of its performance may be based on a study of its variance.

In addition to the sample mean being unbiased as an estimator of the theoretical mean, it also happens that the *sample variance*, defined by

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2,$$

is unbiased as an estimator of the theoretical variance:

$$E(S^2)=\sigma_X^2. \tag{8.4}$$

Details: We wish to evaluate

$$E(S^2)=E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right)=\frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i-\bar{X})^2\right).$$

We write $X_i-\bar{X}=(X_i-\mu_X)+(\mu_X-\bar{X})$ and expand the square

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\bar{X})^2&=\sum_{i=1}^{n}\left((X_i-\mu_X)+(\mu_X-\bar{X})\right)^2\\&=\sum_{i=1}^{n}(X_i-\mu_X)^2+\sum_{i=1}^{n}2(X_i-\mu_X)(\mu_X-\bar{X})\end{aligned}$$

$$+\sum_{i=1}^{n}(\mu_X - \bar{X})^2.$$

We now rewrite the three terms in the last expression above. Because $E(X_i - \mu_X)^2 = \sigma_X^2$, and the expectation of a sum is the sum of the expectations, the first term has expectation

$$E\left(\sum_{i=1}^{n}(X_i - \mu_X)^2\right) = n\sigma_X^2. \tag{8.5}$$

Next, the second term may be rewritten

$$\begin{aligned}\sum_{i=1}^{n} 2(X_i - \mu_X)(\mu_X - \bar{X}) &= 2(\mu_X - \bar{X})\sum_{i=1}^{n}(X_i - \mu_X)\\ &= -2(\bar{X} - \mu_X)\sum_{i=1}^{n}(X_i - \mu_X)\\ &= -2n(\bar{X} - \mu_X)^2,\end{aligned}$$

where the last equality uses $\sum_{i=1}^{n}(X_i - \mu_X) = n(\bar{X} - \mu_X)$, and then, because $E((\bar{X} - \mu_X)^2) = V(\bar{X}) = \sigma_X^2/n$, the expectation of the second term becomes

$$E\left(\sum_{i=1}^{n} 2(X_i - \mu_X)(\mu_X - \bar{X})\right) = -2\sigma_X^2. \tag{8.6}$$

Finally, because again, $E((\bar{X} - \mu_X)^2) = \sigma_X^2/n$, the expectation of the third term is

$$E\left(\sum_{i=1}^{n}(\mu_X - \bar{X})^2)\right) = \sigma_X^2 \tag{8.7}$$

and, combining (8.5), (8.6), and (8.7) we get

$$E\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = (n-1)\sigma_X^2$$

which gives (8.4). □

We use the unbiasedness of the sample mean and sample variance in the following illustration of the way two estimators may be compared theoretically.

Illustration: Poisson Spike Counts On p. 164 we considered 60 spike counts from a motor cortical neuron and found an approximate 95 % CI for the resulting firing

rate using the sample mean. The justification for that approximate CI involved the CLT, and the practical implication was that as long as the sample size is fairly large, and the distribution not too far from normal, the CI would have approximately .95 probability of covering the theoretical mean. In this case, the spike counts do, indeed, appear not too far from normal. Sometimes they are assumed to be Poisson distributed. This is questionable because careful examination of spike trains almost always indicates some departure from the Poisson. On the other hand, the departure is sometimes not large enough to make a practical difference to results. In any case, for the sake of illustrating the *MSE* calculation, let us now *assume* the counts follow a Poisson distribution with mean λ. The sample mean $\bar{X}$ is a reasonable estimator of λ, but one might dream up alternatives. For example, a property of the Poisson distribution is that its variance is also equal to λ; therefore, the sample variance S^2 could also be used to estimate the theoretical variance λ. This may seem odd, and potentially inferior, on intuitive grounds because the whole point is to estimate the mean firing rate, not the variance of the firing rate. On the other hand, once we take the Poisson model seriously the theoretical mean and variance become equal and, from a statistical point of view, it is reasonable to ask whether it is better to estimate one rather than the other from their sample analogues. Our purpose here is to present a simple analysis that demonstrates the inferiority of the sample variance compared with the sample mean as an estimator of the Poisson mean λ. We are going through this exercise so that we can draw an analogy to it later on.

Now, because, as we mentioned immediately before beginning this illustration, $\bar{X}$ and S^2 are unbiased for the theoretical mean and variance they are, in this case, both unbiased as estimators of λ. As a consequence, $MSE(T) = V(T)$ for both $T = \bar{X}$ and $T = S^2$. Analytical calculation of the variance of these estimators (which we omit here) gives

$$V(\bar{X}) = \frac{\lambda}{n}$$
$$V(S^2) = \frac{\lambda}{n} + \frac{2\lambda^2}{n-1}$$

where n is the number of counts (the number of trials). Therefore, the *MSE* of S^2 is always larger than that of $\bar{X}$ so that S^2 tends to be further from the correct value of λ than $\bar{X}$. For example, if we take $n = 100$ trials and $\lambda = 10$, we find $V(\bar{X}) = .10$ while $V(S^2) = 2.12$. The estimator S^2 has about 21 times the variability as $\bar{X}$, so that estimating λ using S^2 would require about 2,100 trials of data to gain the same accuracy as using $\bar{X}$ with 100 trials. Figure 8.2 shows a pair of histograms of $\bar{X}$ and S^2 values calculated from 1,000 randomly-generated samples of size $n = 100$ when the true Poisson mean was $\lambda = 10$. The distribution represented by the histogram on the right is much wider. □

This illustration nicely shows how one method of estimation can be very much better than another, but it is admittedly somewhat artificial; because the distribution of real spike counts may well depart from Poisson, a careful comparison of $\bar{X}$ versus

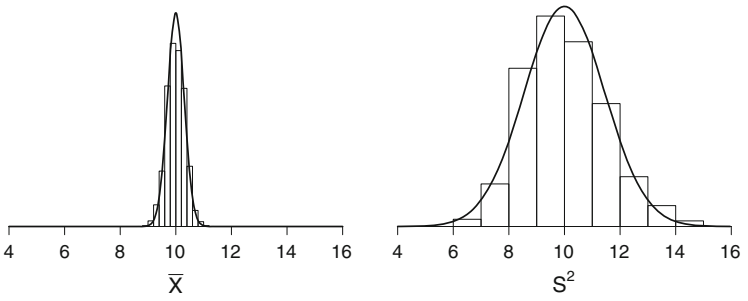

Fig. 8.2 Histograms displaying distributions of $\bar{X}$ and S^2 based on 1,000 randomly-generated samples of size $n = 100$ from a Poisson distribution with mean parameter $\mu = 10$. In these repeated samples both $\bar{X}$ and S^2 have distributions that are approximately normal. Both distributions are centered at 10 (both estimators are unbiased) but the values of S^2 fluctuate much more than do the values of $\bar{X}$.

S^2 should consider their behavior also under alternative assumptions. In this regard, the sample mean remains a reasonably good estimator of the theoretical mean in large samples regardless of the probability distribution of the spike counts. The sample variance, on the other hand, does so only if the theoretical variance is truly equal to the theoretical mean; otherwise, as the sample size increases it will converge to the wrong value. This is likely to be an important consideration. However, even if one were convinced that counts truly followed a Poisson distribution, the analysis above would be compelling. It would be grossly inefficient to use S^2 instead of $\bar{X}$ in estimating λ.

Another thing to notice in Fig. 8.2 is the approximately normal shape of the two histograms. Asymptotic normality of estimators is very common, and we have already relied on it in Section 7.3.5.

8.1.2 Mean squared error may be evaluated by computer simulation of pseudo-data.

In the Poisson spike count illustration on p. 184 we were able to compute the *MSE* exactly. In more complicated situations this is often impossible. Instead we rely on either large-sample arguments, such as those in Section 8.2.2, or numerical simulations. The numerical method uses computer-generated *pseudo-data*, by which we mean numbers or vectors that are generated from known probability distributions in order to mimic the behavior of data. Because the distribution is known, there is a known correct value of θ to which T may be compared.

Suppose we wish to compute $MSE(T)$ in estimating θ under the assumption that a random sample comes from a particular probability distribution having cdf $F(x)$. Assuming we know how to generate random samples from $F(x)$ on the computer, we may use this algorithm:

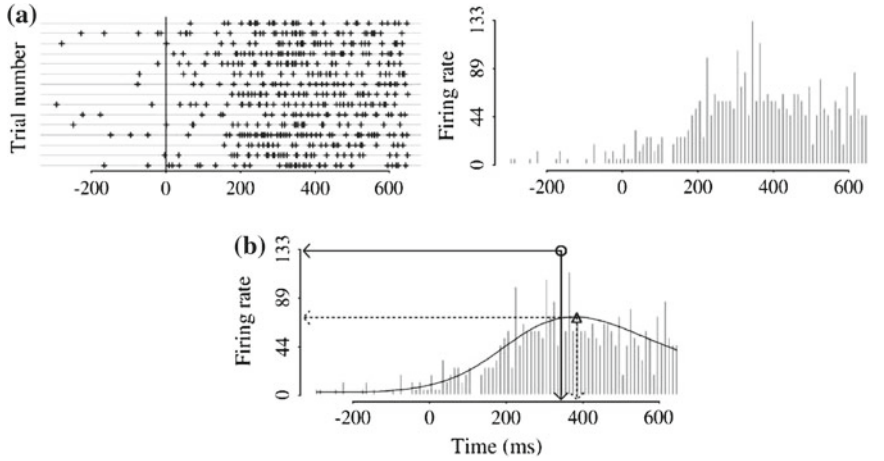

Fig. 8.3 Time of maximal firing rate. **a** displays a raster plot and Peri-Stimulus Time Histogram (PSTH). As explained in Chapter 1, the PSTH represents the firing rate as a function of time. **b** displays the time at which the maximal firing rate occurs, estimated (i) using the PSTH and (ii) using instead a smooth curve. Adapted from Kass et al. (2003).

1. Take G to be a large integer (such as 1,000) and for $g = 1, \ldots, G$ do the following:
   (i) Generate a random sample $X_1^{(g)}, \ldots, X_n^{(g)}$ from $F(x)$.
   (ii) Compute $T^{(g)} = T(X_1^{(1)}, \ldots, X_n^{(g)})$, which is the value of the estimator T based on the gth random sample.
   (iii) Let $Y_g = (T^{(g)} - \theta)^2$.
2. Compute

$$\bar{Y} = \frac{1}{G} \sum_{g=1}^{G} Y_g. \tag{8.8}$$

By the LLN, we have that $\bar{Y}$ converges to the desired $MSE = E((T - \theta)^2)$ in probability. Thus, we take $\bar{Y}$ as our *MSE*.

This kind of computation is used in the following illustration. It involves the statistical efficiency of smoothing, a topic we take up in Chapter 15. In presenting this now we omit details about the method.

Example 1.1 (continued, see p. 3) In Chapter 1 we discussed a study by Olson et al. (2000), in which neuronal spike trains were recorded from the supplementary eye field (SEF) under two different experimental conditions. As is usually the case in stimulus-response studies, the neuronal response—in this case, the firing rate—varied as a function time. For a particular neuron in one of the conditions, the PSTH in Fig. 8.3 displays the way the firing rate changes across time. The data analytic challenge in the Olson et al. study was to characterize the distinctions between the firing rate functions under the two experimental conditions. One of the distinctions,

evident in some of the plots, was that the maximal firing rate occurred somewhat later in one condition than in the other. How should this time of maximal firing rate be computed? One possibility is to use the PSTH, by finding the time bin for which the PSTH is maximized. Panel b of Fig. 8.3 displays the resulting solution: according to the PSTH shown there, the maximal firing rate of about 133 spikes/s occurs at a time marked by the arrow on the left along the time axis. However, this is clearly a noisy estimate. Slight variations in location of time bin, or width, would change this, as would consideration of new data from the same neuron. On the other hand, a second method based on first fitting a smooth curve to the PSTH and then finding its maximum, yields a different answer: the maximum firing rate of about 75 spikes/s (seconds) occurs at a time indicated by the arrow on the right along the time axis. This value is less subject to fluctuations in the data. If we assume that the theoretical firing rate is, in fact, slowly varying in time, then the smooth curve should provide a better estimate. Kass et al. (2003) used MSE to evaluate the extent to which smoothing improves estimation.

Kass et al. (2003) evaluated MSE for the true firing rate function shown in panel a of Fig. 8.4. To do so, they simulated, repeatedly, 16 trials of pseudo-data and then constructed histograms and also fit smooth curves (there are 16 trials in the SEF data shown in Fig. 8.3a). The PSTH and smooth curve from one sample of 16 trials of pseudo-data are shown in panel b of Fig. 8.4. The smoothing method used by Kass et al. (2003) involved regression splines, as discussed in Section 15.2.3. Note that the smooth curve ("estimated rate") is close to the true rate from the simulation, but it misses by a small amount due to the small number of trials we used in the simulation.

To quantify the deviation of both the PSTH and the smooth curve at any one point in time t the *MSE* could be used. That is, we would regard the true firing rate at time t as the value $\theta = \theta_t$ to be estimated, and we would compute $MSE(T) = MSE_t(T)$ when T is based on the PSTH and when T is based on the smooth curve. Here the subscript t is a reminder that we have chosen a particular time point. If $MSE_t(T)$ is evaluated for every time value t the total of all the mean squared errors may be found by integrating across time. This defines what is called the *integrated mean squared error* or *mean integrated squared error* (*MISE*),

$$MISE(T) = \int MSE_t(T)dt$$

where the integration is performed over the time interval of interest. The integral may be evaluated easily simply by calculating the *MSE* along a grid of time values separated by some increment Δt

$$\int MSE_t(T)dt \approx \Delta t \sum_t MSE_t(T).$$

In order to compute the *MSE* at each time value t Kass et al. (2003) used computer simulation: They generated data repeatedly, each time finding both the PSTH and the smooth curve. They simulated 1,000 data sets, each involving 16 randomly-generated

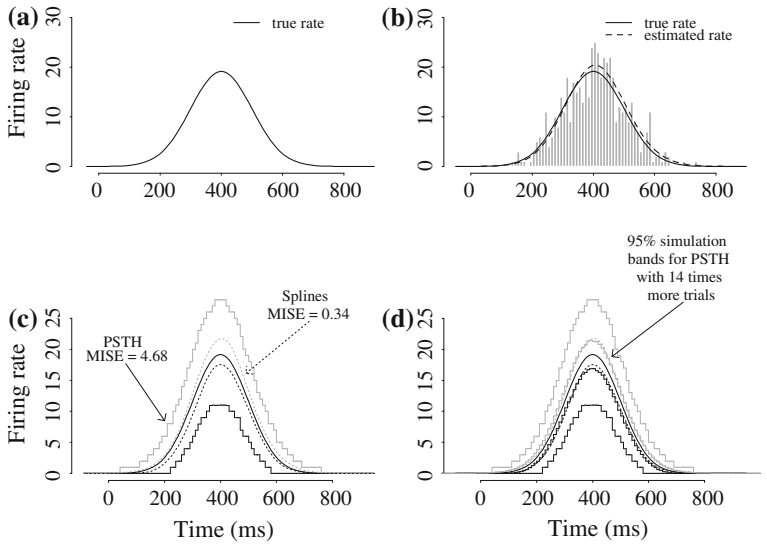

Fig. 8.4 **a** True rate from which 16 trials are simulated; their PSTH is shown in (**b**), with true and estimated firing rates overlaid. **c** shows the true rate and 95 % simulation bands obtained from smoothed and unsmoothed PSTHs. **d** shows the same curves as (**c**), as well as 95 % simulation bands obtained from unsmoothed PSTHs with 16×14 trials instead of 16. Adapted from Kass et al. (2003).

spike trains based on the true firing rate curve shown in Fig. 8.4a, and from these 1,000 data sets they computed the *MISE*. They also computed 95 % bands, within which fall 95 % of the estimated curves. Figure 8.4c shows the two pairs of bands, now labeled with the two values of MISE: the spline-based estimate has a MISE of .34 (in spikes/s squared) while the PSTH has a MISE of 4.68, which is 14 times larger. This means that when the PSTH is used to estimate firing rate, 14 times as much data are needed to achieve the same level of accuracy. Similarly, the 95 % bands for the PSTH are much further from the true firing-rate curve than the bands for the spline-based estimate. Figure 8.4d includes a pair of 95 % bands obtained from the PSTH when 224 trials are used rather than 16 (because $224 = 14 \times 16$). This is another way of showing that the accuracy in estimating the firing rate using spline smoothing based on 16 trials is the same as the accuracy using the PSTH based on 224 trials. Clearly it is very much better to use smoothing when estimating the instantaneous firing rate. □

A detail: One issue that arises in numerical simulation is the accuracy of the computational results, because the value $\bar{Y}$ in (8.8) is itself an estimate of the *MSE*. However, if G is large, the standard error of $\bar{Y}$ will be small. Furthermore, because $\bar{Y}$ is a sample mean, we can apply the method of Section 7.3.4 and use $s/\sqrt{G}$ as its standard error, where $s^2 = \frac{1}{G-1}\sum_{g=1}^{G}(Y_g - \bar{Y})^2$. The standard error lets us determine

whether G is adequately large. For instance, if we wish the MSE to be computed with accuracy δ, we can take G big enough to satisfy

$$\frac{s}{\sqrt{G}} < \frac{\delta}{2}.$$

By the result in Section 7.3.4, an approximate 95 % confidence interval for MSE would be $(\theta - \delta, \theta + \delta)$. Thus, we would have 95 % confidence that the desired accuracy was obtained. □

8.1.3 *In estimating a theoretical mean from observations having differing variances a weighted mean should be used, with weights inversely proportional to the variances.*

In the illustration on Poisson spike counts, p. 184, we used the *MSE* criterion to evaluate alternative estimators, based on an analytical expression. In that case both estimators were unbiased and the comparison was based on variance. Another illustration of this type arises when data are considered collectively across many similarly measured objects, such as neurons or subjects, with the observations from the different individuals contributing varying amounts of information; specifically, with the individual observations having different variances. In combining such discrepant observations, it is preferable not to use the sample mean, but instead to weight each observation according to the amount of information it contributes. Here we provide a theoretical analysis of this problem, and give the basic result.

Suppose we have two independent random variables X_i for $i = 1, 2$, with $E(X_1) = E(X_2) = \mu$ but $V(X_1) = \sigma_1^2$ and $V(X_2) = \sigma_2^2$, with the two variances possibly being different. After analyzing the two-observation case, we will present analogous results for n observations. Let us assume that σ_1 and σ_2 are known and ask how best to combine X_1 and X_2 linearly in order to estimate μ. We write a general weighted combination as

$$Y_w = w_1 \cdot X_1 + w_2 \cdot X_2 \tag{8.9}$$

where $w_1 + w_2 = 1$. It turns out that the optimal special case is

$$\bar{X}_w = w_1 \cdot X_1 + w_2 \cdot X_2 \tag{8.10}$$

where

$$w_i = \frac{\frac{1}{\sigma_i^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} \tag{8.11}$$

for $i = 1, 2$.

Theorem Suppose X_1 and X_2 are independent random variables with $E(X_1) = E(X_2) = \mu$ and $V(X_1) = \sigma_1^2$ and $V(X_2) = \sigma_2^2$, and let Y_w be defined as in (8.9). Then Y_w is unbiased, so that $MSE(Y_w) = V(Y_w)$, and this quantity is minimized among possible weighting pairs by taking $Y_w = \bar{X}_w$, i.e.,

$$V(\bar{X}_w) \leq V(Y_w)$$

or, equivalently,

$$MSE(\bar{X}_w) \leq MSE(Y_w)$$

with equality holding in both cases only if $Y_w = \bar{X}_w$ defined by (8.10) and (8.11).

Proof of Theorem: First, we have

$$\begin{aligned} E(Y_w) &= w_1 \cdot \mu + w_2 \cdot \mu \\ &= (w_1 + w_2)\mu \\ &= \mu. \end{aligned}$$

Thus, Y_w is unbiased and $MSE(Y_w) = V(Y_w)$. To derive the variance result we start with

$$V(w_1 \cdot X_1 + w_2 \cdot X_2) = w_1^2 \cdot \sigma_1^2 + w_2^2 \cdot \sigma_2^2.$$

Now we use $w_1 + w_2 = 1$ and replace w_2 with $1 - w_1$ to get

$$\begin{aligned} V(w_1 \cdot X_1 + w_2 \cdot X_2) &= w_1^2 \cdot \sigma_1^2 + (1 - w_1)^2 \cdot \sigma_2^2 \\ &= \sigma_1^2 w_1^2 + \sigma_2^2 - 2\sigma_2^2 w_1 + \sigma_2^2 w_1^2 \\ &= (\sigma_1^2 + \sigma_2^2) w_1^2 - 2\sigma_2^2 w_1 + \sigma_2^2. \end{aligned}$$

We now minimize this quantity by differentiating with respect to w_1, and setting the derivative equal to zero. We get

$$0 = 2(\sigma_1^2 + \sigma_2^2) w_1 - 2\sigma_2^2$$

and therefore

$$w_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

Dividing the numerator and denominator of this fraction by $\sigma_1^2 \sigma_2^2$ gives

$$w_1 = \frac{\frac{\sigma_2^2}{\sigma_1^2 \sigma_2^2}}{\frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2 \sigma_2^2}} = \frac{\frac{1}{\sigma_1^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}$$

which is the desired result. □

As an illustration, suppose we had 100 independent observations $U_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, 100$, and grouped them unequally defining, say, $X_1 = \frac{1}{10}\sum_{i=1}^{10} U_i$ and $X_2 = \frac{1}{90}\sum_{i=11}^{100} U_i$. It would seem strange to use $\frac{1}{2}(X_1 + X_2)$ in this situation and the intuitive thing to do would be to use the weighted mean: here the weights are $w_1 = 10/100$ and $w_2 = 90/100$ (because $\sigma_1^2 = \sigma^2/10$ and $\sigma_2^2 = \sigma^2/90$) so we get $\bar{X}_w = \bar{U}$.

One way to interpret this is to say that using $\bar{X}$ instead of $\bar{X}_w$ is like throwing away a fraction of the data. For example, suppose X_1 and X_2 both represent means of counts from n trials. If σ_1 is half the size of σ_2 then, from the formula above, the ratio of variances is 1.56. This means that to achieve the same accuracy in the estimator, n would have to be 56 % larger if we used the sample mean instead of the weighted mean. When σ_1 is one-third the size of σ_2 we would have to increase n by a factor of 2.78 (instead of 50 trials, say, we would need 139). In these cases we might say that the weighted mean is, respectively, 1.56 and 2.78 times more efficient than the ordinary sample mean.

Example 8.1 Optimal integration of sensory information Ernst and Banks (2002) considered whether humans might combine two kinds of sensory input optimally, according to (8.10) and (8.11). Subjects were presented with raised bars either visually or by touch (known as haptic input) and had to judge the height of each bar in comparison with a "standard" bar. The experimental apparatus was set up to allow visual or haptic noise to be added to the height of each bar. Subjects were also presented with both visual and haptic input simultaneously. The authors reported evidence that when presented with the simultaneous visual and haptic input, subjects judged heights by combining the two sensory modalities consistently with (8.10) and (8.11). In other words, this was evidence that humans can integrate distinct sensory inputs optimally. □

Here is the result for combining n observations. We have also included here the formula for the standard error of the weighted mean.

Theorem Suppose $X_1, \ldots, X_n$ are independent random variables with $E(X_1) = E(X_2) = \cdots = E(X_n) = \mu$ and $V(X_i) = \sigma_i^2$ for $i = 1, \ldots, n$. Let

$$\bar{X}_w = \sum_{i=1}^{n} w_i \cdot X_i \tag{8.12}$$

where, in (8.12),

$$w_i = \frac{\frac{1}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2}}$$

and for any set of weights $w_1, \ldots, w_n$ for which $\sum_{i=1}^{n} w_i = 1$ define

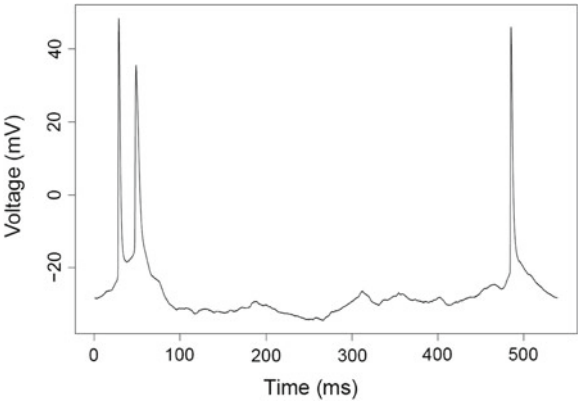

Fig. 8.5 When an action potential follows closely a previous action potential (with small ISI), the second action potential is broader than the first. When a long ISI intervenes, however, the second action potential is very similar to the first.

$$Y_w = \sum_{i=1}^{n} w_i \cdot X_i.$$

Then we have

$$V(\bar{X}_w) \leq V(Y_w)$$

with equality holding if and only if $Y_w = \bar{X}_w$. Furthermore we have

$$SE(\bar{X}_w) = \sqrt{V(\bar{X}_w)} \tag{8.13}$$

where

$$V(\bar{X}_w) = \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2} \right)^{-1}.$$

Proof: The proof is analogous to that for the case $n = 2$. □

Example 8.2 Action potential width and the preceding inter-spike interval As part of a study on the effects of seizure-induced neural activity (Shruti et al. 2008) spike trains were recorded from barrel cortex neurons in slice preparation. One of the interesting findings[2] involved the relationship between the width of each action potential (spike) and its preceding ISI. As is well known, when a spike follows closely on a preceding spike, so that the ISI is relatively short, then the second spike will tend to be wider than the first. If, however, the ISI is sufficiently long, there will not be any effect of the first spike on the second, and the spike widths should be roughly

[2] The results here were obtained by Judy Xi.

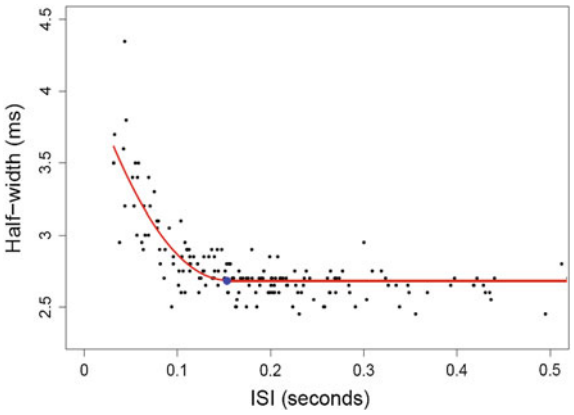

Fig. 8.6 Action potential width varies as function of previous ISI. The data are from many action potentials recorded for a single neuron. A *fitted curve* with a change point is also shown, the change point being indicated as a *large blue dot*.

equal. See (Fig. 8.5). How long is "sufficiently long?" This turns out to be dependent on previous neuronal activity.

Let Y be the spike width and x the preceding ISI length, and let us assume there is an ISI length τ such that, on average, Y is constant for all $x > \tau$. Among neurons taken from animals that had seizures, τ tended to be smaller than its value among control animals. Figure 8.6 displays some of the data, together with a fitted curve. The statistical model used for this curve assumes that, on average, Y decreases with x for $x < \tau$ but remains constant for $x \geq \tau$. In statistical jargon, τ is called a *change point*, because the relationship between Y and x changes at $x = \tau$. The relationship between y and x was assumed to be quadratic for $x < \tau$ (see Section 12.5.4) and constant for $x \geq \tau$. The model was fit using nonlinear least squares. Additional details are given on p. 408 in Section 14.2.1. The parametric bootstrap (Section 9.2.2) was then applied to obtain the $SE(\hat{\tau})$. The method was repeated for neurons from seizure and control animals to see whether there were systematic differences across the two treatment conditions. Figure 8.7 shows results for both groups. Note the very different standard errors across neurons. This suggests that in comparing the two groups it is advisable to use weighted means, as in Eq. (8.12), together with standard errors given by Eq. (8.13). The results were that the control group had weighted mean change point of 190 ($\pm$32) ms and the seizure group reset earlier, with weighted mean change point 108 ($\pm$.012) ms. □

Example 8.3 Neural response to selective perturbation of a brain-machine interface In order to study learning-related changes in a network of neurons, Jarosiewicz et al. (2008) introduced a paradigm in which the output of a cortical network can be perturbed directly and the neural basis of the compensatory changes studied in detail. Using a brain-computer interface (BCI), dozens of simultaneously recorded neurons in the motor cortex of awake, behaving monkeys were used to control the movement of a cursor in a three-dimensional virtual-reality environment.

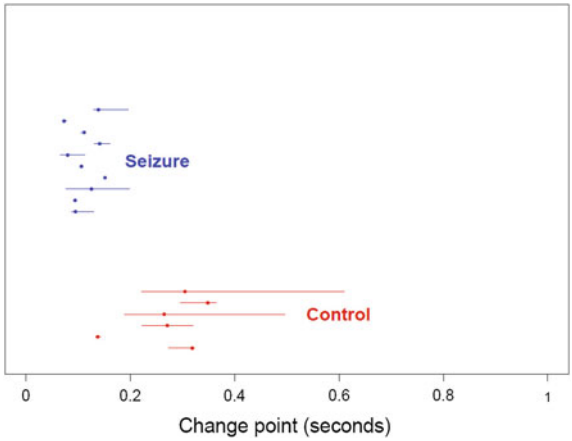

Fig. 8.7 Change points and SEs for neurons of both seizure and control groups. The results for the seizure group appear above those for the control group. The seizure group has change points that occur earlier and they tend to be less variable.

This device creates a precise, well-defined mapping between the firing of the recorded neurons and an expressed behavior (cursor movement). In a series of experiments, they forced the animal to relearn the association between neural firing and cursor movement in a subset of neurons and assessed how the network changes to compensate. Their main finding was that changes in neural activity reflect not only an alteration of behavioral strategy but also the relative contributions of individual neurons to the population error signal. As part of their study the authors compared firing rate modulation among neurons whose BCI signals had been artificially perturbed with that among neurons whose BCI signals remained as determined from their control responses. Because the uncertainties varied substantially across neurons, these comparisons among groups of neurons were carried out using weighted means. □

8.1.4 Decision theory often uses mean squared error to represent risk.

At the end of Section 4.3.4, on p. 102, we mentioned that optimal classification may be considered a problem in decision theory where, in general, the expected loss or *risk* is minimized. In the context of estimation we may consider a decision rule d to be a mapping from each possible vector of observations to a parameter value: we may write $d(X_1, \ldots, X_n) = T$. If we use *squared-error loss* defined by

$$L(d(x_1, \ldots, x_n), \theta) = (d(x_1, \ldots, x_n) - \theta)^2 ,$$

then *MSE* is the risk function

$$MSE(T) = E\left(L(d(X_1, \ldots, X_n), \theta)\right).$$

This terminology, viewing MSE as "risk under squared-error loss," is quite common.

8.2 Estimation in Large Samples

8.2.1 *In large samples, an estimator should be very likely to be close to its estimand.*

In the introduction to this chapter we offered the reminder that the sample mean satisfies

$$\bar{X} \xrightarrow{P} \theta$$

which is the law of large numbers. Suppose T_n is an estimator of θ. If, as $n \to \infty$, we have

$$T_n \xrightarrow{P} \theta \tag{8.14}$$

then T_n is said to be a *consistent* estimator of θ. This means that for every positive ϵ, as $n \to \infty$ we have

$$P(|T_n - \theta| > \epsilon) \to 0.$$

Note that, by (8.3), if $MSE(T_n) \to 0$ then T_n is consistent. Also, if T_n satisfies (8.1) and $\sigma_{T_n} \to 0$ then T_n is consistent.

Details: Multiplying the left-hand side of (8.1) by σ_{T_n} and applying Slutsky's theorem we have $T_n - \theta \xrightarrow{P} 0$, which is equivalent to $T_n \xrightarrow{P} \theta$. □

In words, to say that an estimator is consistent is to say that, for sufficiently large samples, it will be very likely to be close to the quantity it is estimating. This is clearly a desirable property. When T_n satisfies (8.1) and $\sigma_{T_n} \to 0$ we will call T_n *consistent and asymptotically normal*.

8.2.2 *In large samples, the precision with which a parameter may be estimated is bounded by Fisher information.*

Let us consider all estimators of θ that are consistent and asymptotically normal in the sense of Section 8.2.1. For such an estimator $T = T_n$ we may say that its distribution

is approximately normal, and we write

$$T \dot{\sim} N(\theta, \sigma_T^2), \tag{8.15}$$

where the symbol $\dot{\sim}$ means "is approximately distributed as." The expression (8.15) is a convenient way to think of the more explicit Eq. (8.1). From (8.15), σ_T may be considered[3] the standard error of T, and an approximate 95 % CI for θ based on T would be $(T - 2\sigma_T, T + 2\sigma_T)$.

Now, suppose we had two such estimators T^A and T^B that both satisfy (8.15). We would say that T^A is asymptotically more accurate than T^B if $\sigma_{T^A} < \sigma_{T^B}$. An extreme case of this was displayed in Fig. 8.2, where $T^A = \bar{X}$ and $T^B = S^2$, with both histograms being approximately normal in shape and σ_{T^B} being more than four times larger than σ_{T^A}. In general, we would prefer to use an estimator with a small σ_T because it would tend to be closer to θ than an estimator with a larger value of σ_T. In addition, a small σ_T would produce comparatively narrow CIs, indicating improved knowledge about θ. Ideally, we would like to find an estimator T for which σ_T would be as small as possible. Fisher (1922) discovered that this is a soluble problem: there is a minimum value of σ_T and, furthermore, this minimum value is achieved by the method of maximum likelihood.

To understand how this works, we may use some rough heuristics[4] based on the normality in (8.15) to get an expression for σ_T. Let us first note an important fact about normal distributions. Suppose $X \sim N(\mu, \sigma^2)$ with σ known, and consider the loglikelihood function

$$\ell(\mu) = \log f_X(x|\mu).$$

We have

$$f_X(x|\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

so that

$$\ell(\mu) = -\frac{(x-\mu)^2}{2\sigma^2}, \tag{8.16}$$

and when we differentiate twice we get

$$\ell'(\mu) = \frac{x-\mu}{\sigma^2}$$

and

$$\ell''(\mu) = -\frac{1}{\sigma^2}$$

[3] In practice, σ_T may depend on the value of θ, which is unknown, so that a data-based version $\hat{\sigma}_T$ would have to be substituted in forming a confidence interval.

[4] For a rigorous treatment along the lines of the argument here see Kass and Vos (1997, Chapter 2). See also Bickel and Doksum (2001, Chapter 5).

which gives

$$\sigma^2 = \frac{1}{-\ell''(\mu)}. \tag{8.17}$$

That is, the standard deviation of a normal pdf is determined by the second derivative of the loglikelihood function $\ell(\mu)$.

The result (8.17) suggests that when a pdf of an estimator is approximately normal, its standard error may be found in terms of the second derivative of the corresponding loglikelihood function. We now apply this idea to the approximate normal pdf based on (8.15). We write the pdf of the estimator T as $f_T(t|\theta)$ and define its loglikelihood function to be

$$\ell_T(\theta) = \log f_T(t|\theta). \tag{8.18}$$

Using the approximate normality in (8.15) and applying (8.17) we get

$$\sigma_T^2 = \frac{1}{-\ell_T''(\theta)}. \tag{8.19}$$

Equation (8.19) implies that minimizing σ_T is the same as maximizing $-\ell_T''(\theta)$. However, there is an important distinction between (8.19) and (8.17). In (8.17), $\ell''(\mu)$ is a constant whereas, because T is a random variable, $-\ell_T''(\theta)$ is also random (it does not reduce to a constant except when T is exactly normally distributed, so that its loglikelihood becomes exactly quadratic). Thus, regardless of how we were to choose the estimator T, we could not guarantee that $-\ell_T''(\theta)$ would be large because there would be some probability that it might be small. We therefore work with its average value, i.e., its expectation, for which we use the following notation:

$$I^T(\theta) = E\left(-\frac{d^2}{d\theta^2}\log f_T(t|\theta)\right). \tag{8.20}$$

If we replace $-\ell_T''(\theta)$ in (8.19) by its expectation, using (8.20), we get

$$\sigma_T^2 = \frac{1}{I^T(\theta)}. \tag{8.21}$$

The quantity $I^T(\theta)$ is called the *information* about θ contained in the estimator T. Thus, an optimal estimator would be one that makes the information as large as possible.

How large can the information $I^T(\theta)$ be? Fisher's insight was that the information in the estimator can not exceed the analogous quantity derived from the whole sample, which is now known as the *Fisher information*. For a parametric family of distributions having pdf $f(x|\theta)$ the Fisher information is given by

$$I_F(\theta) = E\left(-\frac{d^2}{d\theta^2}\log f(X|\theta)\right).$$

To be clear, for a continuous random variable on (A, B) this expectation is

$$I_F(\theta) = -\int_A^B \left(\frac{d^2}{d\theta^2}\log f(x|\theta)\right) f(x|\theta)dx.$$

For a random sample drawn from this distribution the Fisher information is given by[5]

$$\begin{aligned} I(\theta) &= E\left(-\frac{d^2}{d\theta^2}\log\prod_{i=1}^{n} f(X_i|\theta)\right) \\ &= E\left(-\frac{d^2}{d\theta^2}\sum_{i=1}^{n}\log f(X_i|\theta)\right) \\ &= \sum_{i=1}^{n} E\left(-\frac{d^2}{d\theta^2}\log f(X_i|\theta)\right) \end{aligned}$$

and, because the sample involves identically distributed random variables, all of the expected values in this final expression are the same, and equal to $I_F(\theta)$. Therefore, we have

$$I(\theta) = nI_F(\theta).$$

Result Under certain general conditions, the information in an estimator T satisfies

$$I^T(\theta) \leq I(\theta). \tag{8.22}$$

Therefore, the large-sample variance σ_T^2 of a consistent and asymptotically normal estimator satisfies

$$\sigma_T^2 \geq \frac{1}{I(\theta)}. \tag{8.23}$$

In words, (8.22) says that the information in an estimator can not exceed the information in the whole sample. In Section 8.3.1 we add that the MLE attains this upper bound asymptotically, as $n \to \infty$ and, therefore, has the smallest possible asymptotic variance.

A detail: It is possible for an estimator T to achieve the information bound exactly, in finite samples, i.e.,

$$I^T(\theta) = I(\theta)$$

[5] Because the expectation is used in defining $I(\theta)$, it is often called the *expected information* to distinguish it from the *observed information* which we discuss in Section 8.3.2.

for all n. When this happens the estimator contains all of the information about θ that is available in the data, and it is called a *sufficient statistic*. For instance, if we have a sample from a $N(\mu, \sigma^2)$ distribution with σ known, then the sample mean $\bar{X}$ is sufficient for estimating μ. Sufficiency may be characterized in many ways. If T is a sufficient statistic, then the likelihood function based on T is the same as the likelihood function based on the entire sample. For example, it is not hard to verify that the likelihood function based on a sample $(x_1, \ldots, x_n)$ from a $N(\mu, \sigma^2)$ distribution with σ known is the same as the likelihood function based on $\bar{X}$. This property is sometimes known as *Bayesian sufficiency* (see Bickel and Doksum 2001). In addition, if θ is given a prior distribution as in Section 7.3.9, then T is sufficient when the mutual information between θ and T is equal to the mutual information between θ and the whole sample (see Cover and Thomas 1991). Parametrized families of distributions for which it is possible to find a sufficient statistic with the same dimension as the parameter vector are called *exponential families*. See Section 14.1.6. □

A related result is the following. If we let $\psi(\theta) = E(T)$, where the expectation is based on a random sample from the distribution with pdf $f(x|\theta)$, it may be shown[6] that

$$V(T) \geq \frac{\psi'(\theta)^2}{I(\theta)}.$$

Therefore, if T is an unbiased estimator of θ based on a random sample from the distribution with pdf $f(x|\theta)$ we have $\psi'(\theta) = 1$ and

$$V(T) \geq \frac{1}{I(\theta)}. \tag{8.24}$$

This is usually called the *Cramér-Rao lower bound*. Although Eq. (8.24) is of less practical importance than the asymptotic result (8.23), authors often speak of the bound in (8.23) as a Cramér-Rao lower bound.

Fisher information also arises in theoretical neuroscience, particularly in discussion of neural decoding and optimal properties of tuning curves (see Dayan and Abbott 2001).

8.2.3 *Estimators that minimize large-sample variance are called efficient.*

A consistent and asymptotically normal estimator T satisfies (8.1) and it also satisfies (8.22). In (8.1) we suppressed the dependence of T and σ_T on n. The information $I^T(\theta)$ also depends on n, as does $I(\theta)$. We now consider what happens as $n \to \infty$.

[6] See Bickel and Doksum (2001, Chapter 3).

Suppose we have a consistent and asymptotically normal estimator T which, by definition, satisfies (8.1). If we find a sequence of numbers $c_1, c_2, \ldots, c_n, \ldots$ such that

$$\frac{\sigma_{T_n}}{c_n} \to 1 \tag{8.25}$$

then we have

$$\frac{T_n - \theta}{c_n} \xrightarrow{D} N(0, 1). \tag{8.26}$$

Details: We write

$$\frac{T_n - \theta}{c_n} = \frac{T_n - \theta}{\sigma_{T_n}} \frac{\sigma_{T_n}}{c_n}$$

and apply Slutsky's Theorem (p. 163) using (8.25). □

Equation (8.26) says that c_n can also serve as the large-sample standard error of T. If we have two consistent and asymptotically normal estimators T^A and T^B what matters is the limiting ratio η defined by

$$\frac{\sigma_{T^A}}{\sigma_{T^B}} \to \eta$$

as $n \to \infty$. If $\eta < 1$ then, in large samples, T^A is more accurate than T^B, while if $\eta = 1$ the two estimators are equally accurate. This, together with (8.22), leads us to conclude that the large-sample value of σ_T is minimized if

$$\frac{I^T(\theta)}{I(\theta)} \to 1 \tag{8.27}$$

$n \to \infty$. In this case we also have

$$\sqrt{I(\theta)}(T - \theta) \xrightarrow{D} N(0, 1). \tag{8.28}$$

When an estimator attains (8.27), and therefore (8.28), it is said to be *efficient.*

Details: In general, if $a_1, \ldots, a_n, \ldots$ and $b_1, \ldots, b_n, \ldots$ are positive sequences that satisfy

$$\frac{a_n}{b_n} \to 1$$

then

$$\sqrt{\frac{a_n}{b_n}} \to 1.$$

Applying this to (8.27) we get

$$\sqrt{\frac{I^T(\theta)}{I(\theta)}} \to 1. \tag{8.29}$$

as $n \to \infty$. Let us rewrite $1/\sigma_T$ as

$$\frac{1}{\sigma_T} = \sqrt{I^T(\theta)} = \sqrt{\frac{I^T(\theta)}{I(\theta)}}\sqrt{I(\theta)}. \tag{8.30}$$

Putting (8.30) in (8.1) we get

$$\sqrt{\frac{I^T(\theta)}{I(\theta)}}\sqrt{I(\theta)}(T_n - \theta) \xrightarrow{D} N(0, 1). \tag{8.31}$$

Therefore, by Slutsky's Theorem (p. 163), if (8.27) holds for some estimator T then (8.28) also holds. □

Fisher (1922) described efficient estimators by saying they contain the maximal amount of information supplied by the data about the value of a parameter, and there are rigorous mathematical results that justify Fisher's use of these words. Roughly speaking, the information in the data pertaining to the parameter value may be used well (or poorly) to make an estimator more (or less) accurate; in using as much information about the parameter as is possible, an efficient estimator uses the data most efficiently and reduces to a minimum the uncertainty attached to it. Other definitions of efficiency are sometimes used in statistical theory, but the one based on Fisher information remains most immediately relevant to data analysis, and supports Fisher's observations about maximum likelihood.

8.3 Properties of ML Estimators

8.3.1 In large samples, ML estimation is optimal.

We now state Fisher's main discovery about ML estimation.

Result Under certain general conditions, if T is the MLE then (8.27) and (8.28) hold. That is, ML estimators are consistent, asymptotically normal, and efficient:

$$\sqrt{I(\theta)}(\hat{\theta} - \theta) \xrightarrow{D} N(0, 1). \tag{8.32}$$

In other words, when we consider what happens as $n \to \infty$, among all those "nice" estimators that are consistent and asymptotically normal, ML estimators are the best in the sense of having the smallest possible limiting standard deviation.

Results may also be derived[7] in terms of *MSE*. Under certain conditions, an estimator T_n must satisfy

$$I(\theta) \cdot MSE(T_n) \to c$$

where $c \geq 1$ and for the MLE, where $T = \hat{\theta}$, we have

$$I(\theta) \cdot MSE(\hat{\theta}) \to 1.$$

This is a different way of saying that, for large samples, ML estimation is as accurate as possible.

8.3.2 *The standard error of the MLE is obtained from the second derivative of the loglikelihood function.*

Although we have emphasized the theoretical importance of Eq. (8.28), to be useful for data analysis it must be modified: the quantity $I(\theta)$ depends on the unknown parameter θ, so we must replace $I(\theta)$ with an estimate of it. In other words, when we apply maximum likelihood and want to use (8.32) we must modify it to obtain a confidence interval. One possible such modification is fairly obvious, based on the way we modified initial asymptotic normality results in our discussion of confidence intervals in Section 7.3: we replace θ with the MLE $\hat{\theta}$. Under certain conditions we have

$$\sqrt{I(\hat{\theta})}(\hat{\theta} - \theta) \xrightarrow{D} N(0, 1). \tag{8.33}$$

Details: Because $\hat{\theta} \to \theta$ in probability (i.e., the MLE is consistent), it may be shown that we also have $\sqrt{I(\hat{\theta})/I(\theta)} \to 1$ in probability, so we can again apply Slutsky's Theorem together with (8.28) to get (8.33). □

It turns out that there is a more convenient version of the result. The difficulty with (8.33) is that in some problems it is hard to compute $I(\theta)$ analytically because of the required expectation. Instead, as a general rule, we replace $I(\theta)$ with the *observed information* given by

$$I_{OBS}(\hat{\theta}) = -\ell''(\hat{\theta}). \tag{8.34}$$

[7] See the discussion and references in Kass and Vos (1997).

In other words, instead of the expected information evaluated at $\hat{\theta}$ in (8.33), we use the negative second derivative of the loglikelihood, evaluated at $\hat{\theta}$, without[8] any expectation. Again, under certain conditions, we have

$$\sqrt{I_{OBS}(\hat{\theta})}(\hat{\theta} - \theta) \xrightarrow{D} N(0, 1). \tag{8.35}$$

Details: Note that

$$-\frac{1}{n}\ell''(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\theta^2}\log f(x_i|\theta)$$

and that the expectation of the right-hand side is $I_F(\theta)$. From the LLN we therefore have

$$-\frac{1}{n}\ell''(\theta) \xrightarrow{P} I_F(\theta),$$

and it may also be shown that

$$\sqrt{\frac{I_{OBS}(\hat{\theta})}{I(\hat{\theta})}} \xrightarrow{P} 1,$$

which, again by Slutsky's Theorem, gives (8.35). □

Equation (8.35) provides large-sample standard errors and confidence intervals based on ML estimation, given in the following result.

Result For large samples, under certain general conditions, the MLE $\hat{\theta}$ satisfies (8.35), so that its standard error is given by

$$SE = \frac{1}{\sqrt{-\ell''(\hat{\theta})}} \tag{8.36}$$

and an approximate 95 % CI for θ is given by $(\hat{\theta} - 2SE, \hat{\theta} + 2SE)$.

Additional insight about the observed information can be gained by returning to the derivation of (8.17) and applying it, instead, to the likelihood function based on a sample $x_1, \ldots, x_n$ from a $N(\mu, \sigma^2)$ distribution with σ known, as in Section 7.3.2. There, we found the loglikelihood function to be

[8] For the special class of models known as exponential families, which are used with the generalized linear models discussed in Chapter 14, we have $I(\hat{\theta}) = I_{OBS}(\hat{\theta})$ (see, e.g., Kass and Vos 1997) but this is not true in general.

$$\ell(\theta) = -\sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2\sigma^2}$$

which simplified to Eq. (7.2),

$$\ell(\theta) = -\frac{n}{2\sigma^2}(\theta^2 - 2\bar{x}\theta).$$

Differentiating this twice we get

$$\ell''(\theta) = -\frac{n}{\sigma^2},$$

so that

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{-\ell''(\theta)}}. \tag{8.37}$$

In other words, $1/\sqrt{-\ell''(\theta)}$ gives the standard error of the mean in that case.

Quite generally, for large samples, the likelihood function has an approximately normal form and there is a strong analogy with this paradigm case. Specifically, a quadratic approximation to the loglikelihood function (using a second-order Taylor expansion) produces a normal likelihood (because if $Q(\theta)$ is quadratic then $\exp(Q(\theta))$ is proportional to a normal likelihood function) and in this normal likelihood the value of the standard deviation is $1/\sqrt{-\ell''(\hat{\theta})}$. This heuristic helps explain (8.36).

Details: The quadratic approximation to $\ell(\theta)$ at $\hat{\theta}$ is

$$Q(\theta) = \ell(\hat{\theta}) + \ell'(\hat{\theta})(\theta - \hat{\theta}) + \frac{1}{2}\ell''(\hat{\theta})(\theta - \hat{\theta})^2.$$

Using $\ell'(\hat{\theta}) = 0$ and setting $c = \exp(\ell(\hat{\theta}))$ we have

$$\exp(Q(\theta)) = c \exp\left(-\frac{1}{2}(-\ell(\hat{\theta}))(\hat{\theta} - \theta)^2\right). \tag{8.38}$$

The function on the right-hand side of (8.38) has the form of a likelihood function based on $X \sim N(\theta, \sigma^2)$ where $\hat{\theta}$ plays the role of x and $\sigma = 1/\sqrt{-\ell''(\hat{\theta})}$. □

We now consider two simple illustrations.

Illustration: Exponential distribution Suppose $X_i \sim Exp(\lambda)$ for $i = 1, \ldots, n$, independently. The likelihood function is

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{n} \lambda e^{-\lambda x_i} \\ &= \lambda^n e^{-\lambda \sum x_i} \\ &= \lambda^n e^{-\lambda n \bar{x}} \end{aligned}$$

and the loglikelihood function is

$$\ell(\lambda) = n \log \lambda - n\lambda \bar{x}.$$

Differentiating this and setting equal to zero gives

$$0 = n(\frac{1}{\lambda} - \bar{x})$$

and solving this for λ yields the MLE

$$\hat{\lambda} = \frac{1}{\bar{x}}.$$

Continuing, we compute the observed information:

$$\begin{aligned} -\ell''(\hat{\lambda}) &= \frac{n}{\hat{\lambda}^2} \\ &= n\bar{x}^2 \end{aligned}$$

which gives us the large-sample standard error

$$SE(\hat{\lambda}) = \frac{1}{\bar{x}\sqrt{n}}.$$ □

Illustration: Binomial For a $B(n, p)$ random variable it is straightforward to obtain the observed information

$$-\ell''(\hat{p}) = \frac{n}{\hat{p}(1-\hat{p})}.$$

This gives

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

which is the same as the *SE* found in Section 7.3.5. Therefore, the approximate 95 % CI in (7.22) is an instance of that provided by ML estimation with SE given by (8.36). □

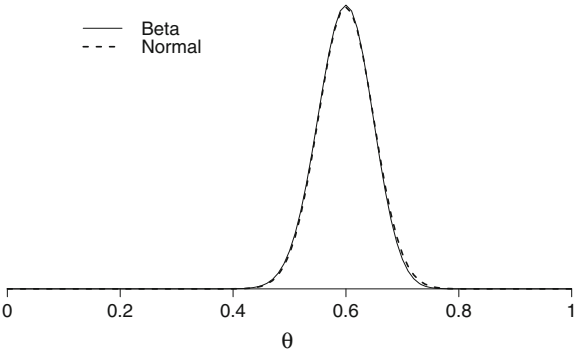

Fig. 8.8 Normal approximation $N(.6, (.049)^2)$ to beta posterior $Beta(61, 41)$.

8.3.3 In large samples, ML estimation is approximately Bayesian.

In Section 7.3.9 we said that Bayes' theorem may be used to provide a form of estimation based on the posterior distribution according to (7.28), i.e.,

$$f_{\theta|x}(\theta|x) = \frac{L(\theta)\pi(\theta)}{\int L(\theta)\pi(\theta)d\theta}.$$

One of the most important results in theoretical statistics is the approximate large-sample equivalence of inference based on ML and inference using Bayes' theorem.

> **Result** For large samples, under certain general conditions, the posterior distribution of θ is approximately normal with mean given by the MLE $\hat{\theta}$ and standard deviation given by the standard error formula (8.36).

We elaborate in Section 16.1.5 and content ourselves here with a simple illustration.

Illustration: Binomial distribution Suppose $Y \sim B(n, \theta)$ with $n = 100$ and we observe $y = 60$. As we said in Section 7.3.9, if we take the prior distribution on θ to be $U(0, 1)$, which is also the $Beta(1, 1)$ distribution, we obtain a $Beta(61, 41)$ posterior. The observed proportion is the MLE $\hat{\theta} = x/n = .6$. The usual standard error then becomes $SE = \sqrt{\hat{\theta}(1-\hat{\theta})/n} = .049$. As shown in Fig. 8.8 the normal distribution with mean $\hat{\theta}$ and standard deviation $\sqrt{\hat{\theta}(1-\hat{\theta})/n}$ is a remarkably good approximation to the posterior. □

For the data from subject P.S. in Example 1.4, which involves a relatively small sample, we already noted (see p. 174) that the approximate 95 % confidence interval (.64, 1.0) found using (8.36) (which is the same as (7.22), see p. 206) differed by

only a modest amount from the exact 95 % posterior probability interval we obtained, which was (.59, .94).

8.3.4 MLEs transform along with parameters.

It sometimes happens that we wish to consider an alternative parameterization of a pdf, say γ rather than θ, and then want find the MLE of γ. If $\gamma = g(\theta)$ for a transformation function g having nonzero derivative, then the MLE of the transformation equals the transformation of the MLE:

$$\hat{\gamma} = g(\hat{\theta}).$$

This is often called *invariance* or *equivariance.* The derivation of invariance of ML is perhaps most easily followed in a concrete example. The argument given next for the exponential distribution could be applied to any parametric family.

Illustration: Exponential distribution (continued from p. 205) Suppose we parameterize the $Exp(\lambda)$ distribution in terms of the mean $\mu = 1/\lambda$ so that its pdf becomes

$$f(x) = \frac{1}{\mu} e^{-x/\mu}.$$

Previously (see p. 205) we found that the MLE of λ based on a sample from $Exp(\lambda)$ is $\hat{\lambda} = 1/\bar{x}$. The invariance property of ML says that

$$\hat{\mu} = 1/\hat{\lambda} = \bar{x}.$$

To see why this works for the exponential distribution, let us use a subscript on the likelihood function to indicate its argument, $L_\lambda(\lambda)$ vs. $L_\mu(\mu)$. We find $L_\mu(\mu)$ by starting with

$$L_\lambda(\lambda) = \lambda^n e^{-\lambda n \bar{x}}$$

and writing

$$L_\mu(\mu) = L_\lambda(\frac{1}{\mu}) = \frac{1}{\mu^n} e^{-n\bar{x}/\mu}.$$

Thus, when we maximize $L_\mu(\mu)$ over μ, we are maximizing $L_\lambda(1/\mu)$ over μ which is the same thing as maximizing $L_\lambda(\lambda)$ over λ. We therefore must have $\hat{\mu} = 1/\hat{\lambda}$. More generally, the same argument shows that when $\gamma = g(\theta)$ we must have $\hat{\gamma} = g(\hat{\theta})$. □

Invariance is by no means a trivial property: some methods of estimation are *not* invariant to transformations of the parameter.

8.3.5 Under normality, ML produces the weighted mean.

We now return to choosing the weights for a weighted mean, discussed in Section 8.1.3. Previously (p. 190) we found the weights that minimized *MSE*. A different way to solve the problem is to introduce a statistical model, and then apply the method of maximum likelihood. Let us do this.

To apply ML, we assume that X_1 and X_2 are both normally distributed. The loglikelihood is

$$\ell(\mu) = -\frac{(x_1-\mu)^2}{2\sigma_1^2} - \frac{(x_2-\mu)^2}{2\sigma_2^2}$$

and setting its derivative equal to zero gives

$$\begin{aligned} 0 &= -\frac{x_1-\mu}{\sigma_1^2} - \frac{x_2-\mu}{\sigma_2^2} \\ &= -\frac{x_1}{\sigma_1^2} - \frac{x_2}{\sigma_2^2} + \mu\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right). \end{aligned}$$

Therefore, dividing through by $\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}$, the MLE is

$$\hat{\mu} = w_1 \cdot X_1 + w_2 \cdot X_2,$$

where

$$w_i = \frac{\frac{1}{\sigma_i^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}$$

for $i = 1, 2$. This is Eq. (8.10).

8.4 Multiparameter Maximum Likelihood

The method of ML estimation was defined for the case of a scalar parameter θ in Section 7.2.2, together with Eqs. (8.35) and (8.36). More generally, when θ is a vector, the definitions of the likelihood function, loglikelihood function, and MLE remain unchanged. The observed information instead becomes a matrix, and the approximate normal distribution mentioned in conjunction with Eq. (8.36) instead becomes an approximate *multivariate* normal distribution.

8.4.1 The MLE solves a set of partial differential equations.

In Section 7.2.2 we computed the MLE by solving the differential equation

$$0 = \ell'(\theta) \tag{8.39}$$

when θ was a scalar. To obtain the MLE of an m-dimensional vector parameter, we must solve precisely the same equation, except that now the derivative in Eq. (8.39) is the vector

$$\ell'(\theta) = \begin{pmatrix} \frac{\partial \ell}{\partial \theta_1} \\ \frac{\partial \ell}{\partial \theta_2} \\ \vdots \\ \frac{\partial \ell}{\partial \theta_m} \end{pmatrix}.$$

This means that Eq. (8.39) is really a set of m equations, often called *the likelihood equations*, which need to be solved simultaneously.

Illustration: Normal MLE Let us return to finding the MLE for a sample $x_1, \ldots, x_n$ from a $N(\mu, \sigma^2)$ distribution. Previously we assumed σ was known, but now we consider the joint estimation of μ and σ. The loglikelihood function now must include a term previously omitted that involves σ. The joint pdf is

$$f(x_1, \ldots, x_n | \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(- \frac{(x_i - \mu)^2}{2\sigma^2} \Big)$$

and the loglikelihood function is

$$\ell(\mu, \sigma) = -n \log \sigma - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}.$$

The partial derivatives are

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \sigma^{-3} \sum_{i=1}^{n} (x_i - \mu)^2$$

Setting the first equation equal to 0 we obtain

$$\hat{\mu} = \bar{x}.$$

Setting the second equation equal to 0 and substituting $\hat{\mu} = \bar{x}$ gives

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

The MLE is thus slightly different than the usual sample standard deviation s, which is defined with the denominator $n-1$ so that the sample variance becomes unbiased as an estimator of σ^2, as in (8.4). We have

$$\hat{\sigma} = \sqrt{\frac{n-1}{n}} \cdot s.$$

Clearly the distinction is unimportant for substantial sample sizes.[9] □

Illustration: Gamma MLE Let us rewrite the gamma loglikelihood function:

$$\ell(\alpha, \beta) = n\alpha \log \beta + (\alpha - 1)\sum_{i=1}^{n} \log x_i - \beta \sum_{i=1}^{n} x_i - n \log \Gamma(\alpha).$$

The partial derivatives are

$$\frac{\partial \ell}{\partial \alpha} = n \log \beta + \sum_{i=1}^{n} \log x_i - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$$
$$\frac{\partial \ell}{\partial \beta} = \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i$$

where $\Gamma'(u)$ is the derivative of the function $\Gamma(u)$ (sometimes called the "digamma function"). Setting the second partial derivative equal to zero we obtain

$$\hat{\beta} = \frac{n\hat{\alpha}}{\sum_{i=1}^{n} x_i}.$$

When we set the first equation equal to zero and substitute this expression for $\hat{\beta}$, we get the nonlinear equation

$$n \log \hat{\alpha} - n \log \bar{x} + \sum_{i=1}^{n} \log x_i - n\frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = 0.$$

To obtain the MLE $(\hat{\alpha}, \hat{\beta})$ we may proceed iteratively: given a value $\hat{\beta}^{(j)}$ we can solve the first equation for $\hat{\alpha}^{(j+1)}$ and solve the second equation to obtain $\hat{\beta}^{(j+1)}$; we

[9] We may obtain $\hat{\sigma} = s$ if we instead integrate out μ from the likelihood and then maximize the resulting function; this function is sometimes called an *integrated* or *marginal* likelihood, and in some situations maximizing the integrated likelihood yields a preferable estimator.

continue until the results converge. The second equation must be solved numerically, but it is not very difficult to use available software to do so. □

8.4.2 *Least squares may be viewed as a special case of ML estimation.*

In Example 1.5 we discussed data collected by Hursh (1939), indicating the linear relationship between a neuron's conduction velocity and its axonal diameter. We also briefly described the method of *least-squares regression*, based on the *linear regression model* (1.4), which is

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{8.40}$$

where $\epsilon_i \sim N(0, \sigma^2)$, independently. Least-squares regression is discussed at length in Chapter 12. Here we show that the method of least squares may be considered a special case of ML estimation.

Least squares may be derived by assuming that the ϵ error variables in (8.40) are normally distributed, and that the problem is to estimate the parameter vector $\theta = (\beta_0, \beta_1)$. Specifically, we assume $\epsilon_i \sim N(0, \sigma^2)$, independently for all i. Calculation then shows that the ML estimate of θ is the least squares estimate. In other words, in the simple linear regression problem, ML based on the assumption of normal errors reproduces the least-squares solution.

Details: In the illustration on p. 210 we wrote down the loglikelihood function for a sample from a $N(\mu, \sigma^2)$ distribution,

$$\ell(\mu, \sigma) = -n \log \sigma - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

and obtained the MLE $\hat{\mu} = \bar{x}$. Notice that, as a function of μ, the loglikelihood is maximized by minimizing the sum of squares $\sum_{i=1}^{n}(x_i - \mu)^2$. Thus, the MLE $\hat{\mu} = \bar{x}$ is also a least-squares estimator in the one-sample problem. For the simple linear regression model (8.40) the loglikelihood function becomes

$$\ell(\beta_0, \beta_1, \sigma) = -n \log \sigma - \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}.$$

We can maximize $\ell(\beta_0, \beta_1, \sigma)$ by first defining $(\hat{\beta}_0(\sigma), \hat{\beta}_1(\sigma))$ to be the maximum of $\ell(\beta_0, \beta_1, \sigma)$ over (β_0, β_1) for fixed σ, and then maximizing $\ell(\hat{\beta}_0(\sigma), \hat{\beta}_1(\sigma), \sigma)$ over σ. However, from inspection of the formula above, for every σ the solution $(\hat{\beta}_0(\sigma), \hat{\beta}_1(\sigma))$ (the maximum of $\ell(\beta_0, \beta_1, \sigma)$) is found by minimizing the sum of squares

$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$. Therefore, the MLE $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma})$ has the least-squares estimate as its first two components. □

8.4.3 *The observed information is the negative of the matrix of second partial derivatives of the loglikelihood function, evaluated at $\hat{\theta}$.*

In the multiparameter case the second derivative $\ell''(\theta)$ becomes a matrix,

$$\ell''(\theta) = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \theta_1^2} & \frac{\partial^2 \ell}{\partial \theta_1 \theta_2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_1 \theta_m} \\ \frac{\partial^2 \ell}{\partial \theta_1 \theta_2} & \frac{\partial^2 \ell}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_2 \theta_m} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 \ell}{\partial \theta_1 \theta_m} & \frac{\partial^2 \ell}{\partial \theta_2 \theta_m} & \cdots & \frac{\partial^2 \ell}{\partial \theta_m^2} \end{pmatrix}.$$

This second-derivative matrix is often called the *Hessian* of $\ell(\theta)$. The *observed information matrix* is $-\ell''(\hat{\theta})$, which generalizes (8.34).

Result For large samples, under certain general conditions, the MLE $\hat{\theta}$ of the m-dimensional parameter θ is distributed approximately as an m-dimensional multivariate normal random vector with variance matrix

$$\hat{\Sigma} = -\ell''(\hat{\theta})^{-1}, \tag{8.41}$$

i.e.,

$$\hat{\Sigma}^{-1/2}(\hat{\theta} - \theta) \xrightarrow{D} N_m(0, I_m) \tag{8.42}$$

as $n \to \infty$.

Example 5.5 (continued from p. 112) In the Hecht et al. experiments on threshold for visual perception of light, the response variable was an indication of whether or not light was observed by a particular subject ("yes" or "no"), and the explanatory variable was the intensity of the light (in units of average number of light quanta per flash). Several different intensities were used, and for each the experiment was repeated many times. The results for one series of trials in one subject are plotted in Fig. 8.9.

As illustrated in Fig. 8.9, the linear regression model (8.40) does not work very well in this example. The proportions vary between 0 and 1 but a line $y = a + bx$ is unrestricted and can not represent the variation accurately, at least not for proportions that get close to 0 or 1. A solution is to replace the line $y = a + bx$ by a sigmoidal curve, which goes to zero as the explanatory variable x goes to $-\infty$ and increases

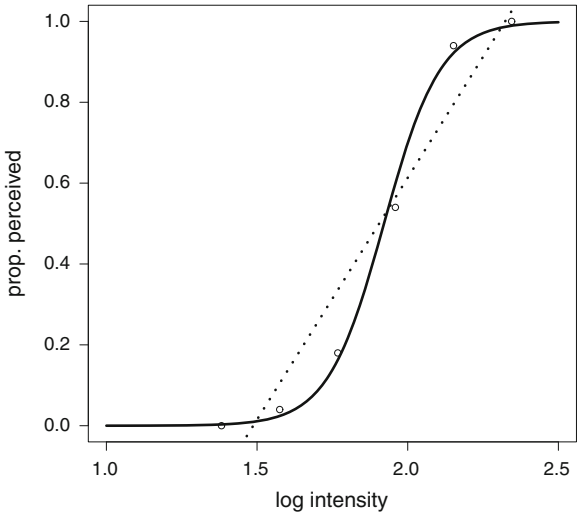

Fig. 8.9 Proportion of trials, out of 50, on which light flashes were perceived by subject S.S. as a function of $\log_{10}$ intensity, together with fits. Data from Hecht et al. (first series of trials) are shown as *circles*. *Dashed line* is the fit obtained by linear regression. *Solid curve* is the fit obtained by logistic regression.

to one as $x \to \infty$. The fitted curve in Fig. 8.9 is based on the following statistical model: for the ith value of light intensity we let Y_i be the number of light flashes on which the subject perceives light and then take

$$Y_i \sim B(n_i, p_i) \tag{8.43}$$

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}. \tag{8.44}$$

This is known as the *logistic regression model*. There are many possible approaches to estimating the parameter vector $\theta = (\beta_0, \beta_1)$ but the usual solution is to apply maximum likelihood. The observed information matrix is then used to get standard errors of the coefficients. These calculations are performed by most statistical software packages. For the data in Fig. 8.9 we obtained $\hat{\beta}_0 = -20.5 \pm 2.4$ and $\hat{\beta}_1 = 10.7 \pm 1.2$. Further discussion of logistic regression, and interpretation of this result, are given in Section 14.1. □

8.4.4 When using numerical methods to implement ML estimation, some care is needed.

There are three issues surrounding the application of numerical maximization to ML estimation. The first is that, while loglikelihood functions are usually well behaved

near their maxima, they may be poorly behaved away from the maxima. In particular, a loglikelihood may have multiple smaller peaks, and numerical methods may get stuck in a region away from the actual maximum. Except in cases where the loglikelihood is known to be concave (see Section 14.1.6.), it is essential to begin an iterative algorithm with a good preliminary estimate. Sometimes models may be altered and simplified in some way to get guesses at the parameter values. In some cases the method of moments may be used to get initial values for an iterative maximization algorithm.

Illustration: Gamma distribution On p. 153 we found the method of moments estimator for the Gamma distribution,

$$\beta^* = \frac{\bar{x}}{s^2}$$
$$\alpha^* = \frac{\bar{x}^2}{s^2}.$$

In order to obtain the MLE of (α, β) we may use an iterative maximization algorithm beginning with $(\hat{\alpha}^{(1)}, \hat{\beta}^{(1)}) = (\alpha^*, \beta^*)$. □

With good initial values, iterative maximization software usually only needs to run for a few iterations, after which the estimates don't change by more than a small fraction of the statistical uncertainty (represented by standard errors). In fact, it may be shown, theoretically, that from any consistent estimator for which the *MSE* vanishes at the rate $1/n$, a single iteration of Newton's method for maximizing the loglikelihood function will produce an efficient estimator (see Lehmann, 1983).

A second important implementation issue is that the second derivatives used in numerical maximization software are often themselves estimated numerically, and they may be estimated rather poorly (because they do not need to be estimated accurately to obtain the maximum). Thus, for the purpose of finding a variance matrix, one should either evaluate second derivatives separately (from an analytical formula, or from special-purpose software), or one should apply the parametric bootstrap (see Section 9.2).

The third issue is that parameterization can be important. Numerical maximization procedures tend to work well when the loglikelihood function is roughly quadratic, which means that the likelihood function is approximately normal. Transformations of parameters can improve this approximation. For example, before running maximization software it is often helpful to transform variance parameters by taking logs.

8.4.5 MLEs are sometimes obtained with the EM algorithm.

Certain statistical models have a structure that lends itself to a special method of likelihood maximization known as the *expectation-maximization (EM) algorithm*. We describe it in one special case.

Illustration: Mixture of Two Gaussians Suppose a random variable X follows either a $N(\mu_1, \sigma_1^2)$ distribution or a $N(\mu_2, \sigma_2^2)$ distribution, and that the selection of the distribution is determined probabilistically: with probability π we have $X \sim N(\mu_1, \sigma_1^2)$ and with probability $1 - \pi$ we have $X \sim N(\mu_2, \sigma_2^2)$. The pdf of X is

$$f_X(x) = \pi f(x; \mu_1, \sigma_1^2) + (1 - \pi) f(x; \mu_2, \sigma_2^2) \tag{8.45}$$

where $f(x; \mu, \sigma^2)$ is the $N(\mu, \sigma^2)$ pdf. If we consider a large sample of values $x_1, \ldots, x_n$ from the distribution of X, some proportion of x_i values (approximately $n\pi$ of them) would be from the $N(\mu_1, \sigma_1^2)$ distribution, while the rest would be from the $N(\mu_2, \sigma_2^2)$ distribution. Such a sample would thus blend the $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ distributions and (8.45) defines a *mixture* of two normal distributions, often called a *mixture of Gaussians* model. Based on a sample of data the problem is to estimate the parameter vector $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \pi)$.

Let us introduce a random variable W_i to represent the selected distribution for X_i in the sense that $W_i = 1$ with probability π and if $W_i = 1$ a value U is drawn from $N(\mu_1, \sigma_1^2)$ and we set $X_i = U$, while $W_i = 0$ with probability $1 - \pi$ and if $W_i = 0$ a value V is drawn from $N(\mu_2, \sigma_2^2)$ and we set $X_i = V$. The variables $W_1, \ldots, W_n$ are not observed. If they were known, however, the problem would be much simpler: we could collect the values of W_i for which $W_i = 1$ and take the sample mean and variance of those as estimates[10] of μ_1 and σ_1^2 and then collect the values of W_i for which $W_i = 0$ and take the sample mean and variance of those as estimates of μ_2 and σ_2^2. Because the W_is are unobserved, they are often called *latent variables* (see Section 16.2). The data $(x_1, \ldots, x_n)$ are said to be *augmented* by $(w_1, \ldots, w_n)$. Let us write $Y = (X_1, \ldots, X_n)$ and $Z = (W_1, \ldots, W_n)$ and then write the loglikelihood function based on the original data y as $\ell_y(\theta)$ and that based on the augmented data (y, z) as $\ell_{(y,z)}(\theta)$. We have

$$\begin{aligned}
\ell_{(y,z)}(\theta) &= \sum_{i=1}^{n} w_i \log f(x_i; \mu_1, \sigma_1^2) + \sum_{i=1}^{n} (1 - w_i) \log f(x_i; \mu_2, \sigma_2^2) \\
&\quad + \sum_{i=1}^{n} w_i \log \pi + \sum_{i=1}^{n} (1 - w_i) \log(1 - \pi) \\
&= \sum_{\{i: w_i = 1\}} \log f(x_i; \mu_1, \sigma_1^2) + \sum_{\{i: w_i = 0\}} \log f(x_i; \mu_2, \sigma_2^2) \\
&\quad + \sum_{\{i: w_i = 1\}} \log \pi + \sum_{\{i: w_i = 0\}} \log(1 - \pi)
\end{aligned} \tag{8.46}$$

and maximizing this with respect to (μ_1, σ_1^2) is the same as maximizing the likelihood for a sample a $N(\mu_1, \sigma_1^2)$ pdf made up of the values x_i for which $w_i = 1$ (and similarly

[10] As we said in Section 8.4.1 (see p. 210), the MLE of the variance has denominator n rather than $n - 1$ but the sample variance is usually preferred.

for (μ_2, σ_2^2)). Thus, the introduction of the latent variables W_i has greatly simplified the problem. However, because these latent variables have not been observed we must get estimates that do not rely on them. To do this we may integrate out the W_i variables (marginalize over them), as we next explain.

If we think of π as a prior probability that $W_i = 1$ then, after observing $X_i = x_i$ we may compute the posterior probability from Bayes' Theorem as

$$P(W_i = 1|X_i = x_i) = \frac{\pi f(x_i; \mu_1, \sigma_1^2)}{\pi f(x_i; \mu_1, \sigma_1^2) + (1 - \pi) f(x_i; \mu_2, \sigma_2^2)}. \tag{8.47}$$

We use the notation

$$\gamma_i = P(W_i = 1|X_i = x_i). \tag{8.48}$$

Note that, because W_i is a binary variable γ_i may also be written

$$\gamma_i = E(W_i|X_i = x_i)$$

and, for later purposes, we make the dependence on θ explicit by writing

$$\gamma_i = E(W_i|X_i = x_i, \theta). \tag{8.49}$$

With this framework in hand, the EM algorithm for this problem may be defined. It produces an iterative sequence $\theta^{(1)}, \theta^{(2)}, \ldots$ that, with good initial values, will converge to the MLE $\hat{\theta}$. Here is the algorithm.

1. Find an initial value $\theta^{(1)}$ for θ and set $j = 1$.
2. Given a current value $\theta^{(j)}$ compute $\gamma_i^{(j)}$ for $i = 1, \ldots, n$ by applying (8.48) using (8.47) where $\theta = \theta^{(j)}$.
3. Using $\gamma_1^{(j)}, \ldots, \gamma_n^{(j)}$ from Step 2 compute the components of $\theta^{(j+1)}$ as follows:

$$\begin{aligned}
\mu_1^{(j+1)} &= \frac{\sum_{i=1}^n \gamma_i^{(j)} x_i}{\sum_{i=1}^n \gamma_i^{(j)}} \\
\mu_2^{(j+1)} &= \frac{\sum_{i=1}^n (1 - \gamma_i^{(j)}) x_i}{\sum_{i=1}^n (1 - \gamma_i^{(j)})} \\
\sigma_1^{2(j+1)} &= \frac{\sum_{i=1}^n \gamma_i^{(j)} (x_i - \mu_1^{(j+1)})^2}{\sum_{i=1}^n \gamma_i^{(j)}} \\
\sigma_2^{2(j+1)} &= \frac{\sum_{i=1}^n (1 - \gamma_i^{(j)}) (x_i - \mu_1^{(j+1)})^2}{\sum_{i=1}^n (1 - \gamma_i^{(j)})} \\
\pi^{(j+1)} &= \frac{1}{n} \sum_{i=1}^n \gamma_i^{(j)}.
\end{aligned}$$

4. Increment j and return to Step 2.
5. Repeat Steps 2–4 until convergence. □

A key step in formulating the EM algorithm in the mixture of two Gaussians model, above, was the introduction of the random variables W_i. In order to maximize the loglikelihood $\ell_Y(\theta)$ defined by the pdf $f_Y(y|\theta)$ we effectively introduced the loglikelihood $\ell_{(Y,Z)}(\theta)$ in (8.46) based on the augmented data pdf $f_{(Y,Z)}(y, z|\theta)$. Step 2 of the algorithm, known as *the expectation step*, is based on the expectation $E(\ell_{(Y,Z)}(\theta)|Y = y, \theta = \theta^{(j)})$. In Step 2 the conditional expectation in (8.49) was evaluated for $\theta = \theta^{(j)}$. In Step 3 the loglikelihood was maximized in terms of the expectations computed in Step 2.

In general, if $Y = y$ is the data vector augmented by $Z = z$ we define

$$Q(\theta, \theta^{(j)}) = E(\ell_{(Y,Z)}(\theta)|Y = y, \theta = \theta^{(j)}). \tag{8.50}$$

Beginning with an initial guess $\theta^{(1)}$, for each j the EM algorithm computes $Q(\theta, \theta^{(j)})$ and sets $\theta^{(j+1)}$ equal to the maximizer of $Q(\theta, \theta^{(j)})$ as a function of θ. The EM algorithm works well for problems in which some kind of data augmentation greatly simplifies the problem, so that $Q(\theta, \theta^{(j)})$ is easy to compute (as in Step 2 of the mixture of two Gaussians illustration above). In addition to models that incorporate latent variables, the EM algorithm is often applied to problems with missing data, where the missing data are treated as augmenting the observed data. (See also the related discussion of Gibbs sampling in Section 16.2.2.)

One way to see that this iterative scheme should work is to apply the formula[11]

$$\frac{d}{d\theta} Q(\theta, \theta^*)|_{\theta=\theta*} = \ell'_Y(\theta^*) \tag{8.51}$$

(see the details below). If $\theta^{(1)}, \theta^{(2)}, \ldots$ is a sequence of EM iterates that converge to a value θ^* then, because each iterate maximizes $Q(\theta, \theta^{(j)})$ its derivative is 0, i.e.,

$$\frac{d}{d\theta} Q(\theta, \theta^*)|_{\theta=\theta*} = 0.$$

From (8.51) we then have

$$\ell'_Y(\theta^*) = 0.$$

Thus, for sufficiently good initial values, when the EM algorithm converges to θ^* we get $\theta^* = \hat{\theta}$, i.e., the EM algorithm converges to the MLE $\hat{\theta}$.

Details: We derive Eq. (8.51). From (8.50) we have

$$Q(\theta, \theta^*) = \int \frac{f(y, z|\theta^*)}{f(y|\theta^*)} \log f(y, z|\theta) dz.$$

[11] This formula was used by Fisher, in his discussion of sufficiency, to substantiate the argument mentioned in Section 8.2.2 (see p. 200 and Kass and Vos 1997, Section 2.5.1)

We differentiate under the integral:

$$\begin{aligned}\frac{d}{d\theta}Q(\theta,\theta^*)|_{\theta=\theta*} &= \int \frac{f(y,z|\theta^*)}{f(y|\theta^*)}\frac{\frac{d}{d\theta}f(y,z|\theta)|_{\theta=\theta*}}{f(y,z|\theta^*)}dz \\ &= \int \frac{\frac{d}{d\theta}f(y,z|\theta)|_{\theta=\theta*}}{f(y|\theta^*)}dz.\end{aligned}$$

We continue using $f(y,z|\theta) = f(z|y,\theta)f(y|\theta)$, differentiate the product, and rewrite:

$$\begin{aligned}&\frac{d}{d\theta}Q(\theta,\theta^*)|_{\theta=\theta*} \\ &= \int \frac{\frac{d}{d\theta}f(z|y,\theta)f(y|\theta)|_{\theta=\theta*}}{f(y|\theta^*)}dz \\ &= \int \frac{f(y|\theta^*)\frac{d}{d\theta}f(z|y,\theta)|_{\theta=\theta*} + f(z|y,\theta^*)\frac{d}{d\theta}f(y|\theta)|_{\theta=\theta*}}{f(y|\theta^*)}dz \\ &= \int \frac{d}{d\theta}f(z|y,\theta)|_{\theta=\theta*} + f(z|y,\theta^*)\frac{d}{d\theta}\log f(y|\theta)|_{\theta=\theta*}dz.\end{aligned} \tag{8.52}$$

In this last expression the integral of the first term vanishes because

$$\int f(z|y,\theta)dz = 1$$

so that

$$\frac{d}{d\theta}\int f(z|y,\theta)dz = 0$$

and taking the derivative under the integral gives

$$\int \frac{d}{d\theta}f(z|y,\theta)|_{\theta=\theta*}dz = 0.$$

Therefore, expression (8.52) reduces to (8.51). □

8.4.6 Maximum likelihood may produce bad estimates.

The method of ML is not universally applicable, nor does it guarantee good statistical results. The most serious concern with ML is that it is predicated on the description of the data according to a particular statistical model. If that model is seriously deficient, the MLE will be misleading. This underscores the essential role of model assessment, and the iterative nature of model building, emphasized in Chapter 1.

The provably good performance of ML estimation also applies only for large samples. What constitutes "large" is difficult to specify precisely, though attempts have been made occasionally. A key observation is that sample size must be judged relative to the number of parameters being estimated. In problems having large numbers of parameters and only modest sample sizes, we should expect neither ML estimates, nor their associated SEs, to be accurate. One standard approach to making progress in such situations is to build models that effectively reduce the number of parameters by restricting them in some way (often by introducing additional probability distributions). In some cases, however, ML must be abandoned. There is a large body of methods that are *nonparametric*, in the sense that they do not posit a statistical model with a finite number of parameters. There are many situations where nonparametric methods perform well, and save the difficulty and worry associated with careful model building.

Chapter 9
Propagation of Uncertainty and the Bootstrap

At the beginning of this book we said that we wanted to lay out the key features of what we called, "the statistical paradigm," which consists of broadly applicable concepts that guide reasoning from data in diverse contexts. One of its foundations is the idea that data may be used to express knowledge and uncertainty about unknown values of model parameters, especially through confidence intervals. This was the focus of Chapter 7. Another is the notion that alternative estimators may be evaluated and compared, which was the main subject of Chapter 8, together with the large-sample optimality and utility of ML estimation. We now turn to a third building block of statistical reasoning, which is a major source of the remarkable reach and flexibility of modern data analysis, especially in complicated settings. It is based on the simple idea that when we have an expression of uncertainty about a random variable or random vector X, in the form of a standard error or variance matrix, we can *propagate* this uncertainty to a new variable Y, where $y = f(x)$ for some[1] function $f(x)$, in order to get a standard error for Y. Let us be concrete by considering a simple example.

Example 5.5 (continued from p. 112) We previously displayed data from Hecht et al. (1942), who investigated the threshold for visual perception by exposing human observers to very weak flashes of light in a darkened room. In the bottom part of Fig. 8.9 we overlaid on the data a sigmoidal curve found from the logistic regression model given by the pair of Eqs. (8.43) and (8.44), using maximum likelihood estimation. We reported the values of the fitted coefficients and their standard errors.

Those data were from a single subject. What if we wanted to compare results across subjects? We would get a set of sigmoidal curves with somewhat different slopes, shifted to some extent to the left or right. One common way such curves are characterized is by the intensity x_{50} at which the subject will perceive the light 50 % of the time. To find x_{50} we begin with Eq. (8.44), which without subscripts on x_i and p_i becomes

[1] We are *not* intending $f(x)$ to be a pdf. We are here, in this chapter, using the notation $y = f(x)$ to refer to some general function.

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_9,

$$p = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}. \tag{9.1}$$

In (9.1) we set $p = .5$ and solve for x_{50}. That is, we solve the equation

$$.5 = \frac{\exp(\beta_0 + \beta_1 x_{50})}{1 + \exp(\beta_0 + \beta_1 x_{50})}$$

for x_{50} as a function of (β_0, β_1). Details given on p. 226 show that

$$x_{50} = \frac{-\beta_0}{\beta_1}. \tag{9.2}$$

In terms of the form $y = f(x)$, here the role of y is played by x_{50}, the role of x is played by (β_0, β_1), and the function is $f(\beta_0, \beta_1) = -\beta_0/\beta_1$.

To estimate x_{50} we replace β_0 and β_1 in (9.2) by their fitted values $\hat{\beta}_0$ and $\hat{\beta}_1$. As discussed in Section 8.4.3, an approximate variance matrix of $(\hat{\beta}_0, \hat{\beta}_1)$ is given by the inverse of observed information matrix for $(\hat{\beta}_0, \hat{\beta}_1)$. It is available from the fitting software. We want to use that variance matrix to express knowledge about x_{50} in the form of a standard error $SE(\hat{x}_{50})$. This is a problem in propagation of uncertainty, which we may write schematically in two ways:

$$\text{uncertainty about } (\beta_0, \beta_1) \overset{propagate}{\longrightarrow} \text{uncertainty about } x_{50} \tag{9.3}$$

which states the uncertainty about x_{50} will have to be obtained from uncertainty about the coefficients (β_0, β_1), and

$$\text{uncertainty attached to } (\hat{\beta}_0, \hat{\beta}_1) \overset{propagate}{\longrightarrow} \text{uncertainty attached to } \hat{x}_{50} \tag{9.4}$$

which, more prescriptively, suggests that we will use the uncertainty we have evaluated along with the estimated coefficients $(\hat{\beta}_0, \hat{\beta}_1)$ (i.e., the inverse of the observed information matrix) to get an expression of uncertainty for $\hat{x}_{50}$. □

Propagation of uncertainty is an old concept[2] but it was given a new, and profoundly important twist with the development of bootstrap methods by Bradley Efron (Efron 1979a). Bootstrap methods for confidence intervals rest on two ideas. First, that the variability in the data, based on the statistical model, may be estimated reasonably accurately and, second, that this variability may be propagated to express uncertainty about any quantities computed from the data, such as functions of the unknown parameters in the model. In the context of estimating x_{50} in Example 5.5 we might write this, schematically, in two steps:

[2] The "law of propagation of error," as it was called, is mentioned as a standard technique by Schultz (1929).

$$\begin{aligned} \text{variation in } (y_1, y_2, \ldots, y_6) &\overset{propagate}{\longrightarrow} \text{uncertainty attached to } (\hat{\beta}_0, \hat{\beta}_1) \\ \text{uncertainty attached to } (\hat{\beta}_0, \hat{\beta}_1) &\overset{propagate}{\longrightarrow} \text{uncertainty attached to } \hat{x}_{50}. \end{aligned} \tag{9.5}$$

Efron's insight was that propagation of uncertainty, from variability in the data to uncertainty in estimates, could be carried out easily on the computer in a wide variety of circumstances, and he followed up with convincing theoretical analysis of the method using some of the principles articulated in Chapter 8. In the 1980s, when desktop computers became available, the use of computers to propagate uncertainty took off (see Efron 1979b).

We discuss propagation of uncertainty in Section 9.1 and then move on to bootstrap methods in Section 9.2. In Section 9.3 we specify the circumstances under which each of the several methods described here might be preferred to the others.

9.1 Propagation of Uncertainty

The problem of transferring uncertainty about a random vector X to a random variable $Y = f(X)$ is the problem of propagation of uncertainty, or what was historically called "propagation of error" and, sometimes, "the delta method." There are several varieties of propagation of uncertainty. The original method, historically, used mathematical analysis with $n \to \infty$ to derive an approximate standard error for Y, which we write as $SE(Y)$, based on an approximate variance matrix for X. In some cases this is easy. We discuss it in Section 9.1.2. It is often even easier to use a brute force computer simulation: if we can generate observations (on the computer) from the approximate distribution of X, we can also immediately obtain the approximate distribution of Y. We explain this method, enumerating the steps, in Section 9.1.1. Propagation of uncertainty is also an essential part of modern Bayesian methods, which appear in Chapter 16.

9.1.1 Simulated observations from the distribution of the random variable X produce simulated observations from the distribution of the random variable $Y = f(X)$.

It is sometimes advantageous to work out analytically the approximate standard error according to (9.19), derived below. However, the calculations can be complicated, which may make them tedious and could also result in math mistakes. A remarkably effective way to propagate uncertainty, which may also reduce the chance of overlooking a math error, is to use simulation. To understand the method, one must first be sure to understand how to work with a probability distribution based on a transformation $y = f(x)$. Let us consider a simple illustration.

Illustration: A random variable having three possible values Suppose X can take the values 2, 4, or 8 with probabilities .2, .5, .3, respectively, and we are interested in the transformation $y = \log_2(x)$. Then Y can take the values 1, 2, or 3. To find the probability distribution of Y we simply note that

$$\begin{aligned} P(Y = 1) &= P(\log_2(X) = 1) = P(X = 2) = .2 \\ P(Y = 2) &= P(\log_2(X) = 2) = P(X = 4) = .5 \\ P(Y = 3) &= P(\log_2(X) = 3) = P(X = 8) = .3. \end{aligned}$$

Thus, for example, if we wanted to find the mean of Y we would obtain

$$\begin{aligned} \mu_Y &= 1 \cdot P(Y = 1) + 2 \cdot P(Y = 2) + 3 \cdot P(Y = 3) \\ &= 1 \cdot (.2) + 2 \cdot (.5) + 3 \cdot (.3). \\ &= 2.1. \end{aligned}$$

□

The calculation in the discrete case (as above) is very simple. In the continuous case, to get the pdf we would have to introduce a derivative factor $|\frac{dy}{dx}|$, as ordinary calculus requires when a variable is transformed (see p. 62). The point, here, is that once we know the probabilities for X, we can obtain them easily for Y using computer simulations. Suppose we can, on the computer, generate observations ("draws") from the distribution of X, and let us denote a set of G such simulated observations by $U^{(1)}, U^{(2)}, \ldots, U^{(G)}$. If we define $W^{(1)} = f(U^{(1)}), W^{(2)} = f(U^{(2)}), \ldots, W^{(G)} = f(U^{(G)})$, we obtain a set of G draws from the distribution of Y. We will refer to these simulated observations as *pseudo-data*.

Illustration: A random variable having three possible values (continued) In the discrete illustration above, suppose we wanted to find $P(Y = 1)$ without using the formula $P(Y = 1) = P(X = 2) = .2$. We could get an approximate answer by the following procedure:

1. For $j = 1$ to 10,000:
   Generate $U^{(g)}$ from the distribution of X.
   Compute $W^{(g)} = \log_2(U^{(g)})$.
2. Let N be the number of $W^{(g)}$ such that $W^{(g)} = 1$ and compute

$$P(Y = 1) \approx \frac{N}{10{,}000}.$$

To compute the mean of Y we could follow the same step 1, and then replace step 2 with

$$\mu_Y \approx \frac{1}{G}\sum_{j=1}^{G} W^{(g)}. \qquad \square$$

This computer-simulation procedure works for discrete and continuous random variables and random vectors.

Algorithm: Simulation-Based Propagation of Uncertainty Suppose the random variable or random vector X has a probability distribution from which we are able to simulate observations, and we wish to find the distribution of a random variable $Y = f(X)$ defined by a real-valued function $f(x)$. Proceed as follows:

1. For $j = 1$ to G:
   Generate $U^{(g)}$ from the distribution of X.
   Compute $W^{(g)} = f(U^{(g)})$.
2. Step 1 gives us a sample $W^{(1)}, W^{(2)}, \ldots, W^{(G)}$ from the distribution of Y. We can obtain whatever information we wish about the distribution of Y by taking G to be sufficiently large. In particular,

   (i) To get $P(a < Y < b)$ let N be the number of $W^{(g)}$ such that $a < W^{(g)} < b$ and compute

   $$P(a < Y < b) \approx \frac{N}{G}.$$

   (ii) To get σ_Y, compute the sample mean $\overline{W} = \frac{1}{G}\sum_{g=1}^{G} W^{(g)}$ and use the sample variance $\frac{1}{G-1}\sum_{g=1}^{G}(W^{(g)} - \overline{W})^2$ to get

   $$\sigma_Y \approx \sqrt{\frac{1}{G-1}\sum_{g=1}^{G}(W^{(g)} - \overline{W})^2}. \tag{9.6}$$

   (iii) To get the qth quantile of the distribution of Y use the qth sample quantile w_q (defined on p. 67) among the pseudo-data values $W^{(1)}, W^{(2)}, \ldots, W^{(G)}$. $\square$

The procedure is very general: it is applicable as long as it is possible to generate observations from the distribution of X. (The problem of creating algorithms that generate observations from a given distribution is itself a sub-specialty field of research; some additional comments about this may be found in Section 16.1.6.) When we use simulation-based propagation of uncertainty together with the approximate normality of Y, due to the results in Section 9.1.2, we have a very powerful inference engine: we can apply them, together, to obtain approximate 95 % CIs in a wide variety of settings.

Result: Simulation-Based Propagation of Uncertainty in Estimation Suppose the random vector X is a consistent estimator of a parameter vector θ having an approximate distribution from which we are able to simulate observations and we wish to estimate $\phi = f(\theta)$ for some real-valued function $f(x)$. If we apply simulation-based propagation of uncertainty, with G large, then an approximate 95 % CI for ϕ is given by $(w_{.025}, w_{.975})$ where $w_{.025}$ and $w_{.975}$ are the .025 and .975 quantiles among the pseudo-data $W^{(1)}, W^{(2)}, \ldots, W^{(G)}$.

The beauty of this simulation-based method of getting approximate confidence intervals is its simplicity and practicality, as long as it is easy to generate observations from the distribution of the estimator X. If, in addition, the estimator $\hat{\phi} = f(\hat{\theta})$ is approximately normal, then we have a slightly different option. Although it will often produce essentially the same answers, it simplifies the reporting of results by producing a standard error, which is connected to the confidence interval by the 95 % rule (p. 117).

Result: Simulation-Based Propagation of Uncertainty in Estimation When the Estimator is Approximately Normal Suppose X is an approximately multivariate normal estimator of θ having estimated variance matrix $\hat{\Sigma}$, and we want to estimate $\phi = f(\theta)$ for some real-valued (univariate) function $f(x)$. Let us take $Y = f(X)$ to be the estimator of ϕ. We will write the observed estimate of θ as $X = \hat{\theta}$ and the observed estimate of ϕ as $Y = \hat{\phi} = f(\hat{\theta})$. If the function $f(x)$ is approximately linear near $x = \hat{\theta}$ and $f'(\hat{\theta})$ is not the zero vector (i.e., not all of its partial derivatives are zero) then

1. Y is approximately normally distributed, and
2. the standard error obtained from (9.6) by simulation-based propagation of uncertainty

$$SE(\hat{\phi}) = \sqrt{\frac{1}{G-1}\sum_{g=1}^{G}(W^{(g)} - \overline{W})^2} \tag{9.7}$$

furnishes approximate inferences. In particular, an approximate 95 % CI is given by $(Y - 2SE(Y), Y + 2SE(Y))$. □

If these two methods differ, it is an indication that the distribution of $\hat{\phi}$ is noticeably non-normal and it is better to use the quantiles as they are likely to be more accurate. The second method, based on approximate normality, is justified by the theorem on p. 235 leading to (9.20).

We illustrate both methods by returning to the example involving perception of dim light.

Example 5.5 (continued from p. 221) At the beginning of the chapter we motivated propagation of uncertainty using the problem of calculating x_{50}, defined on p. 221,

and finding its standard error. If we drop the subscript i in Eq. (8.44), the logistic function used in the logistic regression model may be written in the form

$$p = \frac{\exp(u)}{1 + \exp(u)} \tag{9.8}$$

where $u = \beta_0 + \beta_1 x$. We can solve for u as follows:

$$\frac{p}{1-p} = \frac{\frac{\exp(u)}{1+\exp(u)}}{\frac{1}{1+\exp(u)}} = \exp(u)$$

and taking logs gives

$$u = \log \frac{p}{1-p}. \tag{9.9}$$

If we set $p = .5$ we get $u = 0$. In other words, x_{50} must be the value of x for which

$$\beta_0 + \beta_1 x = 0.$$

Solving for x we get (9.2), and when we plug in $(\hat{\beta}_0, \hat{\beta}_1)$ we obtain

$$\hat{x}_{50} = \frac{-\hat{\beta}_0}{\hat{\beta}_1}. \tag{9.10}$$

To get a standard error for $\hat{x}_{50}$ we propagate the uncertainty from the approximate variance matrix $\hat{\Sigma}$ for $(\hat{\beta}_0, \hat{\beta}_1)$. That is, we assume that statistical software (for logistic regression, which we discuss in Section 14.1.1) has provided the MLE $(\hat{\beta}_0, \hat{\beta}_1)$ and the variance matrix $\hat{V}$ based on the observed information matrix as in (8.41), i.e., $\hat{V} = I_{OBS}(\hat{\beta}_0, \hat{\beta}_1)^{-1}$. We can then set $\hat{\Sigma} = \hat{V}$ and apply the computer-simulation methods.

To obtain a 95 % confidence interval based on quantiles or the standard error of x_{50}, we generate many two-dimensional vectors that represent plausible values of (β_0, β_1) according to the uncertainty in $(\hat{\beta}_0, \hat{\beta}_1)$ and, for each such vector, find x_{50}. That is, we simulate two-dimensional vectors $U^{(g)} = (U_1^{(g)}, U_2^{(g)})$ whose first component corresponds to β_0 and whose second component corresponds to β_1; we then apply (9.10) to these components to get a simulated value

$$W^{(g)} = \frac{-U_1^{(g)}}{U_2^{(g)}}. \tag{9.11}$$

The distribution of $W^{(g)}$ values represents the uncertainty in x_{50} propagated from the uncertainty in $(\hat{\beta}_0, \hat{\beta}_1)$.

We now spell this out in steps. We again assume we have (from software) the MLE $(\hat{\beta}_0, \hat{\beta}_1)$ and the variance matrix $\hat{V}$. The algorithm is as follows:

1. Initialize by setting

   $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$
   $\hat{\Sigma} = \hat{V}$
   G = 1,000 (or some other suitable value).

2. For $g = 1, \ldots, G$

   simulate $U^{(g)} \sim N(\hat{\beta}, \hat{\Sigma})$
   compute $W^{(g)}$ using (9.11).

3. Set $O^{(1)}, O^{(2)}, \ldots, O^{(G)}$ equal to the ordered values of $W^{(1)}, W^{(2)}, \ldots, W^{(G)}$, so that $O^{(1)}$ is the smallest $W^{(g)}$, $O^{(2)}$ is the second smallest, etc., with $O^{(G)}$ being the largest.

   If $.025G$ is an integer, set $r_{.025} = .025G$ and if $.025G$ is not an integer set $r_{.025}$ equal to the smallest integer larger than $.025G$. (If $G = 1{,}000$ then $r_{.025} = 25$.)

   If $.975G$ is an integer, set $r_{.975} = .975G + 1$ and if $.975G$ is not an integer set $r_{.975}$ equal to the smallest integer larger than $.975G$. (If $G = 1{,}000$ then $r_{.975} = 976$.)

   Define

$$w_{.025} = O^{(r_{.025})}$$
$$w_{.975} = O^{(r_{.975})}. \tag{9.12}$$

   (If $G = 1,000$ then $w_{.025}$ is the 25th ordered value of $W^{(g)}$ and $w_{.975}$ is the 976th ordered value of $W^{(g)}$.)
   The approximate 95 % CI for x_{50} is $(w_{.025}, w_{.975})$.

4. Compute

$$SE(x_{50}) = \sqrt{\frac{1}{G-1}\sum (W^{(g)} - \overline{W}^{(g)})^2}.$$

Using the percentile-based simulation algorithm we obtained

$$\text{approx. } 95\,\% \text{ CI for } x_{.50} = (1.88, 1.96).$$

We found the standard error of $\hat{x}_{50}$ to be $SE = .019$. The usual standard-error based approximate 95 % CI is then

$$(1.92 - 2(.019), 1.92 + 2(.019)) = (1.88, 1.96)$$

in agreement with the percentile-based method. This agreement is an indication that the MLE in (9.10) is approximately normally distributed, to a close approximation, for the sample sizes in this data set. The $\log_{10}$ intensity at which subject S.S. (whose

data were shown in Fig. 8.9, the scale on the x-axis having been $\log_{10}$(intensity)) would have perceived half the flashes is estimated to have been $\hat{x}_{50} = 1.921 \pm .019$ with approximate 95 % CI (1.88, 1.96). Note that the logistic regression model (Eqs. (8.43) and (8.44)) could be viewed here as a method of interpolating between the experimental values, while also providing a standard error of the interpolated quantity. □

In the simulation procedure above, a detail left unspecified is the value of the simulation sample size G to be used, i.e., the number of random variables or vectors $U^{(g)}$ to be generated on the computer. Typically we would expect $G = 1{,}000$ to be sufficient, and when the computation is fast we might use $G = 10{,}000$ to be safe. In general the size of G to be used is an empirical matter; if in doubt, one easy way to proceed is to pick a convenient value of G, such as $G = 1{,}000$, and then run the entire procedure several times, each with a new seed to the random number generator (as is typically the default in software). Because new random variables will be generated each time the procedure is run, the several values of the outputs ($w_{.025}$, $w_{.975}$, and SE) will be different. If the output values on different runs are all close to each other then it may be concluded that these quantities of interest are sufficiently accurate. If not, the size of G must be increased.

9.1.2 *In large samples, transformations of consistent and asymptotically normal random variables become approximately linear.*

We now discuss the analytical approach to propagating uncertainty. Let us suppose we have a random variable or vector X, and a function $y = f(x)$, which we wish to apply to X. This will produce a random variable $Y = f(X)$. A handful of special cases have been analyzed in the literature (mostly many years ago), which leads to some standard distributions such as the chi-squared distribution, the t-distribution, and the F-distribution. In practice, however, one often comes across cases that do not fit any specialized framework. Fortunately, there is a simple and powerful method that may be applied in conjunction with a general theoretical result in order to get the approximate distribution of Y.

Suppose, first, that X is a random variable having mean μ_X and standard deviation σ_X. The classical idea behind what is often called *the delta method* assumes, first, that the distribution of X is concentrated around μ_X (so that σ_X is small), and, second, that the function $y = f(x)$ is approximately linear near μ_X. In addition, X is often assumed to be approximately normally distributed. Under these assumptions the linear transformation that approximates $f(x)$ is applied to X to get the approximate distribution of $Y = f(X)$. In particular, if X were normal then the theorem concerning linear transformation of a normal random variable on p. 63 would show that this linear transformation of X would be normally distributed. As a consequence (it may be shown) if X is approximately normal, then Y is approximately normal

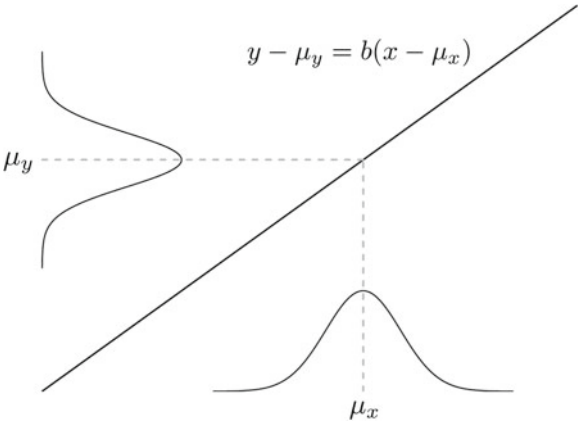

Fig. 9.1 The effect of the transformation $y = a + bx$ operating on a normally distributed random variable X having mean μ_X and standard deviation σ_X. The random variable $Y = a + bX$ is again normally distributed, with mean $\mu_Y = a + b\mu_X$ and standard deviation $\sigma_Y = |b|\sigma_X$. The normal distributions are displayed on the x and y axes; the linear transformation is displayed as a *line*, which passes through the point (μ_X, μ_Y) so that it may be written, equivalently, as $y - \mu_Y = b(x - \mu_X)$.

and the approximate mean and variance of Y is given from the approximating linear transformation, as in the theorem on p. 63.

Theorem Suppose that a sequence of random variables $X_1, X_2, \ldots, X_n, \ldots$ satisfies

$$\frac{X_n - \mu}{\sigma_{X_n}} \xrightarrow{D} N(0, 1)$$

as $n \to \infty$, and that the function $f(x)$ is continuously differentiable with $f'(\mu) \neq 0$. Then

$$\frac{f(X_n) - f(\mu)}{\sigma_{Y_n}} \xrightarrow{D} N(0, 1)$$

with $\sigma_{Y_n} = |f'(\mu)|\sigma_{X_n}$.

Proof: We omit the proof, which is a consequence of Slutsky's theorem (p. 163), but give the essential idea.

First, from the theorem on transformation of a normal random variable (p. 63), if $Y = a + bX$ and $X \sim N(\mu_X, \sigma_X^2)$ then $Y \sim N(\mu_Y, \sigma_Y^2)$ with $\mu_Y = a + b\mu_X$ and $\sigma_Y = |b|\sigma_X$. A pictorial display of this situation is given in Fig. 9.1. Now, suppose that $f(x)$ is not linear, but let us assume that it is only mildly nonlinear within the "most probable" range of X. That is, $f(x)$ is mildly nonlinear within, say, $\mu_X \pm 2.5\sigma_X$, which is the range over which we are assuming X to be approximately normally distributed. Then we may approximate $f(x)$ with the best-fitting linear approximation at $x = \mu_X$:

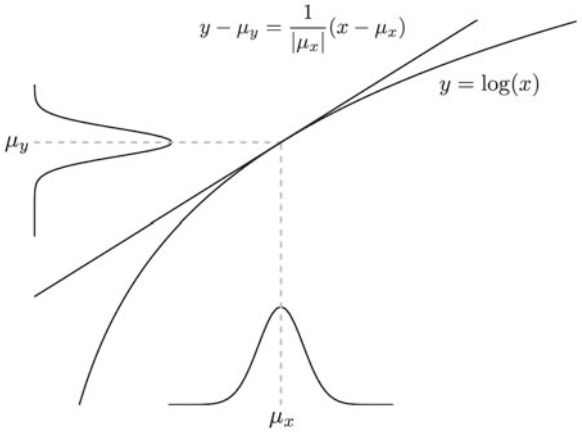

Fig. 9.2 The transformation $y = \log(x)$ operating on a normally distributed (or approximately normally distributed) random variable X having mean μ_X and standard deviation σ_X produces an approximately normally distributed random variable Y with mean and standard deviation approximately given by $\mu_Y = \log(\mu_X)$ and $\sigma_Y = \sigma_X/|\mu_X|$. The approximating *line* could also be written in the form $y - \mu_Y \approx (x - \mu_X)/|\mu_X|$.

$$f(x) \approx f(\mu_X) + f'(\mu_X)(x - \mu_X)$$

which is usually called a first-order Taylor series at $x = \mu_X$. (See the Appendix.) That is, we have

$$f(x) \approx a + bx$$

with $a = f(\mu_X) - f'(\mu_X)\mu_X$ and $b = f'(\mu_X)$. Note that $a + b\mu_X = f(\mu_X)$. As a result, we have that $Y = f(X)$ is approximately normally distributed, with $\mu_Y \approx f(\mu_X)$ and $\sigma_Y \approx |f'(\mu_X)|\sigma_X$. □

We now re-state this theorem in a less mathematically precise but more practical form.

Result: Propagation of Uncertainty in the Scalar Case If X is approximately $N(\mu_X, \sigma_X^2)$ and the function $f(x)$ is approximately linear with $f'(x) \neq 0$ near μ_X ("near" being defined probabilistically, in terms of σ_X), then
(1) $Y = f(X)$ is approximately normal, and
(2) the approximate mean and standard deviation of Y are given by $\mu_Y \approx f(\mu_X)$ and $\sigma_Y \approx |f'(\mu_X)|\sigma_X$.

Note that both conclusions in this result are important: subsequently we will rely on the approximate normality in (1) using computer simulation in place of the analytical formula for the standard deviation appearing in (2). On the other hand, the formulas are sometimes valuable.

A detail: Here is a technical point. In the statement of the theorem the numbers σ_{X_n} do not have to be the standard deviations of X_n. They can, instead, be some numbers that will serve as the approximate standard deviations. In practice, we often do not have the exact standard deviation but we do have a useful approximate value based on large-sample theory, as in Chapter 8. □

Illustration: Log transformation Suppose $g(x) = \log(x)$. Then $f'(x) = 1/x$, so that if X is approximately normal, with small σ_X, then Y is approximately normal with $\mu_Y \approx \log(\mu_X)$ and $\sigma_Y \approx \sigma_X/|\mu_X|$. The picture is given in Fig. 9.2. Careful examination of Fig. 9.2 reveals that the distribution of Y is not exactly normal (it is mildly skewed toward low values), but it is close. □

The illustration above, using the log transformation, serves to show how the analytical calculation works in propagation of uncertainty. As we stressed in Chapter 2, the log transformation is frequently used in practice to make data distributions more symmetrical. An additional benefit of the log transformation comes from its application in statistical procedures such as analysis of variance (Chapter 13) that compare observations across groups or experimental conditions, where it is typically assumed that all the observations have the same variance. Similarly, one of the standard assumptions in linear regression (Chapter 12) is that the noise or error has the same variance for all observations. Sometimes, however, this is clearly violated. Suppose it is found, empirically, that the standard deviation is proportional to the mean. The illustration above may be used to show that the log transformation removes this effect, making the variances approximately homogeneous across observations.

Specifically, suppose we have random variables $X_1, \ldots, X_m$ for which σ_{X_i} is proportional to μ_{X_i}, with all $\mu_{X_i} > 0$. We may write this using the proportionality symbol ($\propto$) as

$$\sigma_{X_i} \propto \mu_{X_i} \tag{9.13}$$

and if the proportionality constant is c we have

$$\sigma_{X_i} = c\mu_{X_i}. \tag{9.14}$$

Now let $Y_i = \log(X_i)$. Then, by the analysis in the previous illustration, using $|\mu_{X_i}| = \mu_{X_i}$ because $\mu_{X_i} > 0$, we obtain

$$\sigma_{Y_i} \approx c.$$

In this context the log transformation is called *variance stabilizing*. Improving homogeneity of variances, making them more nearly equal, is an additional motivation for the log transformation in data analysis. Here is an example.

Example 2.3 (continued from p. 29) As part of their argument that it may be advantageous to transform high-field BOLD signal infMRI data by taking

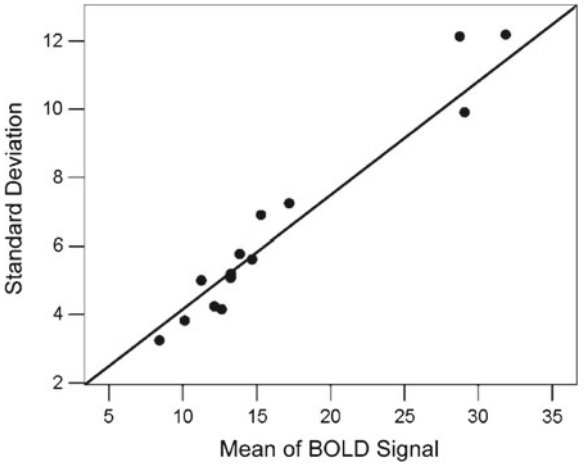

Fig. 9.3 Plot of standard deviation versus mean in BOLD signal across 15 subjects, adapted from Lewis et al. (2005). The plot is nearly linear, so the standard deviation is very nearly proportional to the mean.

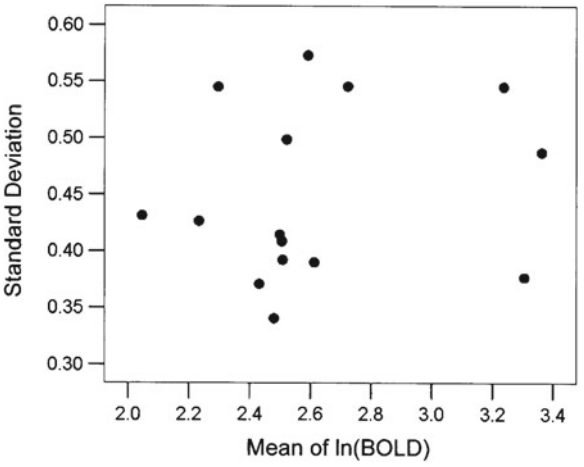

Fig. 9.4 Plot of standard deviation versus mean of log-transformed BOLD signal across 15 subjects, adapted from Lewis et al. (2005). Here, in contrast to Fig. 9.3, the standard deviation is approximately constant and shows no fixed relationship with mean.

logarithms, Lewis et al. (2005) provided plots of the standard deviation versus the mean for the BOLD signal and for the log-transformed BOLD signal. These plots are shown in Figs. 9.3 and 9.4. The standard deviation is nearly proportional to the mean for the BOLD signal, but shows no relationship to the mean of the log-transformed BOLD signal. Because standard statistical procedures assume the standard deviation

is more or less constant regardless of the mean, the authors suggested that taking logs might be a good idea. □

Example 9.1 Square-root transformation of spike counts in motor cortex When the variance of spike counts is plotted against the mean it often happens that they are roughly proportional. That is, the spike counts $X_1, \ldots, X_m$ satisfy

$$\sigma^2_{X_i} \propto \mu_{X_i}, \tag{9.15}$$

at least approximately. (There are many references to this phenomenon; see Shadlen and Newsome 1998, for some of them.) Let us rewrite (9.15) analogously with Eq. (9.14), putting it in the form

$$\sigma^2_{X_i} = c\mu_{X_i} \tag{9.16}$$

for some proportionality constant c. By examining the analysis in the foregoing illustrations of the log transformation it becomes apparent that a similar trick may be used here. From the propagation of uncertainty result $\sigma_Y \approx |f'(\mu_X)|\sigma_X$, together with (9.16) we have

$$\sigma_Y \approx |f'(\mu_X)|c\sqrt{\mu_X}. \tag{9.17}$$

In order to remove the effects in (9.16) we therefore should find $f(x)$ such that

$$f'(x) \propto 1/\sqrt{x} \tag{9.18}$$

because that will force the factors $|f'(\mu_X)|$ and $\sqrt{\mu_X}$ to cancel. The square-root function does the job: if $f(x) = \sqrt{x}$ then (9.18) is satisfied. For this reason, many authors have chosen to use square-root transformations of spike counts in their statistical analyses. In particular, Georgopoulos et al. (2000) reported improvements from a square-root transformation when fitting spike counts to direction of movement by linear regression. For a similar reason, Yu et al. (2009) used square-root transformations of spike counts in studying "neural trajectories" that summarize population activity in motor cortex during movement planning. □

We now extend the propagation of uncertainty argument to the vector case, which involves a multivariate linear approximation (a first-order Taylor series expansion). The idea is to take a sequence of random vectors $X_1, X_2, \ldots$ that are approximately multivariate normal and apply the function $f(x)$ to each of them and, as in the scalar case above, approximate $f(x)$ using a first-order Taylor series based on the derivative of $f(x)$. In this multidimensional case the derivative becomes the vector of partial derivatives. Specifically, for a vector x we let $f'(\mu)$ be the vector of partial derivatives (with respect to all components) of the real-valued function $f(x)$, evaluated at $x = \mu$. That is, the ith component of this derivative is

$$f'(\mu)_i = \left.\frac{\partial f}{\partial x_i}\right|_{x=\mu}.$$

Result: Multivariate Propagation of Uncertainty If X is approximately multivariate normal, given by $N_m(\mu_X, \Sigma_X)$, and the function $f(x)$ is approximately linear with $f'(x) \neq 0$ near μ_X ("near" again being defined probabilistically), then
(1) $Y = f(X)$ is approximately normal, and
(2) the approximate normal mean and standard deviation are given by $\mu_Y \approx f(\mu_X)$ and

$$\sigma_Y \approx \sqrt{f'(\mu_X)^T \Sigma_X f'(\mu_X)}. \tag{9.19}$$

Details: To see how we get this, consider the bivariate case. If we have $z = f(x, y)$ and we apply a first-order Taylor series expansion (a linear approximation) near a point (x_0, y_0), we get

$$z \approx f(x_0, y_0) + \frac{\partial f}{\partial x}\Big|_{(x_0,y_0)}(x - x_0) + \frac{\partial f}{\partial y}\Big|_{(x_0,y_0)}(y - y_0).$$

Analogously to what was done in the scalar case, we insert random variables X and Y and replace (x_0, y_0) with (μ_X, μ_y). With $Z = f(X, Y)$ we note that the first term in the variance $\sigma_Z^2 = V(Z)$ is $V(f(x_0, y_0)) = 0$ (because the variance of a constant is 0), and we then get

$$\sigma_Z^2 = \left(\frac{\partial f}{\partial x}\Big|_{(x,y)=(\mu_X,\mu_Y)}\right)^2 \cdot \sigma_X^2 + \left(\frac{\partial f}{\partial y}\Big|_{(x,y)=(\mu_X,\mu_Y)}\right)^2 \cdot \sigma_Y^2 + 2 \cdot \frac{\partial f}{\partial x}\Big|_{(x,y)=(\mu_X,\mu_Y)} \frac{\partial f}{\partial y}\Big|_{(x,y)=(\mu_X,\mu_Y)} \rho\sigma_X\sigma_Y.$$

The general multidimensional case is analogous. □

The result relies on the following theorem.

Theorem Let μ be an m-dimensional vector, and let $f(x)$ be a differentiable function for which $f'(\mu) \neq 0$. If $X_1, X_2, \ldots, X_n, \ldots$ is a sequence of m-dimensional random vectors and Σ_n is a sequence of positive definite symmetric matrices such that for every nonzero m-dimensional vector w,

$$w^T \Sigma_n^{-1/2}(X_n - \mu) \xrightarrow{D} N(0, 1),$$

then, writing $Y_n = f(X_n)$, we have

$$\frac{(Y_n - f(\mu))}{\sigma_Y} \xrightarrow{D} N(0, 1) \tag{9.20}$$

where

$$\sigma_Y = \sqrt{f'(\mu)^T \Sigma_n f'(\mu)}.$$

Proof: Omitted. □

Example 9.2 Neural firing rate selectivity index In single-unit electrophysiological studies, neural firing rates are often estimated under two experimental conditions. Let us label the conditions A and B, and suppose that for each neuron we have many trials of recordings under each of the conditions. Averaging across the trials gives sample mean firing rates, $\bar{X}_A$ and $\bar{X}_B$, which may be compared. However, comparisons are made across many neurons having quite different firing rates. For this reason, some sort of normalization is usually invoked. One commonly-used comparative measure is the index

$$Y = \frac{\bar{X}_A - \bar{X}_B}{\bar{X}_A + \bar{X}_B}. \tag{9.21}$$

For example, Roesch and Olson (2004) compared activity of neurons in the orbitofrontal (OF) cortex under conditions involving large reward for success in an eye movement task, a large penalty for failure (a time out for the monkey), or neither (i.e., a small reward and a small penalty). The authors compared the large reward to the neutral condition using a measure of the form (9.21),with condition A being large reward and B being neutral. This would identify neurons that tended to respond to expected reward. It would be possible for a neuron to respond not specifically to reward but to the importance of success, which the authors termed "motivation." Both large reward and large penalty should increase the subject's motivation to perform the task.The authors also compared the large penalty to the neutral condition using a measure of the form (9.21), with A representing the large penalty condition and B being neutral. By examining many neurons they concluded that neurons in the OF cortex tend to fire more with large expected reward, and tend to fire less with large expected penalty. They went on to contrast this with premotor cortex where neurons tended to fire more with both large expected reward and large expected penalty. They characterized the results as suggesting that OF cortex was more involved in reward processing while PM activity tended to reflect motivation.

To put this in the general framework we write $X_1 = \bar{X}_A, X_2 = \bar{X}_B, X = (X_1, X_2)$, and then

$$f(x) = \frac{x_1 - x_2}{x_1 + x_2}.$$

The problem of finding the standard error of Y defined by (9.21) then becomes a special case of the general problem of finding the standard error of $Y = f(X)$ when the uncertainty in X is known.

In Example 12.3 we discuss an application of the difference index for firing rates where propagation of uncertainty was used to obtain interesting results.Example 12.3

is based on Behseta et al. (2009), which provides some details about propagation of error for the difference index. □

Additional details: We may also propagate uncertainty analytically to $x_{50} = f(\beta_0, \beta_1)$ using Eq. (9.19) with (9.10), which gives the standard error

$$SE = \sqrt{f'(\hat{\beta}_0, \hat{\beta}_1)^T \hat{\Sigma} f'(\hat{\beta}_0, \hat{\beta}_1)}$$

where the partial derivates are

$$\frac{\partial f}{\partial \beta_0} \Big|_{(\hat{\beta}_0, \hat{\beta}_1)} = -\frac{1}{\hat{\beta}_1}$$
$$\frac{\partial f}{\partial \beta_1} \Big|_{(\hat{\beta}_0, \hat{\beta}_1)} = \frac{\hat{\beta}_0}{\hat{\beta}_1^2}.$$

Plugging into the formulas above the values of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\Sigma}$, the $\log_{10}$ intensity at which subject S.S. would have perceived half the flashes is estimated to have been $\hat{x}_{50} = 1.921 \pm .019$. This agrees with the approximate 95 % CI obtained by the simulation method. □

9.2 The Bootstrap

The *bootstrap* is a very simple way to obtain standard errors and confidence intervals. It has turned out to be one of the great inventions in the field of statistics. In Section 9.2.1 we explain the essential idea, and we contrast the *parametric bootstrap* with the *nonparametric bootstrap*, elaborating on these two distinct methods in Sections 9.2.2 and 9.2.3.

9.2.1 The bootstrap is a general method of assessing uncertainty.

The algorithm for simulation-based propagation of uncertainty (p. 225) began with a random vector X having a known distribution (from which observations could be generated on the computer). In practice, applying the result on p. 226, X becomes an estimator of a parameter vector θ and its distribution is known approximately; typically it is a normal distribution. From this, uncertainty can be propagated from X to an estimator $\hat{\phi}$ of $\phi = f(\theta)$. As illustrated in Example 5.5 on p. 226, an essential input to the algorithm is the variance matrix of X (in Example 5.5 we had $X = (\hat{\beta}_0, \hat{\beta}_1)$ and used $\hat{\Sigma} = I_{OBS}(\hat{\beta}_0, \hat{\beta}_1)^{-1}$). But what if it is difficult to compute the variance matrix of X? The bootstrap instead backs up a step, using the variation in the data

themselves so that an explicit form for the variance matrix of X becomes unnecessary (and the variance matrix of X can, in fact, also be obtained from the bootstrap).

Here is the idea. Let us suppose $X_1, \ldots, X_n$ is a random sample from a distribution having distribution function $F_X(x)$. We write this as $X_i \sim F_X$, independently, for $i = 1, \ldots, n$. We wish to find the standard error of a scalar statistic $T = T(X_1, \ldots, X_n)$. Notice, as we have said before, that T is obtained by applying some mapping to the random variables. Let us emphasize this still further by using the function $h(x_1, x_2, \ldots, x_n)$ to denote that mapping so that $T(X_1, X_2, \ldots, X_n) = h(X_1, X_2, \ldots, X_n)$. In the case of ML estimation, for instance, $h(x_1, x_2, \ldots, x_n)$ would be the function that gives the value of the MLE for a particular set of data $x_1, \ldots, x_n$. In some cases the function $h(x_1, x_2, \ldots, x_n)$ is explicit, as in ML estimation of the binomial propensity p, while in other cases it is implicit—the result of solving a differential equation, as in ML estimation of β_1 in the logistic regression model of Example 5.5 (p. 214). In either situation, however, $SE(T)$ is defined as the standard deviation of $T = h(X_1, X_2, \ldots, X_n)$ when the X_i random variables follow the distribution with cdf F_X. Now, if we were able to simulate observations from F_X on the computer, we could simulate G samples where G is a large number, proceeding as follows:

1. For $g = 1$ to G
   Generate a sample $U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)}$ from F_X
   Compute $W^{(g)} = h(U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)})$
2. Compute $\overline{W} = \frac{1}{G}\sum_{i=1}^{G} W^{(g)}$ and then

$$SE_{sim}(T) = \sqrt{\frac{1}{G-1}\sum_{g=1}^{G}(W^{(g)} - \overline{W})^2}.$$

Step 1 of this scheme would evaluate the estimator T on all the sets of *pseudo-data* $U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)}$ for $g = 1, \ldots, G$. Each set of simulated values $U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)}$ may also be called a *sample of pseudo-data*. The squared value $SE_{sim}(T)^2$ is simply the sample variance of the $W^{(g)}$ random variables, and for large G it would become close to the variance $V(T)$ (because, in general, the sample variance converges to the theoretical variance, in probability, as in Section 7.3.4). Thus, for large G we would get $SE_{sim}(T) \approx SE(T)$.

The only problem with the scheme as we have described it so far is that, in practice, we don't know the distribution F_X, so we don't know how to generate the pseudo-data. This situation is similar to the one we found in Section 7.3.4 where we could not compute $SE(\bar{X}) = \sigma_X/\sqrt{n}$ because we did not know σ_X. There, we solved the problem by substituting s for σ_X, which is often called a *plug-in* estimate, and this worked because the plug-in estimate is consistent, i.e.,

$$S \xrightarrow{P} \sigma_X \tag{9.22}$$

which is the same as (7.19). The idea of the bootstrap is analogous: we replace F_X by an estimate of it and then apply the algorithm above. If we have a parametric model and we use ML estimation to estimate the parameters, we can use the model with the fitted parameters to generate the pseudo-data $U_1^{(g)}, \ldots, U_n^{(g)}$. This scheme is called the *parametric bootstrap*. Otherwise, we replace F_X by the empirical cdf $\hat{F}_n$ and draw the pseudo-data $U_1^{(g)}, \ldots, U_n^{(g)}$ from $\hat{F}_n$. This is the *nonparametric bootstrap.* Both methods extend to cases in which we replace scalar estimates (e.g., $\hat{\beta}_1$) by vectors of estimated quantities (e.g., $(\hat{\beta}_0, \hat{\beta}_1)$).

The parametric bootstrap and nonparametric bootstrap both begin, conceptually, by estimating the data distribution F_X. The parametric bootstrap uses a specific assumption, such as normality of the data. The nonparametric bootstrap does not require any specific data distributional assumption, and this is the sense in which it is "nonparametric." The nonparametric bootstrap is also usually easier to implement. Its disadvantage is that it requires i.i.d. random variables to represent the variation in the data. There are many cases where the data are not modeled as i.i.d., such as in regression, time series, and point processes. Sometimes a clever transformation makes the nonparametric bootstrap applicable (see Davison and Hinkley 1997, for examples), but in other cases the parametric bootstrap is either the only available approach or at least a more straightforward methodology to apply. Both forms of bootstrap use propagation of uncertainty.

9.2.2 The parametric bootstrap draws pseudo-data from an estimated parametric distribution.

Suppose we assume that a set of data $x_1, x_2, \ldots, x_n$ is a random sample from a distribution with pdf $f(x_i|\theta)$, and we estimate θ with the MLE $\hat{\theta}$. If we assume for the moment that the parameter θ is a scalar then, according to the scheme in Section 9.2.1, we may obtain the standard error of $\hat{\theta}$ as $SE_{sim}(\hat{\theta})$ by generating pseudo-samples $U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)}$ from the distribution with pdf $f(x_i|\theta)$. Because we do not know the value of θ we plug in the MLE $\hat{\theta}$ and instead generate pseudo-samples from the distribution with pdf $f(x_i|\hat{\theta})$. This is a *parametric bootstrap*, and the resulting value of $SE_{sim}(\hat{\theta})$ is a *parametric bootstrap* standard error.

Algorithm: Parametric bootstrap estimate of standard error To obtain the standard error $SE(\hat{\theta})$ we proceed as follows:

1. For $g = 1$ to G
   Generate a random sample $U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)}$ from the distribution having pdf $f(x_i|\hat{\theta})$.
   Find the MLE $\hat{\theta}^{(g)}$ based on $U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)}$ and set $W^{(g)} = \hat{\theta}^{(g)}$.
2. Compute $\overline{W} = \frac{1}{G}\sum_{i=1}^{G} W^{(g)}$ and then

$$SE(\hat{\theta}) = \sqrt{\frac{1}{G-1}\sum_{g=1}^{G}(W^{(g)} - \overline{W})^2}.$$

□

Why does the parametric bootstrap work? As in (9.22), the plug-in estimator $\hat{\theta}$ satisfies

$$\hat{\theta} \xrightarrow{P} \theta \tag{9.23}$$

which is part of the statement in (8.32). Let us write the cdf corresponding to $f(x_i|\theta)$ in the form $F_X(x|\theta)$. From (9.23) it follows that

$$F_X(x|\hat{\theta}) \xrightarrow{P} F_X(x|\theta) \tag{9.24}$$

for all x (we omit details), which is a formal way of saying that the distribution of pseudo-data based on the distribution having pdf $f(x_i|\hat{\theta})$ will be close to the distribution of the data (which has pdf $f(x_i|\theta)$). Thus, simulating pseudo-data is very much like simulating new data from the same distribution as the original data.

When θ is a vector, the same method may be used to estimate the value $f(\theta)$ of any real-valued function $f(x)$. We modify the procedure as follows.

Algorithm: Parametric bootstrap when estimating $f(\theta)$ Suppose we want to find the standard error of $f(\hat{\theta})$ and get an approximate 95 % CI for $f(\theta)$. We proceed as follows:

1. For $g = 1$ to G
   Generate a random sample $U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)}$ from the distribution having pdf $f(x_i|\hat{\theta})$.
   Find the MLE $\hat{\theta}^{(g)}$ based on $U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)}$ and set $W^{(g)} = f(\hat{\theta}^{(g)})$.
2. Compute $\overline{W} = \frac{1}{G}\sum_{i=1}^{G} W^{(g)}$ and then

$$SE(f(\hat{\theta})) = \sqrt{\frac{1}{G-1}\sum_{g=1}^{G}(W^{(g)} - \overline{W})^2}. \tag{9.25}$$

In addition, an approximate 95 % CI for $f(\theta)$ is given by

$$\text{approx. } 95\,\%\text{ CI} = (w_{.025}, w_{.975}) \tag{9.26}$$

where $w_{.025}$ and $w_{.975}$ are the sample quantiles defined from the ordered $W^{(g)}$ values as in (9.12).

If we have several functions $f_1(\theta), f_2(\theta), \ldots, f_k(\theta)$ we may obtain approximate 95 % CIs for each using (9.26) and we can get an approximate variance matrix

$$\hat{V} = \hat{V}(f_1(\hat{\theta}), f_2(\hat{\theta}), \ldots, f_k(\hat{\theta})),$$

by following step 1, above, for each of $f_1(\theta), f_2(\theta), \ldots, f_k(\theta)$ to get

$$W_j^{(g)} = f_j(\hat{\theta}^{(g)})$$

for $j = 1, \ldots, k$, and then setting $\hat{V}$ equal to the sample variance matrix (see p. 90) of the k-dimensional vectors $W^{(g)} = (W_1^{(g)}, \ldots, W_k^{(g)})$. □

Example 8.2 (continued from p. 193) In discussing the way previous seizures affect the relationship between spike width and preceding inter-spike interval length we displayed results based on change-point models. The statistical model assumed that, on average, Y decreases quadratically with x for $x < \tau$ but remains constant for $x \geq \tau$, with τ being the change point. In Fig. 8.7 we displayed fitted change-points together with standard errors, which led to the conclusion that the seizure group reset to baseline average spike widths earlier than the control group. We said that the standard errors shown in Fig. 8.7 were based on a parametric bootstrap. The specifics of computing the bootstrap standard errors followed the steps given above: based on the fitted $\hat{\tau}$, together with the fitted parameters for the quadratic relationship when $x < \tau$ and the constant relationship when $x \geq \tau$ (see p. 408), pseudo-data samples were generated and for the gth such sample a value $\hat{\tau}^{(g)}$ was calculated following the same procedure that had been used with the real data; then formula (9.25) was applied. □

There are modifications of the bootstrap confidence interval procedure that offer improvements. These are reviewed by DiCiccio and Efron (1996). Particularly effective[3] are the *bias-corrected and accelerated* (or BC_a) intervals, which are often used as defaults in bootstrap software.

9.2.3 The nonparametric bootstrap draws pseudo-data from the empirical cdf.

In Section 9.2.2 we showed how the parametric bootstrap is used to get standard errors and confidence intervals. The key theoretical point was captured by Eq. (9.24), which says that, for large samples, the distribution of the pseudo-data based on the MLE plug-in estimate will be close to the distribution of the data. The idea of the nonparametric bootstrap is to generate pseudo-data, instead, from the empirical cdf

[3] The bootstrap approximate 95% CI based on percentiles in Eq. (9.26) has the property that as $n \to \infty$ the probability of coverage is $.95 + \eta_n$ where η_n vanishes at the rate of $1/\sqrt{n}$. The BC_a intervals have the analogous property with η_n vanishing at the rate $1/n$, which means the theoretical coverage probability should be closer to .95.

$\hat{F}_n(x)$, defined on p. 64. The theoretical justification for this is given by the theorem on p. 145, which says that[4] for i.i.d. random variables

$$\hat{F}_n(x) \xrightarrow{P} F_X(x). \tag{9.27}$$

This has a form very similar to (9.24). In words, for large samples, the distribution of pseudo-data generated from the empirical cdf will be close to the distribution of the data. The advantage of this nonparametric formulation is the reduction of assumptions: we do not have to rely on a specific parametric model, but rather can assume only that we are dealing with an i.i.d. sample.

How do we generate observations from the empirical cdf $\hat{F}_n$? This turns out to be very easy. According to its definition (on p. 64), the empirical cdf assigns probability $\frac{1}{n}$ to each observation in the sample $x_1, x_2, \ldots, x_n$. This means that in order to draw a single observation from the distribution $\hat{F}_n$, we randomly select one of the values $x_1, x_2, \ldots, x_n$, with each value having probability $\frac{1}{n}$. In order to draw a set of pseudo-data, we simply repeat this process n times. In doing so the procedure is likely to produce repeats: we are sampling the values $x_1, x_2, \ldots, x_n$ each time; this is called *sampling with replacement*; we "replace" each value after sampling it, before drawing again from all the values $x_1, x_2, \ldots, x_n$. Using standard statistical software it is easy to draw samples with replacement from a set of data.

Because we are sampling the sample of data, the process is often called *resampling*. Bootstrap resampling is beautifully simple. We define the algorithm in terms of any consistent estimator T of an unknown quantity ϕ. Here, ϕ could be defined in terms of a parameter vector $\phi = f(\theta)$ or it could be defined from the data distribution F_X without reference to any parameter vector (e.g., ϕ could be the median of the distribution F_X). The algorithm is as follows:

Algorithm: Nonparametric bootstrap for an estimator T of ϕ To get a nonparametric bootstrap approximate 95 % CI for ϕ from a sample $x_1, \ldots, x_n$ based on $T = h(X_1, \ldots, X_n)$, and to get the nonparametric bootstrap $SE(T)$, we proceed as follows:

1. For $g = 1$ to G

   Generate a sample $U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)}$ by resampling, with replacement, the observations $x_1, \ldots, x_n$.

   Compute $T^{(g)} = h(U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)})$.

---

[4] Actually, a stronger result is needed, and it is stated in terms of the *supremum* (also known as the *least upper bound*). The supremum of a set of numbers $S(x)$, written $\sup_x S(x)$, is the smallest value c such that $S(x) < c$. (Thus the alternative name, "least upper bound.") It is used when $S(x)$ is bounded but does not reach a maximum across the range of x. The stronger version of the result in the theorem is that the convergence is uniform in the sense that

$$\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{P} 0.$$

This holds when $F(x)$ is a continuous cdf, and in many other cases.

2. Set $O^{(1)}, O^{(2)}, \ldots, O^{(G)}$ equal to the ordered values of $T^{(1)}, T^{(2)}, \ldots, T^{(G)}$, so that $O^{(1)}$ is the smallest $T^{(g)}$, $O^{(2)}$ is the second smallest, etc., with $O^{(G)}$ being the largest.

   If $.025G$ is an integer, set $r_{.025} = .025G$ and if $.025G$ is not an integer set $r_{.025}$ equal to the smallest integer larger than $.025G$.

   If $.975G$ is an integer, set $r_{.975} = .975G + 1$ and[5] if $.975G$ is not an integer set $r_{.975}$ equal to the smallest integer larger than $.975G$.

   Define

$$\begin{aligned} t_{.025} &= O^{(r_{.025})} \\ t_{.975} &= O^{(r_{.975})}. \end{aligned} \tag{9.28}$$

   The approximate 95 % CI for ϕ is $(t_{.025}, t_{.975})$.
3. Compute $\bar{T} = \frac{1}{G}\sum_{i=1}^{G} T^{(g)}$ and then

$$SE(T) = \sqrt{\frac{1}{G-1}\sum_{g=1}^{G}(T^{(g)} - \bar{T})^2}.$$

□

This extends immediately to the case in which each X_i, and thus each $U_i^{(g)}$, is a random vector; the algorithm above is unchanged. As with the parametric bootstrap (see p. 241), modifications to the percentile-based intervals can offer improvements and the bias-corrected and accelerated (or BC_a) intervals are often used as defaults in bootstrap software.

In practice, the parametric and nonparametric bootstraps often produce very similar confidence intervals and standard error assessments, so that the choice between them may depend on convenience. There are important examples (e.g., in time series) where the data do not form an i.i.d. sample and it can be difficult or impossible to use the nonparametric bootstrap, but in many situations it is easy to take advantage of theoretically identical replications, and resample the data.

Example 9.2 (continued from p. 236) In the SEF example introduced in Chapter 1 there were two experimental conditions, and the problem was to compare the firing rates of a neuron under each of these conditions based on a limited number of trials. In a particular time interval we found mean firing rates of 48 spikes per second for the spatial condition versus 70 spikes per second for the pattern condition. As we have noted previously, because studies involve many neurons with varying firing rates, it is common to examine the difference index

[5] With this convention, if $G = 1{,}000$ then there are 24 values smaller than $r_{.025}$ and 24 values larger than $r_{.975}$. If $G = 100$ there are 2 values smaller than $r_{.025}$ and 2 values larger than $r_{.975}$.

$$Y = \frac{\bar{X}_A - \bar{X}_B}{\bar{X}_A + \bar{X}_B}.$$

In Section 9.1.1 we discussed generation of a standard error for T using propagation of uncertainty based on the asymptotic normality of $\bar{X}_A$ and $\bar{X}_B$. An alternative would be to apply the nonparametric bootstrap procedure given above. These would give very similar results, but let us make sure it is clear how the bootstrap would be applied. For each g in step 1 we would first draw a random samples of size 15 from the 15 firing rates under the spatial condition and another random sample of size 15 from the 15 firing rates under the pattern condition; we would compute the two sample means to get $\bar{X}_A^{(g)}$ and $\bar{X}_B^{(g)}$; then we would apply the difference index formula to get

$$Y^{(g)} = \frac{\bar{X}_A^{(g)} - \bar{X}_B^{(g)}}{\bar{X}_A^{(g)} + \bar{X}_B^{(g)}}.$$

Having obtained $Y^{(1)}, Y^{(2)}, \ldots, Y^{(G)}$ (where we would take something like $G = 1{,}000$), we would go to step 2 and, to find an approximate 95 % CI, we would order the values $Y^{(1)}, Y^{(2)}, \ldots, Y^{(G)}$ and compute the resulting 2.5 and 97.5 percentiles. In Step 3 we would compute the mean and apply the formula for the standard error. □

Example 9.3 (continued from p. 187) As we said in Section 8.1, one of the questions asked by Olson et al. was whether SEF neurons tend to reach their maximal firing rate later under one of the experimental conditions (the "pattern" condition) than under the other (the "spatial" condition). To answer this, each neuron's PSTH, under each condition, was smoothed as in Fig. 8.3 (with methods described in Chapter 15), and then the time $t_{\max}$ at which the maximum occurred was computed. This was regarded as an estimator of the time τ of maximal firing rate. Olson et al. applied bootstrap methods. To get a bootstrap confidence interval for τ the nonparametric bootstrap algorithm above can be applied: we set $\phi = \tau$ and in step 1, for each g, the individual trials (each of which provides a spike train, as in Fig. 8.3) would be resampled, then the resulting pseudo-data would be used to get a PSTH, this PSTH would be smoothed, and a value $T^{(g)} = t_{\max}^{(g)}$ would be computed; then step 2 would be carried out. □

The point to be taken from these examples is that the nonparametric bootstrap, like the parametric bootstrap, can produce confidence intervals relatively easily, even for complicated estimation procedures: in step 1 of the algorithm we simply re-run the estimation procedure from start to finish using each set of pseudo-data rather than the original data. Step 2 is then accomplished with just a few software commands. When the data may be considered i.i.d. samples the nonparametric bootstrap is typically even easier than the parametric bootstrap because resampling the data may be accomplished with a single software command.

The nonparametric bootstrap has been studied extensively, and has been shown to work well in a variety of theoretical and empirical senses. For more information about the bootstrap, see Efron and Tibshirani (1993) and Davison and Hinkley(1997).

An important caveat is that arbitrary shuffles of the data do not necessarily produce bootstrap samples. The key assumption is *independent and identically distributed* sampling of $X_1, \ldots, X_n$, so that the key result (9.27) applies. Many problems may be put in this form, but the nonparametric bootstrap only applies once they are.

9.3 Discussion of Alternative Methods

At the beginning of this chapter we considered the data on perception of dim light to illustrate propagation of uncertainty according to the diagram in (9.4). We went on to discuss analytical propagation of uncertainty, simulation-based propagation of uncertainty, and then both the parametric and non-parametric bootstrap methods of obtaining uncertainty about the target estimand, in this case x_{50}, the intensity at which a flash of light is perceived 50 % of the time.

The choice among these methods is largely a matter of convenience. It is often easy to obtain the variance matrix of the parameter MLEs and then simulation-based propagation of uncertainty is easy to implement. Sometimes it is also easy to get the derivatives analytically, and the analytical approach becomes an option. The percentile method of getting confidence intervals from simulation becomes more accurate than that based on $\pm 2SE$ when the nonlinearity in the target estimand as a function of the parameters is pronounced (relative to the uncertainty in the parameters, as explained in Section 9.1.2). With i.i.d. data the nonparametric bootstrap is very easy to apply, and is often the preferred method. But many examples involve non-i.i.d. data. In regression or time series contexts, for instance, nonparametric bootstrap methods require modification and may be difficult or impossible to apply (this is the case for some point process models of neural spike train data). In such settings the parametric bootstrap is often used.

These methods can produce valid 95 % confidence intervals, which cover the estimand 95 % of the time, when the statistical model is correct and the sample size is sufficiently large. The statistical model used with the nonparametric bootstrap, in the form we have presented, assumes i.i.d. sampling but is otherwise very general. All of the methods aim to provide an appropriate spread of the confidence interval about the estimate, which is what leads to the correct coverage probability. The bias in the estimator is ignored because, for sufficiently large samples, it becomes vanishingly small. Furthermore, as we noted in Chapter 8, the bias squared often becomes vanishingly small faster than the variance becomes vanishingly small, so that the MSE is dominated by the variance. In practice, however, it is worth remembering that nontrivial bias in the estimator can greatly diminish the coverage probability of a putatively 95 % confidence interval. If a statistical model is grossly incorrect because, for example, some important explanatory factor has not been considered, then these procedures will not perform well. For reasonably good models bootstrap

methods are remarkably reliable, with large samples. Of course, with small samples the coverage probability can be highly inaccurate but, in such cases, there may be too little information to draw useful statistical inferences.

Chapter 10
Models, Hypotheses, and Statistical Significance

The notion of *hypothesis* is fundamental to science. Typically it refers to an idea that might plausibly be true, and that is to be examined or "tested" with some experimental data. Sometimes, the expectation is that the data will conform to the hypothesis. In other situations, the hypothesis is introduced with the goal of refuting it. In either case, however, variation and experimental noise prevent a perfect determination of the veracity of the hypothesis. In reality, the hypothesis will at best predict only approximately the results of an experiment. But then, one might ask, in order to be judged favorably, how close to the data should a theoretical prediction be? Development of a systematic method of answering this question, the chi-squared *goodness-of-fit* test, was one of the great advances in the early part of the twentieth century.

We describe chi-squared tests in Section 10.1. The idea is to use a statistical model to represent the theoretical predictions of the hypothesis. In this setting the model embodies the hypothesis, and we usually speak of assessing the fit of the model, as opposed to the accuracy of the hypothesis. The statistical model assigns probabilities to possible data outcomes, and if the experimental data turn out to be very rare—according to the model—then the model is deemed a poor fit. Because the chi-squared procedure analyzes the discrepancy between model prediction and data outcome, it might better be called, as John Tukey suggested, a "badness-of-fit" test. On the other hand, it is often applied as a way of checking that a model fits reasonably well—the expectation, or hope, being that it does.

When, instead, there is great interest in the possibility that the hypothesis may be wrong, we usually label it a *null hypothesis*, and if the data provide sufficient evidence against the null hypotheses we speak of *rejecting* it. Ronald Fisher introduced the general concept of *p-value*, with *p* standing for probability, to quantify the rarity of the data outcome under a null hypothesis. The notion is that when *p* is small, the data outcome is rare under the hypothesis, and thus casts doubt on the hypothesis. Fisher worked out specific procedures for obtaining *p*-values in many important problems, and his methodology became standard practice. We introduce *p*-values in the context of chi-squared tests, in Section 10.1.3, and we discuss the general framework and methodology in Section 10.3.

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_10,

The null hypothesis and p-value are only part of the standard approach to testing hypotheses. An additional idea is to introduce a specific *alternative hypothesis*, which has the potential to replace the null. In the 1930s Jerzy Neyman and Egon Pearson provided a theoretical framework that explicitly included an alternative hypothesis. Specifically, Neyman and Pearson defined *type one error* (usually written *Type I*) as the probability of incorrectly rejecting the null hypothesis and *type two error* (*Type II*) as the probability of incorrectly rejecting the alternative hypothesis. The theory considers both kinds of errors, and analyzes statistical hypothesis tests according to the probabilities of making these errors. We go over the fundamental elements of the Neyman-Pearson framework in Section 10.4, and we also discuss several different points of view about the statistical assessment of hypotheses. The terminology *hypothesis test* sometimes connotes application of the Neyman-Pearson framework. In our discussion here we use "hypothesis test" and *significance test*, interchangeably, without meaning to imply any particular theoretical posture.

It is unconventional to present goodness-of-fit tests before other hypothesis tests. Our preference for this ordering[1] is due to the smaller number of concepts and issues that arise in goodness-of-fit testing: from a pedagogical point of view, in this context it is easier to concentrate on the logic of p-values. We discuss other kinds of null hypotheses in Section 10.2.

10.1 Chi-Squared Statistics

We have described several studies where a theoretical model seemed to fit the data well and was then used for scientific inference. For instance, the Hardy-Weinberg binomial model fit well the nicotinic acetylcholine receptor and ADHD data in Example 5.1, the Poisson distribution was used to fit quantal response in synaptic transmission data in Example 5.6, the normal distribution fit well the background noise in MEG in Example 1.2, and the exponential and gamma distributions were used to fit ion channel opening duration data in Example 3.5. Previously we judged fit simply by looking at tables and graphs, informally. The chi-squared procedure provides a probabilistic quantification of the observed discrepancy between theoretical prediction and data.

The essence of goodness-of-fit assessment is as follows:

(i) We define a statistical model that assigns probabilities to potentially-observed outcomes;
(ii) We compute the discrepancy between the data values and the values obtained from the fitted model; and

[1] This order of presentation is the one followed by Fisher in his immensely influential *Statistical Methods for Research Workers*, but it seems to have been abandoned later in the twentieth century as the Neyman-Pearson approach became dominant.

(iii) Assuming the data were generated by the hypothetical model, we determine whether the observed discrepancy would be considered rare; if observing such a large discrepancy constitutes a sufficiently rare event, then we consider this to be evidence that the model does *not* hold.

The discrepancy between observed data and fit is evaluated using a statistic, here a *chi-squared statistic*, and its rarity is judged by comparing the observed value to a suitable probability distribution, here a chi-squared distribution, according to the *p*-value. The chi-squared statistic is used when each observation may be considered to arise as one of several possible categories.

10.1.1 The chi-squared statistic compares model-fitted values to observed values.

To assess the fit of a theoretical model to a set of data we begin with the obvious idea of examining the discrepancy between the model predictions and the data values.

Example 5.1 (continued, see p. 107) In Chapter 5, on p. 107, we displayed data from a study of genotype frequencies for the nicotinic acetylcholine receptor subunit $\alpha 4$ gene among children with ADHD and their parents. The table of frequencies (for a $T \rightarrow C$ exchange in one base in the gene sequence) among the 136 parents in the Kent et al. study is given again below:

	TT	CT	CC
Number	48	71	17
Frequency	.35	.52	.13
Hardy-Weinberg probability	.38	.47	.15
Hardy-Weinberg expected number	51.7	63.9	20.4

We noted previously that the frequencies and Hardy-Weinberg probabilities are quite close. We have now added a fourth line in the table to indicate the predicted or "expected" number of each genotype. To judge the fit of the model we evaluate the discrepancy between the values in the first and last lines of this table. □

In Example 5.1 there are many possible ways to measure the discrepancy between the vector of observed values (48, 71, 17) and the vector of theoretically-expected values (51.7, 63.9, 20.4). The most common assessment is based on the chi-squared statistic. Let us denote observed values by O and theoretically-expected values by E, so that the first pair of O and E values are 48 and 51.7, the second pair are 71 and 63.9, and the third pair are 17 and 20.4. The chi-squared statistic is

$$\chi^2_{obs} = \sum \frac{(O-E)^2}{E} \tag{10.1}$$

where the sum is over all pairs of values, in this case the three pairs, and we have used the subscript on χ^2_{obs} to indicate that it is calculated from the observed data. A large χ^2_{obs} indicates a failure of the model to fit the data. But how do we know when χ^2_{obs} should be considered large? The O values surely will, by chance fluctuation, deviate from the theoretical E values. The key is that when the theoretical model is valid the magnitude of this chance fluctuation becomes predictable.

To motivate χ^2_{obs} let us note that each O value is a count, counts are usually modeled as Poisson random variables, and for a Poisson random variable Y we have $V(Y) = E(Y)$. A reasonable way to combine the counts is to standardize each O value by subtracting the corresponding expected value, which we here take to be E, and dividing by the standard deviation which, if the observed value were Poisson would be the square root of the expectation, here $\sqrt{E}$. Each contribution $(O-E)^2/E$ may thus be considered the square of a standardized variable. It turns out that, for large samples, these standardized variables approximately follow a standard normal distribution. Recalling that the chi-squared distribution arises as a sum of squares of standard normal variables it then becomes at least plausible that a chi-squared distribution might be used to judge the magnitude of the chi-squared statistic. This argument may be made rigorous. We comment further on theoretical aspects of the method in Section 11.1.4.

To obtain the p-value for the chi-squared procedure we consider a random variable X having a χ^2_ν distribution and evaluate $p = P(X > \chi^2_{obs})$. This provides an approximate p-value (approximate because the chi-squared statistic approximately follows a chi-squared distribution, for large samples). We discuss the selection of ν in Section 10.1.2. If p is sufficiently small we consider the observed value to be rare. Typically, $p < .05$ is taken as modest evidence and $p < .01$ is taken as strong evidence that the model doesn't fit.

Example 5.1 (continued from p. 249) For the ADHD data we get

$$\chi^2_{obs} = \frac{(48-51.7)^2}{51.7} + \frac{(71-63.9)^2}{63.9} + \frac{(17-20.4)^2}{20.4} = 1.62.$$

We compare this to a χ^2_1 distribution by taking X to be a random variable having a χ^2_1 distribution and then computing $P(X > 1.62)$. We find $P(X > 1.62) = .20$, so that an approximate p-value is $p = .20$. This indicates a good fit of the Hardy-Weinberg model to these data. □

10.1.2 For multinomial data, the chi-squared statistic follows, approximately, a χ^2 distribution.

In Example 1.4 we introduced a binary random variable to analyze the variation across outcomes where each outcome was one of two possibilities, "burning house" or "non-burning house." In Example 5.1, we have a similar situation, except instead

of two possible outcomes we have three: each of the 136 subjects contributed a genotype that was classified as TT, TC, or CC. As discussed on p. 119, this leads to the assumption of a multinomial distribution across the three categories of data, which is the fundamental assumption for the application of the chi-squared test on p. 250. More generally, the theoretical starting point of every chi-squared test is the idea that the given set of counts may be considered an observation of a multinomial random vector. Here is a particularly straightforward example where the genetic model completely specifies the set of multinomial probabilities, leaving no free parameters.

Example 10.1 Allele frequencies in fruit flies Some basic genetic investigations have involved the "vestigial" (vg) and "ebony" (e) strains of fruit flies. The vestigial flies have small wings so that the animal can not fly, while the ebony flies are very dark in color. Kempthorne (1957, p. 131) cites an investigation involving cross breeding of vg with e flies. According to Mendelian equilibrium theory, the four possible results (denoted +, vg, e, vge) should be in the proportions 9:3:3:1. The four respective frequencies among 465 flies were 268, 94, 79, 24. The theoretical proportions are (.563, .188, .188, .0625) while the observed proportions were (.576, .202, .170, .0516). For instance, $.576 = 268/465$. In this case, we model the vector of numbers of phenotypes among 465 flies as a $M(n, p_1, p_2, p_3, p_4)$ distribution, where $n = 465$ and p_1 is the probability that a given fly would be of type +, p_2 the probability the fly would be of type vg, etc. We would assume that the phenotypes are independent of each other across flies (so that knowing one fly's phenotype does not change another fly's phenotype probability distribution), and each has the same set of four probabilities. Thus, under the model, the vector (268, 94, 79, 24) is treated as if it were an observed value of the multinomial random vector. □

In applications of chi-squared methodology each O is a count associated with a particular data *category*. In Example 5.1, for instance, the categories were TT, CT, CC. The number of categories is important in determining the degrees of freedom ν. The value to use for ν depends on the problem. If we take the number of categories to be k and the number of estimated parameters to be m then ν is found from the formula

$$\nu = k - 1 - m. \tag{10.2}$$

The degrees of freedom, often abbreviated *d.f.*, may be considered the number of free parameters. The idea and terminology of degrees of freedom come from mechanics: we count the number of dimensions in which the random variable is "free to move," often beginning with some apparent maximal number of dimensions and subtracting off constraints. The examples below should help clear this up, and there are general formulas for each type of problem. In Eq. (10.2) we begin with a multinomial distribution that has k categories with probabilities $p_1, \ldots, p_k$. Because these sum to 1, there are only $k - 1$ free parameters. Then, after estimating m parameters for the null hypothetical model we are left with $\nu = k - 1 - m$ free parameters.

Example 10.1 (continued from p. 251) Returning to the allele frequencies example, the "observed values" O are 268, 94, 79, 24. The "expected values" E values must be

calculated. If the ratios were 9:3:3:1, the corresponding proportions would be 9/16, 3/16, 3/16, 1/16. With 465 flies, we would therefore expect to see $\frac{9}{16} \cdot 465 = 261.6$, $\frac{3}{16} \cdot 465 = 87.2$, $\frac{3}{16} \cdot 465 = 87.2$, $\frac{1}{16} \cdot 465 = 29.1$. The O and E values are compared and summarized by the chi-squared statistic using (10.1):

$$\chi^2_{obs} = \frac{(268 - 261.6)^2}{261.6} + \cdots + \frac{(24 - 29.1)^2}{29.1} = 2.34.$$

Here there are four categories, so three degrees of freedom. □

Just as the binomial may be approximated by a normal distribution for large n, so too may the multinomial be approximated by a multivariate normal for large n. This leads to the general result that the chi-squared statistic follows, approximately, a chi-squared distribution.

Result Suppose $X \sim M(n, p_1, p_2, \ldots, p_k)$ and we have a statistical model $p_1 = p_1(\theta), p_2 = p_2(\theta), \ldots p_k = p_k(\theta)$ based on an m-dimensional parameter vector θ. Let $\hat{\theta}$ be the MLE and let Y_n be a random variable representing χ^2_{obs} according to (10.1), i.e.,

$$Y_n = \sum_{i=1}^{k} \frac{\left(X_i - np_i(\hat{\theta})\right)^2}{np_i(\hat{\theta})}. \tag{10.3}$$

Then, assuming suitable general conditions on the statistical model, as $n \to \infty$ we have

$$Y_n \xrightarrow{D} \chi^2_\nu \tag{10.4}$$

where $\nu = k - 1 - m$.

A detail: The "suitable general conditions" on the model are that the mapping $\theta \to (p_1(\theta), p_2(\theta), \ldots, p_k(\theta))$ must be one-to-one and differentiable with the derivative matrix having rank m. □

In practice, the most important input to this theoretical result, which leads to the calculation of the p-value, is the assumption that the data may be represented by a multinomial random vector. As in the binomial case, the multinomial assumption will make sense when it is reasonable to assume the classification variables are independent across observations (across subjects in Example 5.1). Thus, as before, it is the judgment of independence that must be considered most carefully.

10.1.3 The rarity of a large chi-squared is judged by its p-value.

The conventional cut-offs for the p-value are .05 and .01, with $p < .05$ and $p < .01$ reflecting modest and strong evidence. These two particular numbers were handed down from Fisher and are now imbedded in standard practice, but they are somewhat arbitrary and should be considered rough guides rather than finely tuned criteria.[2] Articles in the literature often include statements in the form $p < .05$, with the result typically being called *statistically significant*, or $p < .01$, which may be labeled *highly significant*. However, it is not unusual to obtain a very small p-value (e.g., 10^{-4}), which is quite different than .01. Rather than saying $p < .01$, it is preferable to report the p-value, and it is also good practice to say what statistic was computed, e.g., in Example 5.1 on p. 250, one would report $p = .20$ for chi-squared on one degree of freedom.

Example 10.1 (continued from p. 251) We use the computer to find $p = P(X > 2.34) = 1 - P(X \leq 2.43)$ where X has a χ^2_3 distribution. We obtain $P(X \leq 2.43) = .4951$ and therefore $p = .50$. This p-value is large, much larger than the conventional values .05 and .01. Thus, data that deviate from expected values as much as these would not be rare and we conclude there is a good fit of the theoretical model to these data. □

Example 5.4 (continued from p. 111) In the radioactive disintegration example, the statistical model is that the data are a sample from a $P(\lambda)$ distribution. Here, we have $\theta = \lambda$ so that $p_i(\theta) = p_i(\lambda)$. The O and E values are given in Table 10.1. The E values are obtained as $E_i = np_i(\hat{\lambda})$ where $p_i(\lambda) = P(X = i) = e^{-\lambda}\lambda^i/i!$ and we then substitute $\lambda = \hat{\lambda} = \bar{x}$. Thus, after computing $\hat{\lambda} = \bar{x} = 3.87$ we obtain the values $\hat{p}_i(\hat{\theta}) = e^{-\hat{\lambda}}\hat{\lambda}^i/i!$, which appear in the theoretical statement (10.3) and the values $E_i = np_i(\hat{\lambda})$, which appear without the subscript i in (10.1). For example, the expected number of times we would observe one particle emitted is 2608 times the probability of getting one particle emitted, i.e., $2{,}608 \cdot e^{-3.87}(3.87) = 210.523$.

Calculation of (10.1) gives $\chi^2_{obs} = 12.9$ and here there are $\nu = 11 - 1 - 1 = 9$ degrees of freedom: we start with $11 - 1 = 10$ degrees of freedom, because there are 11 categories, but we lose one degree of freedom from estimating λ. From the chi-squared cdf we find that when $X \sim \chi^2_{10}$, $P(X > 12.9) = .17$. Thus, $p = .17$ and there is no evidence of departure from the Poisson distribution despite the large sample size, which would have given an opportunity to detect even a small departure. □

> *A detail:* A technical point arises in Example 5.4, above, from the observation that the number of categories here is actually somewhat arbitrary: we chose to use 11 categories, but could have chosen a different number. As a result, the large-sample distribution is not the claimed chi-squared, but a slightly different approximation (a pair of

[2] Our characterization of $p < .05$ as "modest evidence" is consistent with Fisher's view. In particular, he felt $p = .05$ was inconclusive. See the footnote on p. 298.

Table 10.1 Fit of Poisson distribution to the counts of α-particle emissions from a specimen during 2,608 intervals.

k	Observed counts	Poisson fitted counts
0	57	54.399
1	203	210.523
2	383	407.361
3	525	525.496
4	532	508.418
5	408	393.515
6	273	253.817
7	139	140.325
8	45	67.882
9	27	29.189
≥ 10	16	17.075

bounds) may be used for the p-value. In this case, using 11 categories, the p-value would be somewhere between those obtained for 9 and 10 degrees of freedom. This would make the p-value a bit bigger than our reported $p = .17$. Many texts emphasize this technicality but, for models such as these, with a single parameter, it has little effect on the conclusions. □

10.1.4 Chi-squared may be used to test independence of two traits.

Many studies seek to evaluate the association of two traits. In genetic epidemiology, for instance, it is useful to know whether a particular genotype may be associated with a disease. When the occurrence of each trait is considered a random variable, the traits will fail to be associated if the two random variables are independent. Thus, the issue becomes one of evaluating the fit of a statistical model based on independence.

Example 10.2 Alzheimer's and APOE As part of a study of markers for late-onset Alzheimer's disease, Yu et al. (2007) looked for the presence of the ε_4 allele of the apolipoprotein E gene (*APOE*), which had previously been associated with increased risk of Alzheimer's, among both Alzheimer's patients and controls. The following table summarizes some of the data they presented from 193 Alzheimer's patients (AD) and 232 controls:

	ε_4 absent	ε_4 present
AD	58	135
Controls	162	70

At first glance it appears that the ε_4 allele is far more prevalent among the Alzheimer's patients than among the controls—and that this is probably not due to chance. This may be verified using a χ^2 test. □

Example 10.2 involves what is called a *two-by-two table* (written 2×2). In general, the probabilities for a 2×2 table may be represented as follows:

	1 absent	1 present	
2 absent	p_{11}	p_{12}	p_{1+}
2 present	p_{21}	p_{22}	p_{2+}
	p_{+1}	p_{+2}	

Here the subscript ij corresponds to the (i, j) element in the table, meaning that p_{ij} is the probability in row i and column j. For example, p_{22} is the probability that a random individual has both trait 1 and trait 2 (e.g., in Example 10.2 both ε_4 and AD). The probabilities along the margins of the table come from summing the probabilities along rows or columns. For example, $p_{+2} = p_{12} + p_{22}$ is the probability that the individual has trait 1 (e.g., ε_4) and $p_{2+} = p_{21} + p_{22}$ is the probability that the individual has trait 2 (e.g., AD). Now, if independence holds, then the probability of having both trait 1 and trait 2 must equal the probability of having trait 1 times the probability of having trait 2, i.e., $p_{22} = p_{2+}p_{+2}$. Filling out the rest of the table of probabilities the same way gives the independence model

$$p_{ij} = p_{i+}p_{+j}$$

for all i, j.

In order to apply χ^2_{obs} we need to compute the expected values, each of which is the number of individuals we would expect in a particular entry of the table. In principle, the expected value for the (i, j) entry in the table is $E = n \cdot p_{ij} = n \cdot p_{i+}p_{+j}$ for each of the four p_{ij}'s, but we don't know the values of p_{i+} and p_{+j}. Here we resort to the standard "plug-in" method: we estimate these marginal probabilities from the data. For instance, in the Alzheimer's example there are a total of 425 individuals so we use $\hat{p}_{1+} = (58 + 135)/425$, for the probability of having AD, etc. ($\hat{p}_{2+} = (162 + 70)/425$, $\hat{p}_{+1} = (58 + 162)/425$, $\hat{p}_{+2} = (135 + 70)/425$).

This estimation process causes the chi-squared distribution to lose degrees of freedom, as in Example 5.4. In general, if there are r rows and c columns we begin with $rc - 1$ degrees of freedom: there are rc probabilities in the table but they must sum to 1, which means we lose one degree of freedom. We then lose another $r - 1$ degrees of freedom for estimating row marginal probabilities and $c - 1$ for estimating column marginal probabilities. This leaves $rc - 1 - (r - 1) - (c - 1) = rc - r - c + 1 = (r - 1)(c - 1)$ degrees of freedom.

Example: 10.2 (continued from p. 254) In this example $r = 2$ and $c = 2$ so there is one degree of freedom. Entering the data into an appropriate statistical software package produces $\chi^2_{obs} = 65$ on one degree of freedom, and $p = 7 \times 10^{-16}$, which is truly tiny. Clearly there is an association here. □

Software used to get chi-squared results, as in Example 10.2 above, typically applies a variation of the chi-squared statistic that includes a "continuity correction." This adjusts the statistic slightly to make the continuous chi-squared distribution match more closely the distribution of the discrete chi-squared statistic in small samples. It is also possible to use so-called "exact" methods, which avoid the χ^2 distribution altogether. While such methods are commonly applied, it is important to keep in mind that we are usually looking for clear and compelling results, either not significant or strongly significant, and borderline cases should be interpreted as such. That is, when the continuity correction—or the distinction between exact and approximate methods—is important to conclusions, this signals a case in which a careful investigator ought to recognize the ambiguity of the data.

Example 10.2 (continued, introduced on p. 254) The Alzheimer's and *APOE* data may be examined further to see if there is a difference between men and women. Here is the table for the AD patients:

	ε_4 absent	ε_4 present
Women AD	32	70
Men AD	26	65

The proportions appear to be about the same, and this time we get $\chi^2 = .071$ again on one degree of freedom, and $p = .79$, so there is no evidence of any discrepancy in ε_4 prevalence among the male and female AD patients. □

One final subtlety should be noted. The logic we have described here assumes that all subjects have the same underlying (theoretical) probabilities p_{ij}, as would occur if each subject in the study were drawn randomly from a population of potential subjects. That could be a good rough description of what happened in the Alzheimer's study. However, often a set of diseased patients is selected and then a set of controls is chosen separately. In epidemiology this is called a *case-control* study. It generates a different statistical model, but it turns out to give the same χ^2 test. (The cited study did not say which way the subjects were collected.) We return to the issue of data collection strategies and their effects on scientific inference in Section 13.4.

10.2 Null Hypotheses

10.2.1 Statistical models are often considered null hypotheses.

In talking about assessing fit we have used a "hypothesized model," i.e., the model being fit to the data. The standard terminology is to take such a model to be the "null" model, or the *null hypothesis*, often written as H_0. Sometimes the null hypothesis completely specifies the probability distribution, as in Example 10.1 (p. 251). In other cases it merely identifies a family of distributions, as in the α-particle emissions example (where there is still a free parameter λ), and in the Alzheimer's and *APOE*

example (where there remain two free parameters p_{1+} and p_{+1}). The "null" here indicates that such a hypothesis is often used with an eye toward collecting evidence against the hypothesis, the implicit understanding being that H_0 would eventually be replaced with something that could describe such data better.

10.2.2 Null hypotheses sometimes specify a particular value of a parameter within a statistical model.

Another possibility is that the null hypothesis specifies a particular value of a parameter within a family of distributions.

Example 1.4 (continued, see p. 13) In the investigation of blindsight in patient P.S. the possibility that P.S. was guessing corresponds to taking $p = .5$ in the binomial model. We write this as $X \sim B(17, p)$ with H_0: $p = .5$. One way to test this is with χ^2. We take the observed values to be 14 and 3 (for the two categories "non-burning preferred" and "burning preferred") and take the expected values to be np_0 and $n(1-p_0)$, with $n = 17$ and $p_0 = .5$, which gives $np_0 = 8.5$ and $n(1-p_0) = 8.5$. The chi-squared statistic is then

$$\chi^2_{obs} = \frac{(14 - 8.5)^2}{8.5} + \frac{(3 - 8.5)^2}{8.5} = 7.12.$$

Here we have two categories and 0 estimated parameters, so $\nu = 1$. Comparing 7.12 to a χ^2_1 distribution gives a p-value of $p = .0076$, which[3] is strong evidence against H_0. □

In Example 1.4 there is a simple null hypothesis and a chi-squared procedure to test it. Because the sample size there is small, however, the continuity correction mentioned on p. 256 would change the p-value somewhat. We will obtain a more accurate p-value for Example 1.4 on p. 267.

10.2.3 Null hypotheses may also specify a constraint on two or more parameters.

In the blindsight example (p. 257) we had a single binomial and tested $H_0 : p = .5$. Now suppose we have two binomials, $X_1 \sim B(n_1, p_1)$ and $X_2 \sim B(n_2, p_2)$ and we wish to test $H_0 : p_1 = p_2$. This is a special case of a widely-applied type of null hypothesis, namely one that corresponds to a constraint on some parameters in a

[3] In this example we use the notation p in two different ways: at first p stands for the probability that P.S. would choose the non-burning house, and then later it stands for the p-value. These are both such common notations that we felt we couldn't change either of them. We hope our double use of p is not confusing.

statistical model. In the case of two binomials, $H_0 : p_1 = p_2$ may be assessed by comparing χ^2_{obs} to a χ^2_1 distribution: we begin with two free parameters p_1 and p_2 and lose a degree of freedom due to the constraint. In fact, this special case of χ^2_{obs} turns out to be mathematically equivalent to the test of independence examined above.

Example 10.2 (continued, see p. 254) On p. 256 the way the Alzheimer's data were collected would affect the way the statistical problem would be posed. If AD patients and controls were collected separately, then we would examine whether the probability of having the ε_4 genotype was the same in each population, i.e, we would have two binomials and would test $H_0 : p_1 = p_2$. To repeat, this test may be carried out using χ^2_{obs}, exactly as done previously, on p. 255. □

In a similar way, data from two independent samples $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$ may be used to test the hypothesis that the corresponding means μ_1 and μ_2 are equal, $H_0 : \mu_1 = \mu_2$.

Example 1.1 (continued from p. 3) In the case of the SEF neuronal activity under two conditions there were 15 trials in both experimental conditions, generating mean firing rates of 48 spikes per second for the spatial condition and 70 spikes per second for the pattern condition across the time interval from 200 to 600 milliseconds after the onset of the cue. The null hypothesis $H_0 : \mu_1 = \mu_2$ would say that the two mean firing rates are equal. □

The standard statistical procedure for testing $H_0 : \mu_1 = \mu_2$ is called a t-test, because it relies on the t distribution. We discuss this in Section 10.3.4. Example 7.2 provides another example.

Example 7.2 (continued from p. 167) For the test-enhanced learning study we previously showed how to get a confidence interval for $\mu_1 - \mu_2$, where μ_1 and μ_2 were the mean scores within the SSSS and SSST conditions. As an alternative we may test the null hypothesis $H_0 : \mu_1 = \mu_2$, which says that the theoretical mean scores in the SSSS and SSST conditions are identical. We present results based on the t-test on p. 265. □

10.3 Testing Null Hypotheses

10.3.1 The hypothesis $H_0 : \mu = \mu_0$ for a normal random variable is a paradigm case.

We have already noted that a null hypotheses may specify a particular value of a parameter. To establish intuition based on a widely-used form of test statistic, let us return to the prototypical situation we considered in Section 7.3.2, where we have a sample $X_1, \ldots, X_n$ from a $N(\mu, \sigma^2)$ distribution with σ known. To test $H_0 : \mu = \mu_0$ we may form the ratio

$$Z = \frac{\bar{X} - \mu_0}{SE(\bar{X})} \tag{10.5}$$

where

$$SE(\bar{X}) = \sigma/\sqrt{n} \tag{10.6}$$

is the standard error of the mean, as in Eq. (7.9). The data-based analogue, computed from a sample $x_1, \ldots, x_n$, is

$$z_{obs} = \frac{\bar{x} - \mu_0}{SE(\bar{x})} \tag{10.7}$$

where $\bar{x}$ is the sample mean computed from the data and $SE(\bar{x}) = \sigma/\sqrt{n}$. (The *SE* value is the same for the data-based mean and its theoretical counterpart because the formula in this simple case does not depend on the actual values of the data.) If the magnitude $|z_{obs}|$ is sufficiently large we would say there is evidence against H_0. To analyze this procedure we return to the theoretical statement (10.5). Because $\bar{X} \sim N(\mu, \sigma^2/n)$, under $H_0 \colon \mu = \mu_0$ we also have

$$Z \sim N(0, 1). \tag{10.8}$$

We therefore obtain a *p*-value from

$$p = P(|Z| \geq |z_{obs}|). \tag{10.9}$$

Together, (10.7) and (10.9) define a *z-test* for normal data with σ known.

As in Section 7.3.2 we have presented the *z*-test first in this special case for conceptual simplicity. In practice, the data are typically not normally distributed and σ is not known. We may treat the more general setting by approximation, analogously to what was done in Section 7.3.4. The procedure is to replace σ with the sample standard deviation s in $SE(\bar{x})$, as in Eq. (7.17) and, having done so, invoke (10.7) as above. For the purpose of formalizing the argument in theoretical terms let us replace Z, in (10.5) with Y,

$$Y = \frac{\bar{X} - \mu_0}{SE(\bar{X})}. \tag{10.10}$$

We do this because when the observations are non-normal Y will also typically be non-normal and we want to reserve the notation Z for the case $Z \sim N(0, 1)$.

Result If $X_1, \ldots, X_n$ is a random sample from a distribution having mean μ and standard deviation σ, and n is sufficiently large, then a test of the null hypothesis $H_0: \mu = \mu_0$ may be carried out by applying (10.7) with $SE(\bar{x})$ defined by (7.17) and computing an approximate p-value using (10.9). That is, under $H_0: \mu = \mu_0$, for sufficiently large n we have

$$P(|Y| \geq |z_{obs}|) \approx P(|Z| \geq |z_{obs}|) \tag{10.11}$$

where Y is defined by (10.10) and $Z \sim N(0, 1)$, so that the p-value based on (10.7), where $SE(\bar{x})$ is defined by (7.17), together with (10.9) is approximately correct.

This result is an immediate consequence of the theorem following (7.18).

10.3.2 For large samples the hypothesis H_0: $\theta = \theta_0$ may be tested using the ratio $(\hat{\theta} - \theta_0)/SE(\hat{\theta})$.

The uncertainty associated with an estimate is quantified by the estimate's standard error, as defined in Eq. (7.6) on p. 159. In Example 1.4, concerning blindsight in patient P.S. we reported on p. 13 an approximate 95 % confidence interval (.64, 1.0) (based on calculations given on p. 158) and we noted that this was inconsistent with the probability of .5, which would correspond to guessing. But if we are mainly interested in whether the data are consistent with guessing, we could rephrase the problem using the observed discrepancy between $\frac{14}{17}$ and .5. The proportion $\hat{\theta} = \frac{14}{17}$ seems much too big to be consistent with guessing. So we may ask this question: If P.S. were guessing, how unlikely would it be that $\hat{\theta}$ would be as far from .5 as was $\frac{14}{17}$?

We will present several different procedures that provide slightly different numerical answers to this question, all of which lead to the same conclusion. The one most closely related to the approximate confidence interval in (7.8) assesses the discrepancy between $\hat{\theta}$ and .5 in units of $SE(\hat{\theta})$. This relies on the approximate normality of the MLE $\hat{\theta}$.

Result: Suppose $X_1, \ldots, X_n$ has joint pdf $f(x_1, \ldots x_n|\theta)$, with θ a scalar, and suppose further that T_n is an asymptotically normal estimator of θ with standard error $SE(T_n) = \hat{\sigma}_{T_n}$. Then the null hypothesis H_0: $\theta = \theta_0$ may be tested by using the statistic

$$z_{obs} = \frac{T_n - \theta_0}{SE(T_n)}, \tag{10.12}$$

with large values of $|z_{obs}|$ indicating evidence against H_0. If the sample size is large, an approximate p-value may be obtained from

$$p = P(|Z| \geq |z_{obs}|) \tag{10.13}$$

where $Z \sim N(0, 1)$.

This result follows from the theorem in Section 7.3.5, which said that if $\hat{\sigma}_{T_n}$ is the standard error of T_n in the sense that

$$\frac{\hat{\sigma}_{T_n}}{\sigma_{T_n}} \xrightarrow{P} 1$$

then

$$\frac{(T_n - \theta)}{\hat{\sigma}_{T_n}} \xrightarrow{D} N(0, 1).$$

If $\theta = \theta_0$ then the random variable

$$Z = \frac{T_n - \theta_0}{SE(T_n)}$$

follows, approximately, for large n, a $N(0, 1)$ distribution and the p-value based on $Z \sim N(0, 1)$ will be approximately correct. Because Z is a common notation for a $N(0, 1)$ random variable, the value z_{obs} in (10.12) is often called a *z-score* and the procedure in (10.12) and (10.13) is a *z-test*.

Example 1.4 (continued from p. 257) Suppose $X \sim B(n, \theta)$ and we wish to test $H_0 : \theta = \theta_0$. The usual formula for SE is $SE(\hat{\theta}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$. It is customary to find SE under the null hypothesis, $\theta_0 = .5$, i.e., we replace[4] $\hat{\theta}$ with $\theta_0 = .5$ in the calculation of SE. In the case of the data from P.S. we had $n = 17$ so we get $SE = \sqrt{(.5)(.5)/17} = .121$, and $z_{obs} = (.824 - .5)/.121 = 2.68$. This gives us a

[4] The logic of the procedure does not demand that we use θ_0 in place of $\hat{\theta}$. The justification of the large-sample significance test, the Theorem in Section 7.3.5 that says Z is approximately $N(0, 1)$, is not refined enough to distinguish between the two alternative choices for $SE(T_n)$ (both would satisfy the theorem). However, because we are doing the calculation under the assumption that $\theta = \theta_0$, it makes some sense to use the value $\theta = \theta_0$ in computing the standard error.

p-value of .0074, which is nearly the same as the value .0076 obtained from the chi-squared analysis (see p. 257). In fact, in this case, a little bit of manipulation shows that we have the arithmetic identity $z_{obs}^2 = \chi_{obs}^2$, where z_{obs} is defined in (10.12) and χ_{obs}^2 is defined by (10.1) with (10.2). □

The identity above provides a way of understanding the chi-squared procedure. The definition of a χ_1^2 distribution is that it results from squaring a $N(0, 1)$ random variable. When we replace the data with random variables we get the theoretical counterpart of the observed value z_{obs},

$$Z = \frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})},$$

which has an approximate $N(0, 1)$ distribution. Therefore, its square has an approximate χ_1^2 distribution, but its square is the theoretical counterpart of the observed value $z_{obs}^2 = \chi_{obs}^2$. In other words, the theoretical chi-squared statistic follows, approximately, a chi-squared distribution.

When θ is a vector essentially the same result as in (10.12) and (10.13) holds again for each component. That is, if θ_i is one component of θ and $T_{n,i}$ is the corresponding component of an asymptotically normal vector estimator T_n (which would be asymptotically multivariate normal as in (8.42)), then we can test $H_0 : \theta_i = \theta_{i,0}$ by replacing T_n by $T_{n,i}$ and θ_i by $\theta_{i,0}$ in (10.12) and again using (10.13). For example, in simple linear regression we may have both an intercept and a slope, but we may wish to test the null hypothesis that the slope is zero—which would correspond to there being no linear relationship between the response and explanatory variables. We return to this case in Chapter 12.

10.3.3 For small samples it is customary to test $H_0 : \mu = \mu_0$ using a t statistic.

In Section 7.3.10 we presented the usual t-based confidence interval for a mean μ of a normal distribution. The point was that, for small samples of observations that are truly normal, the normal distribution of the standardized sample mean should be replaced by a t distribution (with degrees of freedom given by the degrees of freedom used in the estimation of σ by s). In the case of testing $H_0 : \mu = \mu_0$ with truly normal observations the normal distribution in (10.9) is replaced by a t-based counterpart:

$$p = P(|T| \geq |t_{obs}|) \tag{10.14}$$

where t_{obs} is defined by replacing σ with s in (10.6) and (10.7), i.e.,

$$t_{obs} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \tag{10.15}$$

and T follows a t distribution, $T \sim t_\nu$ where $\nu = n - 1$. This is called a *t-test*. We could consider n to be "large" and apply (10.9) with $\hat{\theta} = \bar{x}$ and $SE(\bar{x}) = s/\sqrt{n}$. As in Section 7.3.10, using the t distribution instead of the standard normal distribution has the effect of making extreme values more probable; therefore, the p-value using the t distribution with (10.14) will be larger (providing less evidence against H_0) than that found using the normal distribution (10.9), and the distinction vanishes as n increases.

The t-test defined in Eq. (10.14) is often used when paired data of the form u_i and w_i are observed and their differences $x_i = u_i - w_i$ are analyzed. The conception is that $U_1, \ldots, U_n$ is a random sample from a $N(\mu_1, \sigma_1^2)$ distribution and $W_1, \ldots, W_n$ is a random sample from a $N(\mu_1, \sigma_1^2)$ distribution and the problem is to test $H_0 : \mu_1 = \mu_2$. The differences $X_i = U_i - W_i$, for $i = 1, \ldots, n$ then form a random sample from a $N(\mu, \sigma^2)$ distribution with $\mu = \mu_1 - \mu_2$. The null hypothesis then may be rewritten $H_0 : \mu = 0$, so that we obtain a normal random sample with null hypothesis of the form $H_0 : \mu = \mu_0$ (where $\mu_0 = 0$), which is the problem solved by the t-test in Eq. (10.14). In this setting the procedure is called a *paired t-test*.

Example 10.3 Glutamate increase in response to pain Mullins et al. (2005) used proton magnetic resonance spectroscopy to study brain response to pain in humans. The authors obtained spectra from the anterior cingulate cortex during application of painfully cold compress to the subject's foot and during several rest periods. One analysis used the magnitude of the response associated with glutamate. This involved a pair of measurements of the form u_i and w_i, for subject i, with u_i being the glutamate concentration during pain and w_i being the glutamate concentration during rest. The differences $x_i = u_i - w_i$, for $i = 1, \ldots, n$ were then analyzed with a paired t-test. In this study, which the authors called "preliminary," results from only seven subjects were reported. The authors reported a 9.3 % increase in glutamate concentration during pain, with $t_{obs} = 3.85$, yielding $p = .006$, which is highly significant. In other words, even with only seven subjects, these data appear to provide strong evidence of an increase in glutamate in anterior cingulate cortex during administration of a painful stimulus. □

The t-test is justified by the following theorem.

Theorem If $X_1, \ldots, X_n$ is a sample from a $N(\mu, \sigma^2)$ distribution and $H_0 : \mu = \mu_0$ holds, then

$$P(|Y| \geq |t_{obs}|) = P(|T| \geq |t_{obs}|) \tag{10.16}$$

where Y is defined by (10.10) with $SE(\bar{X}) = S/\sqrt{n}$, $t_{obs} = z_{obs}$ is given by (10.7) with $SE(\bar{x})$ defined by (7.17), and T follows a t_ν distribution with $\nu = n - 1$.

Proof: The proof is the same as that of the theorem containing Eq. (7.31). □

In practice, as we said in Section 7.3.10 calculations based on t distributions often agree pretty well with those based on normal distributions. However, for large values of $|t_{obs}|$ the tails of the distribution come into play, and the p-values computed with the t distribution may be quite a bit different than those based on the normal

distribution. In any case, throughout the scientific literature the t-test is considered a standard approach, as long as the data do not deviate too far from normality. The small sample size in Example 10.3 is worrisome because departures from normality could affect the results. The p-value of .006, however, is sufficiently small to be reassuring: substantial departures from normality would be required to change the conclusion we would draw from the data. In Section 13.3 we discuss methods that depend on neither the normality of the data, as in (10.16), nor normality of the sample mean, as in (10.11).

10.3.4 For two independent samples, the hypothesis H_0: $\mu_1 = \mu_2$ may be tested using the t-ratio.

Let us next apply the idea in Section 10.3.2 to the problem of testing $H_0 : \mu_1 = \mu_2$ based on two independent samples $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$. The obvious starting point is the difference between the sample means $\bar{X}_1 - \bar{X}_2$, which should then be divided by its standard error.

Now, what is the standard error of $\bar{X}_1 - \bar{X}_2$? Because the two samples are independent we have

$$V(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \tag{10.17}$$

where σ_1^2 and σ_2^2 are the respective variances of each X_{1i} and X_{2i}, within each of the two samples. The standard error will be the square-root of the variance in (10.17) after we plug in suitable estimates of σ_1 and σ_2 (as in Eq. (7.24)). The most common procedure, the ordinary *t-test,* makes the assumption that $\sigma_1 = \sigma_2$, which greatly simplifies the theoretical results. We now label these standard deviations by σ (so that $\sigma = \sigma_1 = \sigma_2$). With this assumption, the two sample standard deviations s_1 and s_2 both estimate σ. We then pool the data together by calculating

$$S^2_{pooled} = \frac{1}{n_1 + n_2 - 2}\left(\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2 + \sum_{i=1}^{n_2}(X_{2i} - \bar{X}_2)^2\right)$$

which is taken as an estimator of σ^2 and gets plugged into (10.17) for σ_1 and σ_2. The test statistic becomes

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{10.18}$$

and, assuming $\mu_1 = \mu_2$, as n_1 and n_2 become infinite T converges in distribution to $N(0, 1)$. This gives the following method (where the notation converts the capital T, X and S to lower case once T is applied to observed data).

Result: Suppose $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$ are independent random samples from distributions having means μ_1 and μ_2 and standard deviations $\sigma_1 = \sigma_2$. The null hypothesis $H_0 : \mu_1 = \mu_2$ may be tested using

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{10.19}$$

with large values of $|t_{obs}|$ indicating evidence against H_0. If the sample sizes are large, an approximate p-value may be obtained from

$$p = P(|Z| \geq |t_{obs}|) \tag{10.20}$$

where $Z \sim N(0, 1)$.

The result above, using (10.20), is justified by the Central Limit Theorem. If, in addition, we are willing to assume normality of the distributions then we have a theoretically exact result, which applies in small samples.

Result: Suppose $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$ are independent random samples from normal distributions having means μ_1 and μ_2 and standard deviations $\sigma_1 = \sigma_2$. The null hypothesis $H_0 : \mu_1 = \mu_2$ may be tested using (10.19) with large values of $|t_{obs}|$ indicating evidence against H_0. A p-value may be obtained from

$$p = P(|T| \geq |t_{obs}|) \tag{10.21}$$

where $T \sim t_\nu$, with $\nu = n_1 + n_2 - 2$.

The method above, using (10.21) with (10.19), is called the *two-sample t-test*. Sometimes the two samples are called "independent" to emphasize the distinction between this setting and that of the paired t-test in Section 10.3.3. To be concrete, suppose that the data come from human subjects. Typically, the data in the paired case are paired because two observations come from the same subject, as in Example 10.3. It is then natural to take advantage of the pairing by analyzing differences. In contrast, the two samples in (10.19) come from separate subjects[5] and there is no natural way to identify a particular x_1 observation with an x_2 observation. Here is an example.

Example 7.2 (continued from p. 258) In the test-enhanced learning study Roediger and Karpicke (2006) found strong evidence against H_0, the hypothesis the theoretical mean scores in the learning-test group and the restudy groups were identical. Applying the two-sample t-test to the data displayed in Fig. 7.3 we obtained $t_{obs} = -3.19$ on 58 degrees of freedom. Using the normal approximation this gives

[5] We discuss this distinction again in Section 13.1.

$p = .0014$ while using the t distribution we get $p = .0023$. Either way there is strong evidence against H_0, indicating strong evidence that the mean assessment score under the SSST condition is greater than the mean assessment score under the SSSS condition. □

In Example 7.2 the p-value is larger when the t distribution is used than when the normal distribution is used. This is generally the case, as the t distribution has thicker tails, so that it gives higher probability to values with large magnitudes. Standard practice is to report the t-based p-value.

Deviations from the assumption that $\sigma_1 = \sigma_2$, which motivates the use of (10.18), typically must be quite large in order to have a strong effect on the p-value in (10.20) or (10.21). (A rough rule of thumb would be that, for substantial sample sizes, the conclusions are likely to be valid when the standard deviations are within a factor of 3 of each other.) However, a simple alternative is to define $S_1 = s_1$ and $S_2 = s_2$ to be the sample standard deviations of the two respective samples and then define

$$t_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}. \tag{10.22}$$

Replacing T in (10.18) with

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}, \tag{10.23}$$

the large-sample result based on the central limit theorem again holds, with p-value given by (10.20). This version of the two-sample t-test is often called[6] *Welch's* t-test, or the *unequal variance* t-test. We provide simulation-based methods of computing the p-value for this test in Sections 11.2.1 and 11.2.2.

10.3.5 Computer simulation may be used to find p-values.

We have gone over several examples of p-values. Let us now summarize the essential logic we have applied, and show how they may be obtained using computer simulation.

In each case we have an observed value of some test statistic, which we now write in generic form as q_{obs}. The examples so far have involved various formulas for χ^2_{obs}, z_{obs} and t_{obs}, with context determining the formula. We then introduce a theoretical

[6] Welch provided an approximate distribution from which p-values could be computed, which is more accurate than the normal.

statistic Q, and use its distribution under the null hypothesis (chi-squared, normal, t_ν) in a relevant statistical model to compute the p-value

$$p = P(Q \geq q_{obs}|H_0) \tag{10.24}$$

where we have used the conditioning notation to emphasize that[7] the probability is computed under the assumption that H_0 holds.

In many situations it is possible to use the computer to generate artificial data under the null hypothesis. That is, the statistical model specified by the null hypothesis contains certain probability distributions, and it is often relatively easy to generate observations from these probability distributions. We have done this previously, in Section 9.1.1, and produced *simulated* data, which we have also called *pseudo-data*. Each set of pseudo-data should resemble the real data in many respects that are crucial to analysis, such as having the same number of observations as the real data. On the other hand, the pseudo-data will have known variation with all the characteristics we assume in our theoretical world of statistical modeling. If we can create sets of pseudo-data repeatedly, a large number of times (each set of pseudo-data being different due to the randomness specified by the statistical model) then we can also compute the p value numerically.

The idea is to generate a large number G of pseudo-data sets (e.g., $G = 10{,}000$) and apply the statistic Q to each set of pseudo-data. This produces G computer-generated observations from the probability distribution of Q (under H_0). To find $p = P(Q \geq q_{obs})$ we then simply have to get the proportion of such generated observations (out of $G = 10{,}000$) for which Q is as large as q_{obs}. Let us use $Q^{(g)}$ to denote a value of Q computed from a set of pseudo-data, where $g = 1, 2, \ldots, G$. Here is the algorithm.

Finding the p-value by simulation

1. Generate G sets of pseudo-data labelled $g = 1, \ldots, G$ and for the gth set of pseudo-data compute $Q^{(g)}$.

2. Let N be the number of sets of pseudo-data for which $Q^{(g)} \geq q_{obs}$.

3. The p-value is given by $p = \frac{N}{G}$.

Example 1.4 (continued from p. 261) Let us take X to be a random variable representing the number of non-burning house preferences. Under the null hypothesis we have $X \sim B(17, .5)$. As our test statistic we may use $Q = |X - 8.5|$, where 8.5

[7] This may be considered an abuse of the notation because we usually consider H_0 to be a fixed, non-random entity, so we are not really "conditioning" on it in the usual sense developed in Chapter 3. The exception occurs under the Bayesian interpretation given in Section 10.4.5, where H_0 is formally considered to be an event. In that scenario the probability in (10.24) does become a conditional probability.

is the expected value of X and we are here judging small and large deviations from 8.5 to be equally important. We have $q_{obs} = 14 - 8.5 = 5.5$. We may then simulate 10,000 observations from a $B(17, .5)$ distribution and count the number N for which $Q \geq q_{obs}$. Doing this, we obtained $N = 126$ and $p \approx .013$. □

One issue is that the accuracy of such computer-generated p-values depends on the number of data sets generated. If we take G to be extremely large we can get a very accurate p-value, but in complicated problems the computing time may get too long. In most problems $G = 10{,}000$ is large enough to obtain reasonable accuracy.

> *Details:* In fact, we may compute the accuracy of such computer-generated p-values quite generally from the binomial standard error. If we generate G data sets, we have $N \sim B(G, p)$ where p is the desired p-value, which is estimated by $\hat{p} = N/G$. The standard error for this binomial proportion is $SE(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/G}$. Thus, in the example above, the accuracy would be $SE = \sqrt{(.0126)(.9874)/10{,}000} = .0011$. Doubling this we get a 95 % CI for p of $.013 \pm .002$. □

We used Example 1.4 to demonstrate the idea of simulation-based computation of p-values. The great virtue of p-values based on pseudo-data is that they can be easy to compute even in very complicated situations where direct calculation is impossible. However, the binomial setting shares with some other common problems sufficient simplicity that the exact p-value may be computed more directly.

Example 1.4 (continued) We have that $Q \geq q_{obs}$ precisely when $x \geq 14$ or $x \leq 3$. Thus, we have

$$p = P(X \leq 3) + P(X \geq 14)$$

where $X \sim B(17, .5)$, which may be computed by evaluating the binomial cdf from statistical software. Specifically, if $F(x)$ is the $B(17, .5)$ cdf, then

$$p = F(3) + 1 - F(13).$$

In this special case the $B(17, .5)$ distribution is symmetrical so that $P(X \geq 14) = P(X \leq 3)$ and we also have

$$p = 2F(3) = .013$$

which agrees with the value obtained above, by simulation. □

10.3.6 The Rayleigh test can provide evidence against a uniform distribution of angles.

When a random sample $X_1, \ldots, X_n$ consists of angles, i.e., measurements between 0 and 360 degrees (with 0 being the same as 360) or, equivalently, 0 and 2π radians (with 0 being the same as 2π) a common question is whether the angles tend to be

clustered around a particular direction. A natural null hypothesis is that the angles are uniformly distributed on the unit circle, i.e., $X_i \sim U(0, 2\pi)$ with the understanding that $x = 0$ is the same as $x = 2\pi$.

There is a body of methods devoted to analyzing data on the unit circle, i.e., angles, which are usually called *circular data*. For data $x_1, \ldots, x_n$ let us define

$$\begin{aligned} \bar{C} &= \frac{1}{n}\sum_{i=1}^{n}\cos(x_i) \\ \bar{S} &= \frac{1}{n}\sum_{i=1}^{n}\sin(x_i) \\ R_{obs} &= \sqrt{\bar{C}^2 + \bar{S}^2}. \end{aligned} \tag{10.25}$$

Recall (see p. 610 of the Appendix) that the cosine and sine of an angle α are the (x, y) coordinates of a point found by rotating the vector (1, 0) counter-clockwise through an angle α. This implies that $\bar{C}$ and $\bar{S}$ are, respectively, the mean of the x coordinate and the mean of the y coordinate when the data are plotted as points on the unit circle. The vector $(\bar{C}, \bar{S})$ is called the sample mean *resultant vector* and R is its magnitude. Note that $(n\bar{C}, n\bar{S})$ is the sum of n unit vectors so[8] its maximal length is n, which occurs when all the vectors (x_i, y_i) are equal. In this case we get $R = 1$. When the vectors tend to be clustered together, R gets close to 1. The *Rayleigh test* uses R as a test statistic and computes

$$p = P(R > R_{obs}),$$

where R is the random variable defined as in (10.25) with random variables $X_1, \ldots, X_n$ replacing data values $x_1, \ldots, x_n$, under the assumption that $X_1, \ldots, X_n$ form a sample from the uniform distribution on the unit circle.

Example 10.4 Hippocampal hemispheric differences among homing pigeons Gagliardo et al. (2001) examined directional orienting after release among groups of homing pigeons in three experimental conditions. In the first condition, at one month of age each pigeon was subjected to left unilateral ablation of the hippocampal formation. In the second condition, at 1 month of age each pigeon was subjected to right unilateral ablation of the hippocampal formation. The third condition was a control, with no ablation. At around four months of age the birds were released from one of three locations and their direction of flight was recorded. This generated samples with sizes ranging from $n = 11$ to $n = 30$ across the nine groups (Three locations for each of three treatments.) Each sample was a set of flight direction angles and the initial question was whether the birds tended to follow a particular direction (homeward). In this case the null hypothesis was no orientation

[8] This generalizes (and follows from) Eq. (A.29), which says that the maximal length of the sum of two unit vectors is 2 and it occurs when the vectors are equal.

at all, i.e., a uniform distribution of angles. The authors applied the Rayleigh test and found strong evidence against H_0 for the control and right-side ablation groups ($p < .001$ for four groups and $p < .02$ for two groups) but no evidence against H_0 for the left-side ablation group. This, together with other analyses, led them to conclude that the left hippocampal formation appears to be critical for navigational map learning. □

Statistical software is available for computing this p-value, but it is easy to compute using simulation as in Section 10.3.5.

10.3.7 The fit of a continuous distribution may be assessed with the Kolmogorov-Smirnov test.

The framework for statistical hypothesis testing, which includes specifying a null hypothesis, choosing a statistic for evaluating it, and then computing a p-value, is very flexible. Here is another case which, like the chi-squared tests of Section 10.1, may be used to assess fit of a statistical model.

Suppose we have a sample of i.i.d. random variables $X_1, \ldots, X_n$ each having distribution function $F(x)$, and we wish to examine whether $F(x)$ takes a specified form, such as $N(0, 1)$ or $Exp(1)$. Testing whether a batch of observations follow an $Exp(1)$ distribution is important in the analysis of spike train data (see Section 19.3.5). We write the specified distribution function as $F_0(x)$ and consider the null hypothesis H_0: $F(x) = F_0(x)$, and we assume $F(x)$ and $F_0(x)$ are continuous.

To test H_0 the discrepancy between empirical cdf $\hat{F}_n(x)$, which satisfies $F_n(x) \to F(x)$ for all x as $n \to \infty$ (see Section 6.2.2), and $F_0(x)$ may be examined. A standard procedure is to consider the largest possible value of the magnitude $|\hat{F}_n(x) - F_0(x)|$, over all x. This is called the *Kolmogorov-Smirnov (KS) statistic*.

> *A detail:* Strictly speaking, because x ranges from $-\infty$ to ∞ there may not be a value of x at which the magnitude $|\hat{F}_n(x) - F_0(x)|$ achieves a maximum. Instead, the supremum is used. (See p. 242.) Therefore, the KS statistic is
>
> $$KS = \sup_x |\hat{F}_n(x) - F_0(x)|.$$
>
> □

The distribution of the KS statistic under H_0 has been studied and, it turns out, does not depend on the choice of null cdf $F_0(x)$. Many statistical software packages provide p-values for the KS test. In particular, for large n we have $p < .05$ when the KS statistic is greater than $1.36/\sqrt{n}$. See Bickel and Doksum (2001, Section 4.1).

10.4 Interpretation and Properties of Tests

We now turn to some theoretical aspects of significance tests. In practice, new situations arise where no standard test is available. Researchers then invent significance tests, and sometimes they are not valid. What do we mean by this? The key property is Eq. (10.24). For an evaluation of statistical significance to be correct, theoretically, (10.24) must be satisfied.

Let $F_Q(x)$ be the cdf of Q under the statistical model specified by H_0 and let us assume that Q follows a continuous distribution. We then have $P(Q \leq q) = 1 - P(Q \geq q)$ and we obtain from (10.24) the equivalent form

$$p = 1 - F_Q(q_{obs}). \tag{10.26}$$

This will help below. Sometimes (10.24) does not hold exactly, but it does hold approximately, as in the case of chi-squared tests. In Section 10.4.1 we derive two consequences that allow us to check whether (10.24) is approximately true. That section describes the behavior of a valid significance test when H_0 is true. In Section 10.4.3 we consider what happens when H_0 is false.

10.4.1 Statistical tests should have the correct probability of falsely rejecting H_0, at least approximately.

The criteria for determining statistical significance, usually taken to be .05 or .01, are called *significance levels.* Fisher suggested[9] that research workers might routinely use $p < .05$ as a "convenient convention" to summarize the evidence against H_0. Indeed, this became standard practice. Neyman and Pearson then considered, formally, the behavior of such a procedure. They began by saying one might *reject* H_0 for sufficiently large values of the test statistic Q. If we let c be the cut-off value for which H_0 is rejected whenever $Q \geq c$, then c is called the *critical value* and

$$\alpha = P(Q \geq c)$$

is called the *level* of the test for the critical value c. Now, for the t-test on p. 265 based on $Q = |T|$ and $q_{obs} = t_{obs}$ defined in (10.19), at a particular level, such as $\alpha = .05$, we may reverse the process and, for any α, we can find a critical value c_α such that

$$\alpha = P(Q \geq c_\alpha). \tag{10.27}$$

[9] See pages 114 and 128 of the fourteenth (1970) edition of Fisher (1925).

For example, the probability of falsely rejecting H_0 based on the criterion $p < .05$ is $\alpha = .05$. Equation (10.27) should hold for any valid test, at least if Q has a continuous distribution (and it should hold approximately for the discrete case).

> *A detail:* For continuous statistics like that in the t-test we can find $c_{.05}$ for which $P(Q \geq c_{.05}) = .05$ and $P(Q \geq c) < .05$ whenever $c > c_{.05}$. In the discrete case, however, only particular values of probabilities actually occur, so there may not exist $c_{.05}$ for which $P(Q \geq c_{.05}) = .05$ and, furthermore, there will be values $a > b$ such that $P(Q > a) = P(Q > b)$. We ignore this technical point here. □

Equation (10.27) gives us a way of checking any test to see whether the fundamental property (10.24) holds: we pick values of c_α, compute the probability $P(Q \geq c_\alpha)$, and see whether the answer is α. For instance, when H_0 holds, we should find $p < .05$ (i.e., $Q \geq c_{.05}$) 5 % of the time and we should find $p < .01$ (i.e., $Q \geq c_{.01}$) 1 % of the time. Another way to say this[10] would be, "if we use $p < .05$ we will be making an incorrect decision 5 % of the time and if we use $p < .01$ we will be making an incorrect decision 1 % of the time."

This calibration of p-values in terms of significance levels is satisfied when (10.24) holds. That is, for any α between 0 and 1, a test that rejects H_0 whenever $p < \alpha$ will have α as its significance level. Formula (10.24) holds for the t-test under the assumption of normality, but without the assumption of normality (10.24) is only approximately correct, as in the first version in Section 10.3.4. Likewise, (10.24) holds only approximately for the p-values computed from the chi-squared distribution based on the chi-squared statistics in Section 10.1. Similarly, when a new statistical test is proposed to deal with a complicated or unusual situation, it may provide approximate p-values. For approximate tests it is good to know how close the p-value is to being correct. In this case it is valuable to verify, by computer simulation, that the test has approximately the level $\alpha = .05$ when $p < .05$, and similarly for other levels such as $\alpha = .01$. For illustrative purposes we carried out the calculation in the case of the example on blindsight of patient P.S.

Example: Blindsight of P.S. Let us consider the use of χ^2_{obs} as we did on p. 257. For a χ^2_1 distribution we have $c_{.05} = 3.84$, i.e., if $X \sim \chi^2_1$ then $P(X \geq 3.84) = .05$. For the case $n = 17$ and $p_0 = .5$ we may compute the value of $\alpha = P(Q \geq 3.84)$ where Q is the chi-squared statistic. This is easily done by computer simulation. We obtained $\alpha = .049$. Repeating this for $c_{.01} = 6.63$ we obtained $\alpha = .013$. For these standard cut-off values for p, and for this sample size, we conclude that the χ^2_1 distribution furnishes an accurate approximation.[11] □

---

[10] Fisher objected to the idea that statistical significance should be equated with decision making about hypotheses. From our modern perspective this is an objection about the words used to describe (10.27) but the formula itself is crucial. We say more about this in Section 10.4.7.

[11] On the other hand, we should recall that the p-value we obtained for the data $x = 14$ was $p = .0076$ based on χ^2_{obs} and the chi-squared distribution while the exact p-value was $p = .0127$. The discrepancy between approximate and exact values is a bit larger; the approximation apparently gets worse as we move further out into the tails.

Equations (10.24) and (10.27) provide explicit statements of the behavior of a significance test under the assumption that H_0 is true. Let us continue to assume that H_0 is true and go a step further by observing that the p-value is, itself, a random variable and inquiring about its distribution. If we ask, "How often do we get $p < .05$?" the answer, for any valid test, according to (10.27), is 5 % of the time; if we ask "How often do we get $p < .01$?" the answer is 1 % of the time; if we ask "How often do we get $p < .25$?" the answer is 25 % of the time. In general, we must get $p < \alpha$ with probability α. But if a random variable X satisfies $P(X < \alpha) = \alpha$ then $X \sim U(0, 1)$. (Assuming X is continuous then $P(X < x) = P(X \leq x) = F_X(x) = x$, which is the cdf of the $U(0, 1)$ distribution.) Therefore, when H_0 holds, the p-values from a valid significance test will be uniformly distributed between 0 and 1.

Details: If we were to repeatedly sample data according to the statistical model specified by H_0, then we would get random values of q_{obs}. Let us denote such random values by the random variable Y. By the way we are constructing Y it has the same distribution as Q. To be even more specific, let us denote the mapping from data values $x_1, \ldots, x_n$ to y values by $y = T(x_1, \ldots, x_n)$ so that $Y = T(X_1, \ldots, X_n)$. The definition (10.24) could be rewritten in terms of y as

$$p = P(Q \geq y|H_0) = P(Q \geq T(x_1, \ldots, x_n)|H_0). \tag{10.28}$$

Now, just as repeated samples would give random values of y so, too, would repeated samples give random values of p. Let us denote such random values by the random variable P. The random variable P satisfies

$$P = P(Q \geq Y|H_0) = P(Q \geq T(X_1, \ldots, X_n)|H_0). \tag{10.29}$$

With this notation in hand, we show that the theoretical distribution of p-values under H_0 is uniform.

Theorem Let $X_1, \ldots, X_n$ be a random sample from which P is defined from (10.29), and assume Q follows a continuous distribution. If H_0 holds then $P \sim U(0, 1)$.

Proof: From the first equality in (10.29) we have

$$P = 1 - F_Q(Y),$$

which is the random variable version of (10.26). Because Y follows the same distribution as Q, $F_Q(y) = F_Y(y)$, so that

$$P = 1 - F_Y(Y)$$

and

$$1 - P = F_Y(Y).$$

From the probability integral transform given in Section 3.2.5 it follows that $1 - P$ has a $U(0, 1)$ distribution. It is an easy exercise to show that $X \sim U(0, 1)$ if and only if $1 - X$ is $U(0, 1)$. Therefore, $P \sim U(0, 1)$. □

We also have the following.

Theorem Let $X_1, \ldots, X_n$ be a random sample from which P is defined from (10.29), and assume Q follows a continuous distribution. Then, under H_0, the probability that $P < \alpha$ is equal to α, i.e.,

$$P(P < \alpha | H_0) = \alpha. \tag{10.30}$$

Proof: This is a corollary to the previous theorem: because $P \sim U(0, 1)$ we have $F_P(x) = x$ which, because Q is continuous, is the same as (10.30). □

Example 7.2 (continued from p. 265) To illustrate the uniformity of p-values guaranteed by the theorem, we generated samples of pseudo-data based on the real data used in the t-test on p. 265. The idea was to begin with the 60 data values under the SSSS and SSST conditions and create 10,000 sets of pseudo-data like the real data except that for each set of pseudo-data H_0 was true. To force H_0 to hold we sampled the 60 data values and then arbitrarily put them into two groups of 30 values each, so that each of the two groups of pseudo-data would follow the same distribution.[12] We repeated this to get the 10,000 sets of pseudo-data, and then ran the t-test and computed the p-value for each set of pseudo-data. Figure 10.1 is a histogram of the resulting 10,000 p-values. The distribution is uniform. □

10.4.2 A confidence interval for θ may be used to test H_0: $\theta = \theta_0$.

Let us return to the "paradigm case" of Section 7.3.2 in which $X_1, \ldots, X_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution with the value of σ known. In Section 7.3.2 we found a confidence interval for μ. Now let us consider, instead, the null hypothesis H_0: $\mu = 0$. This hypothesis comes up frequently because many experiments generate, for each subject, one observation under each of two conditions, and the data may be reduced by taking the difference of the two observations. Thus, instead of n pairs of observations we analyze n single-number differences X_i and the null hypothetical question becomes whether the mean of these differences is zero. In practice, the value of σ is unknown but here, as in Section 7.3.2, we assume it is known in order to simplify the derivation below.

As in Section 7.3.2 we have standard error $SE(\bar{X}) = \sigma/\sqrt{n}$. In Section 7.3.2 we showed that the interval $(\bar{X} - 2 \cdot SE(\bar{X}), \bar{X} + 2 \cdot SE(\bar{X}))$ is a 95 % CI for μ, which

[12] Specifically, both groups followed the distribution specified by the empirical cdf based on the 60 data values. This is an example of *bootstrap sampling* and will lead to a *bootstrap test* discussed in Chapter 11.

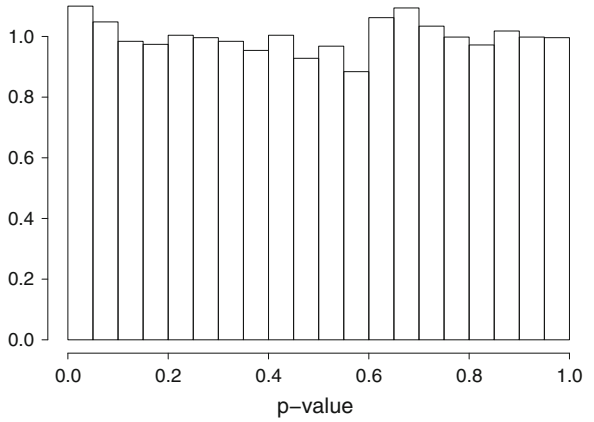

Fig. 10.1 Histogram of test-enhanced learning p-values under H_0. Each p-value was computed by sampling at random the 60 data values under the SSSS and SSST conditions and arbitrarily putting them into two groups of 30 each, then running a t-test, as in the t-test on p. 265. This was repeated 10,000 times. The histogram was normalized by dividing the number of observed p-values, in each bin, by the number expected if they followed a $U(0, 1)$ distribution.

means

$$P(\bar{X} - 2 \cdot SE(\bar{X}) \leq \mu \leq \bar{X} + 2 \cdot SE(\bar{X})) = .95.$$

To test H_0: $\mu = 0$ we can check whether our 95 % CI contains 0. If it does not, we have evidence against H_0.

Theorem Suppose $X_1, \ldots, X_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution, with the value of σ known. If $H_0 : \mu = 0$ holds, then we have

$$P(0 \notin (\bar{X} - 2 \cdot SE(\bar{X}), \bar{X} + 2 \cdot SE(\bar{X}))) = .05.$$

Proof: For every μ we have

$$\begin{aligned} & P(\mu \notin (\bar{X} - 2 \cdot SE(\bar{X}), \bar{X} + 2 \cdot SE(\bar{X}))) \\ &= 1 - P(\mu \in (\bar{X} - 2 \cdot SE(\bar{X}), \bar{X} + 2 \cdot SE(\bar{X}))) \\ &= 1 - .95 = .05. \end{aligned}$$

The result follows by taking $\mu = 0$. □

This theorem says that the confidence interval for μ may be *inverted* to produce a test of H_0: $\mu = 0$. We use the term "inverted" because instead of looking *within* the interval, as we do in the usual application of a confidence interval, in testing H_0 we are seeing whether it lies *outside* the confidence interval. When $\mu = 0$ lies outside the confidence interval we reject H_0 with significance level $\alpha = .05$, and can report $p < .05$.

The same logic may be used to state a version of the theorem in more general form.

Theorem Suppose $X_1, \ldots, X_n$ is a random sample from a distribution that depends on a single parameter θ, and suppose $(\hat{\theta} - 2 \cdot SE(\hat{\theta}), \hat{\theta} + 2 \cdot SE(\hat{\theta}))$ is a 95 % CI, i.e.,

$$P(\hat{\theta} - 2 \cdot SE(\hat{\theta}) \leq \theta \leq \hat{\theta} + 2 \cdot SE(\hat{\theta})) = .95.$$

If $H_0 : \theta = \theta_0$ holds, then we have

$$P(\theta_0 \notin (\hat{\theta} - 2 \cdot SE(\hat{\theta}), \hat{\theta} + 2 \cdot SE(\hat{\theta}))) = .05.$$

Proof: The argument is the same here as for the previous theorem. □

This theorem says that whenever we have a 95 % confidence interval for a parameter, we may invert it to get a test of a null hypothesis that takes the form $H_0 : \theta = \theta_0$. We stated the theorem to indicate generality, but actually the paradigm case of the normal sample with σ known furnishes one of the rare situations in which a standard confidence interval has exactly the correct overage probability of .95. More commonly we rely on intervals that have *approximate* coverage probability .95. The method of using an approximate 95 % confidence interval to test a hypothesis produces a significance level of approximately $p = .05$ (we might write $p \approx .05$). In practice, the null-hypothetical value should be far outside the confidence interval, as in Example 1.4 in Chapter 1.

10.4.3 Statistical tests are evaluated in terms of their probability of correctly rejecting H_0.

In Section 10.4.1 we pointed out that a statistical test should have its significance levels match reasonably well its reported p-values, at least in the case of .05 and .01, and that this results in incorrect rejection of H_0 with the putative frequency (e.g., 5 or 1 % of the time). But suppose we have two different ways of testing a hypothesis. How should we judge which way is better?

To answer this question, we may consider not only incorrect rejection of H_0 but also an incorrect decision not to reject. The two possible decisions may be identified as "reject H_0" and "accept H_0." There are then two types of errors: incorrectly rejecting H_0 when it is in fact true, and incorrectly accepting H_0 when it is in fact false. These are called *type I* and *type II* errors. A good test would be one with small *type I* and *type II* errors. In order to evaluate the type II error we must introduce a particular non-null hypothesis. This is called the *alternative hypothesis* and is usually denoted H_A (or H_1). The *power* of the test is then the probability of correctly rejecting H_0

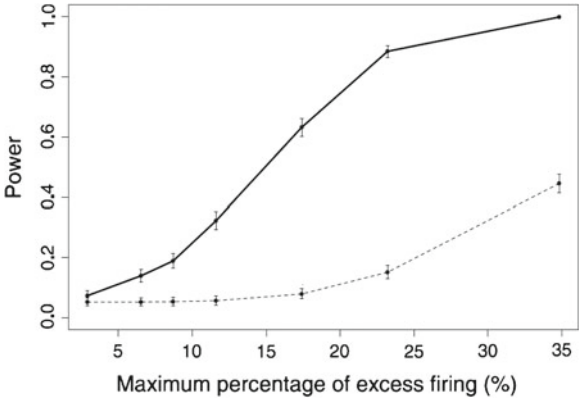

Fig. 10.2 Power of the method proposed by Ventura et al. (2005a), shown in the *black line* (*the upper curve*), compared with an alternative method, shown in the *thin line* (*the lower curve*). Power is plotted against the maximum percentage excess firing above that predicted by independence. Both tests have the same value $\alpha = .05$, indicated by the y-axis value when the percentage excess firing is zero (so that H_0 holds). The power of the new method is much greater than the power of the alternative method. Adapted from Ventura et al. (2005).

when H_A is true, i.e., it is one minus the probability of a type II error. The probability of a type II error is usually denoted by β. Thus, for a test based on a statistic Q, using (10.27), we have

$$\alpha = P(Q \geq c_\alpha | H_0) \tag{10.31}$$

$$\beta = P(Q < c_\alpha | H_A) \tag{10.32}$$

and

$$\text{power} = 1 - \beta. \tag{10.33}$$

If we have two different tests that we want to compare, we may choose a value α, such as $\alpha = .05$, find for each test its critical values c_α, such as $c_{.05}$, and then ask, for a particular H_A, which test is more powerful in the sense of having a larger value of $1 - \beta$ given by (10.32). This is the general program laid out by Neyman and Pearson, and it is the standard way to evaluate competing statistical tests of hypotheses.

Example 10.5 Time-varying dependence between spike trains Ventura et al. (2005a) proposed a bootstrap method of testing the null hypothesis of independence between two spike trains. Their method not only tested independence but also found a window of time over which the two spike trains had increased joint activity. To compare the new method to an existing procedure (which instead used contiguous time bins in the joint peri-stimulus time histogram), Ventura et al. computed power using (10.32) and (10.33) for a particular series of scenarios as the excess joint firing, above that predicted by independence, was increased. Figure 10.2 is a plot of power as a function of excess firing rate for the two methods. The purpose of such a plot is

to demonstrate the superiority of a proposed method to an existing alternative. The plot in Fig. 10.2 indicates especially large gains in power for 15–20 % excess joint activity. □

Another use of power is to determine sample size. The idea is to choose an alternative H_A, considered to be plausible, and ask how big a sample size would be needed to achieve both a particular level α and a particular power $1 - \beta$. The values $\alpha = .05$ and $1 - \beta = .8$ are often used in medical applications, and planners of clinical trials typically must show to reviewers their calculation that the proposed sample size meets such specifications under reasonable assumptions.

10.4.4 The performance of a statistical test may be displayed by the ROC curve.

In (10.31), (10.32), and (10.33), the critical value c_α was determined by the level, as in (10.27). We can turn things around and instead think of varying the critical value $c = c_\alpha$, which then makes the level and power depend on c, and we then make this dependence explicit by writing

$$\alpha(c) = P(Q \geq c|H_0) \tag{10.34}$$

and

$$\beta(c) = P(Q < c|H_A). \tag{10.35}$$

The choice of c determines a trade-off of type I and type II errors: when c is increased, $\alpha(c)$ gets smaller so type I error probability decreases and type II error probability $\beta(c)$ increases. An ideal test would have small level and large power. As c is increased the level gets smaller—which is desirable—but the power $1-\beta(c)$ also gets smaller. The performance of a test may be examined by plotting $1 - \beta(c)$ versus $\alpha(c)$ for a range of values of c. The function $y = f(x)$ that traces values $(x, y) = (\alpha(c), 1-\beta(c))$ is called the *receiver-operating characteristic (ROC)* curve.

The simplest setting is the paradigm case of Section 10.3.1, where $\bar{X} \sim N(\mu, \sigma^2/n)$ and we wish to test $H_0 : \mu = \mu_0$. If H_0 holds, then the ratio Z defined in (10.5) satisfies $Z \sim N(0, 1)$ but if $H_A : \mu_1$ holds with $\mu_1 \neq \mu_0$, then $Z \sim N(\delta, 1)$ where $\delta = (\mu_1 - \mu_0)/SE(\bar{X})$. The ROC curves for $\delta = 2$ and $\delta = 1$ are shown in Fig. 10.3. When $\delta = 1$ it is more difficult to discriminate between the two alternatives; the power $(y = 1 - \beta)$ is lower for a given value of the level $(x = \alpha)$ and the ROC curve is closer to the line $y = x$ (which is the ROC curve when $\delta = 0$). If we were instead to pick a very small value of δ the ROC curve would essentially fall on the line $y = x$, while if we picked a very large value of δ the ROC curve would hug the y-axis near $x = 0$ and hug the asymptote $y = 1$ for increasing values of x. Thus, the higher the curve, the better its overall performance. Sometimes tests are compared by plotting their ROC curves. In addition, the area under the curve is often

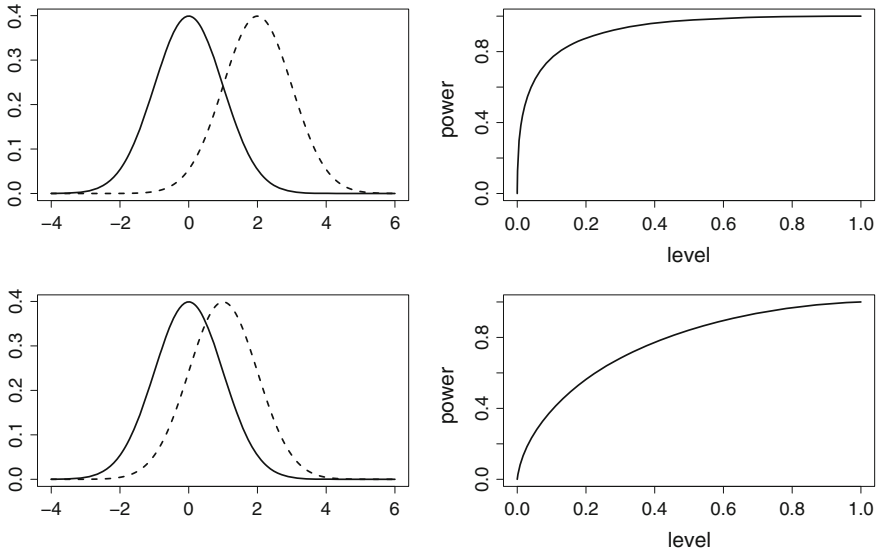

Fig. 10.3 Two pairs of normal distributions and the resulting ROC curves. The *Left-hand side* shows the pair of pdfs for $N(0, 1)$ (*solid*) and $N(\delta, 1)$ (*dashed*) and to the *Right* are the corresponding ROC curves. *Top* $\delta = 2$. *Bottom* $\delta = 1$.

evaluated: it is 1 (the area of the 0–1 square) for a perfect test and .5 (the area within the square under the line $y = x$) for tests with no predictive ability at all (in the normal case corresponding to $\delta = 0$).

The ROC curve is also used in psychophysical analysis of perceptual detection of stimuli, called *signal detection theory* (SDT). According to SDT, perception involves a noisy unknown process in the brain of a subject, which may be considered a random variable X. In SDT two stimuli are considered. The first is taken to be a null stimulus, in analogy with a null hypothesis, and the second involves a stimulus of interest. For example, a subject may be repeatedly shown images and asked to detect whether a particular object appeared in the image. In null cases the images would not contain the object, but in cases involving the stimulus of interest—those involving the *signal* in addition to the noise—they would. The two stimuli, null (noise alone) and signal-plus-noise, generate two distinct probability distributions for X, as in Fig. 10.3. If X is sufficiently large, say $X > c$, then the subject responds. In the setting of object recognition, the subject would respond that the object is present whenever $X > c$. This will incur two types of errors, analogous to type I and type II errors. The SDT problem is then to characterize the null and signal-plus-noise distributions of X.

To characterize the null and alternative distributions of X points along an ROC curve are found by using judgment confidence to vary the cutoff c. That is, the subject is asked to report the confidence with which the judgment of perception is made: when confidence in perception is high, the probability of a type II error should be low, which corresponds to c being small. A confidence rating scale such

as 1–6 is typically used; from many repeated trials, with a 6 point scale, 5 points are obtained along an empirical ROC curve. (For the lowest confidence value there is never a perception of response at all, so it is considered to correspond to $c = \infty$.) The common terminology used in SDT replaces the y-axis label of power with "hit rate," or "hits," and the x-axis label of level with "false alarm rate," or simply "false alarms." As in the case of statistical tests, the null and signal-plus-noise distributions are often assumed to be normal, but that is not essential to the logic of the method.

Example 10.6 Dual-process theory of memory One method of studying memory has involved recall of words taken from a list that was previously studied. A variant of this uses a list of words consisting of some previously studied (or "old") words together with some new words; then, for each word taken from the composite list the experimental subject is asked to say whether the word is new or old. This produces a series of binary judgments to which SDT may be applied: the old words define the signal-plus-noise condition, while the new words define the null condition. According to certain dual-process theories of memory, there is a distinction between remembering based on some set of details or related events, and remembering without such corresponding details being available and, instead, there is only a sense of "familiarity." Yonelinas (2001) reported an experiment in which 19 subjects were each given a list of 58 words to study, and then were tested on a composite list of 75 words. Half of the old words were studied under "full attention" and half were studied under "divided attention." In the full attention condition subjects saw each word for 1.5 s (seconds) and were instructed to try to remember it. In the divided attention condition the subjects also had to judge the magnitude of a number, presented on the same screen as the word. The composite list of 75 test words consisted of 25 old words studied under full attention, 25 studied under divided attention, and 25 new words. The subjects were required to judge whether each test word was new or old using a 6 point scale (ranging from "sure it was new" to "sure it was old") and then, after the judgment had been made (and the word was no longer visible), they were also required to indicate whether they could remember details about the word, such as what it looked like or sounded like, and whether they would be able to report such details. Words were considered to be recognized based on familiarity when no details could be recalled.

According to the dual-process model of Yonelinas, ROC curves for familiar objects should be similar to those obtained from a pair of displaced normal distributions, as in Fig. 10.3, whereas words recollected with details would have a constant probability of memory retrieval once a minimal confidence threshold was exceeded. A pair of ROC curves for the familiarity words, in both the full attention and divided attention conditions, are shown in the left-hand part of Fig. 10.4. As support for the dual-process theory, Yonelinas also presented ROC curves for the words recognized with detailed recollection, and these curves were quite flat, with an apparent threshold at which recollection occurred. These are in the right-hand part of Fig. 10.4. □

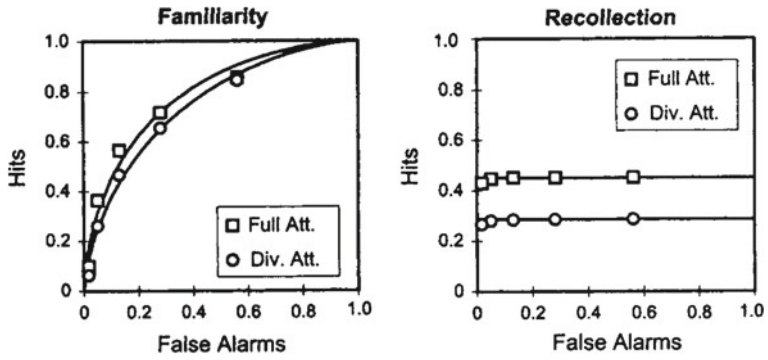

Fig. 10.4 ROC curves, adapted from Yonelinas (2001). On the *Left* are curves from words recognized based only on familiarity, and on the *Right* are curves from words for which recognition was based on detailed recollection. Curves for full attention and divided attention words are plotted separately.

10.4.5 The p-value is not the probability that H_0 is true.

The p-value is commonly misinterpreted as the probability that the null hypothesis is true. A correct statement is necessarily rather cumbersome. Let us continue to write a generic test statistic as Q and the value it takes when calculated from data as q_{obs}. In the case of the chi-squared tests we used $Q = X \sim \chi^2_\nu$ with $x_{obs} = \chi^2_{obs}$ and for the two-sided t test (10.19) we used $Q = |T|$ with $q_{obs} = |t_{obs}|$. We chose the notation q_{obs} so that we can clearly distinguish the observed value from the theoretical random variable Q. The p-value is then given by Eq. (10.24). In words, p is the probability that one *would observe* a value of the test statistic as discrepant from the null hypothesis as the one observed from the data, *if the null hypothesis were true*. Or, again, in slightly different words: if the null hypothesis were true, the test statistic Q would have a probability distribution; the p-value is the resulting probability that Q would be as discrepant from the null hypothesis as the value q_{obs} actually observed. There is no substantially simpler way to say this. The important point about the correct interpretation is its subjunctive nature: the p-value is a probability based on what *might* have happened if a random sample had been drawn under H_0.

Because the logic behind p-values is somewhat convoluted, they are very often misinterpreted to mean something much simpler and more direct, namely the probability that H_0 is true based on the data. That is, a value $p = .05$ is often misinterpreted as meaning that .05 is the probability that H_0 is true, which we would write as $P(H_0|data) = .05$. This is sometimes called the *p-value fallacy* (Goodman 1999a, b). There is no denying how nice it would be to have $P(H_0|data)$. In principle, that probability may be obtained, instead, from Bayes' Theorem:

$$P(H_0|data) = \frac{P(data|H_0)P(H_0)}{P(data|H_0)P(H_0) + P(data|H_A)P(H_A)}. \tag{10.36}$$

From a practical point of view, however, application of formula (10.36) requires considerable care. We return to it in Section 16.3. Bayes' Theorem does provide some guidance on the calibration of p-values. As we discuss in Section 16.3.4, under reasonable assumptions $p = .05$ corresponds to values of $P(H_0|data)$ much larger than .05 and, in fact, only a little smaller than $1/2$. In other words, in terms of $P(H_0|data)$, a p-value of .05 is likely to provide only marginal evidence against H_0.

10.4.6 Borderline p-values are especially worrisome with low power.

To better understand the meaning of p-values it is also worth keeping in mind a different Bayesian analysis than that described in Section 16.3 and alluded to in Section 10.4.5. In Section 3.1.4 we applied Bayes' Theorem to screening tests. There, we calculated the probability that a patient might have a disease, which we denoted as the event A, based on a positive screening test outcome B. That is, we computed $P(A|B)$ based on the sensitivity of the test $P(B|A)$, the specificity of the test $P(B^c|A^c)$ and the prevalence of the disease $P(A)$. We showed that when $P(A)$ is small, $P(A|B)$ will be much smaller than one might expect based on seemingly good values of the specificity and sensitivity.

The same kind of analysis may be applied to significance tests by viewing them as analogous to screening tests. In this case A becomes the event that H_0 is false, B becomes the event that the test rejects H_0, the sensitivity $P(B|A)$ becomes the power of the test, $1 - \beta$ (see (10.33)), the specificity $P(B^c|A^c)$ becomes $1 - \alpha$, and the prevalence $P(A)$ becomes the prevalence of false null hypotheses. Plugging these into Bayes' Theorem, and using the definition of positive predictive value (PPV) on p. 44 with Eq. (3.2), we get

$$\begin{aligned} P(H_0 \text{ false}|\text{test rejects } H_0) &= PPV \\ &= \frac{(\text{power})(\text{prevalence})}{(\text{power})(\text{prevalence}) + \alpha(1 - \text{ prevalence})}. \end{aligned}$$

If we assume that prevalence of false null hypotheses is less than .5, so that truly false null hypotheses are somewhat rare, and further that the power is low, then a borderline p-value slightly less than .05, corresponding to a rejection rule with $\alpha = .05$, will lead to very weak evidence against H_0. For example, if the prevalence of false null hypotheses is .2 and the power is also .2, then using $\alpha = .05$ we find

$$P(H_0 \text{ false}|\text{test rejects } H_0) = \frac{1}{2}.$$

In other words, in this situation $p < .05$ provides no evidence that the given null hypothesis is false. More generally, when a study is based on a small sample size and therefore lacks power, while false null hypotheses are uncommon, a finding of marginal statistical significance is likely to provide little or no evidence against H_0. This argument has been used by Button et al. (2013) to suggest that many statistically significant results in neuroscience are likely to be spurious.

These Bayesian analyses, here and in Section 16.3, referenced in Section 10.4.5, lead us to advise that, in most situations, to be safely taken as supplying substantial evidence against H_0, p-values should be very much smaller than .05.

10.4.7 The p-value is conceptually distinct from type one error.

We began by presenting p-values as a way of assessing evidence against a null hypothesis, and then reviewed the basic elements of the additional hypothesis testing framework based on evaluation of the performance of a test under both null and alternative hypotheses. The latter was introduced originally by Neyman and Pearson. Fisher disliked the Neyman-Pearson conception because he thought the alternative hypothesis was artificial and unnecessary—more than that, he thought it was counter-productive. In the Neyman-Pearson scheme there was no apparent role for p-values: in principle, one would pick a level α (such as $\alpha = .05$) *a priori* and then determine whether $p < \alpha$ rather than reporting the p-value itself. Furthermore, the implication was that, in practice, the null hypothesis might routinely be accepted rather than rejected. This was the point that Fisher found most troubling. He said, "It is certain that the interest of statistical tests for scientific workers depends entirely [on] their use in rejecting hypotheses which are thereby judged to be incompatible with the observations." (Fisher 1935.) From our current vantage point it is easy enough to step back from that early controversy. On the one hand, Fisher was correct that p-values and the rejection of statistical hypotheses would become a major activity of everyday science. On the other hand, the Neyman-Pearson conceptions have proven their worth in theoretical work, where evaluation of type I and type II errors have been important in understanding alternative testing procedures. The modern point of view is thus a synthesis of Fisher's "significance testing" and the Neyman-Pearson "hypothesis testing." There is no longer a compelling need to distinguish between these separate notions, which were once considered incompatible. We use the terms "significance testing" and "hypothesis testing" interchangeably.

10.4.8 A non-significant test does not, by itself, indicate evidence in support of H_0.

In previous subsections we have laid out the logic of significance testing using p-values. As we noted at the beginning of Section 10.4.1, Fisher's original

conception was that small p-values could provide evidence against H_0, and in Section 10.4.7 we cited his concern that they not be used for "accepting" a null hypothesis. In this regard, the modern view is consistent with Fisher's interpretation of p-values: they can only be used to show how the data appear to be inconsistent with H_0; they do not supply support for H_0. A non-significant test of $H_0 : \theta = \theta_0$ could occur either because H_0 holds or because the variability is so large that it is difficult to determine the value of the unknown parameter. The latter possibility must be considered.

As an illustration, let us return to the blindsight example, Example 1.4, once again and imagine a different outcome. Suppose that, instead of 14/17 "non-burning" house selections, patient P.S. had chosen the non-burning house 12 out of 17 times. If $X \sim B(17, .5)$, an exact calculation like that on p. 268 gives

$$p = 2F(5) = .14.$$

In this circumstance it would be *incorrect* to say that there is evidence in favor of H_0. In fact, for 12 out of 17, the estimate of the propensity of P.S. to choose the non-burning house would be $\hat{p} = 12/17 = .71$ with standard error $SE(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n} = .22$. While it is true that the value $H_0: p = .5$ is clearly consistent with the data, the standard error is so large that a wide range of non-null values are also consistent with the data.

It is very common for investigators to interpret failure of a test to reach significance as an indication that $H_0 : \theta = \theta_0$ holds. This is reasonable *only* if, in addition, the standard error of the estimate $SE(\hat{\theta})$ is small: a confidence interval would have to include only those values of θ that are, for practical purposes, essentially the same as θ_0.

It is especially tempting to misinterpret a non-significant test when results from two situations are being compared, and statistical significance is obtained in one situation but not the other. We return to this point in Section 13.2.2 when we discuss interaction effects in ANOVA.

Example 10.7 Synchronous firing of V1 neurons Synchronous neural activity is widely believed to play an important role in neural computation (e.g., Uhlhaas et al. 2009) but its statistical assessment is subtle (see Harrison et al. 2013). Suppose we have two spike trains, each represented as binary time series using some small windows of time, as in Fig. 5.2, where a 1 signifies that a spike has occurred and a 0 that no spike has occurred. When both time series have a 1 in the same time bin we say that the two neurons have fired synchronously. Under reasonable statistical models, some synchronous spikes will occur by chance even if the two neurons are firing independently. The statistical problem is to identify synchronous firing that occurs more frequently than predicted by chance alone. Kass et al. (2011) provided a statistical framework for evaluating synchronous spikes (see also Kelly and Kass 2012). They analyzed two pairs of neurons recorded from primary visual cortex (V1) in an anesthetized monkey during visual exposure to moving grating stimuli. They defined a quantity ξ_H that represented the proportional gain in synchronous firing rate above that expected under independence (actually, conditional independence

given measured network activity). The null hypothetical value under independence was $H_0 : \xi_H = 1$, which they restated as $H_0 : \log \xi_H = 0$. For one pair of neurons they reported $\log \hat{\xi}_H = .06$ with $SE = .15$ giving a t-ratio of .39. Their conclusion was that these data were consistent with H_0. Here, they were not relying on the significance test alone: a confidence interval would exclude substantial values of $\log \xi_H$. Specifically, an approximate 95 % confidence interval for $\log \xi_H$ based on (7.8) is (−.24, .36) and when transformed to the ξ_H scale it becomes (.79,1.4), which eliminates as highly unlikely excess synchronous firing rates of 40 % above independence. (Here $\exp(-.24) = .79$, $\exp(.36) = 1.4$, and the 40 % figure comes from the right-hand CI limit of 1.4.) The authors contrasted this pair of neurons with a different pair, for which they obtained $\log \hat{\xi}_H = .82$ with $SE = .23$ giving a t-ratio of 3.57, which leads to an approximate 95 % confidence interval for ξ_H of (1.4, 3.6).

The physiological point was that distinct pairs of neurons in V1 may respond quite differently with regard to synchronous spiking in excess of that produced by network activity: the first pair produced synchronous spikes at roughly the rate they would be produced under independence, while the second pair produced synchronous spikes at roughly double the rate expected under independence ($\exp(.82) = 2.3$, with confidence interval (1.4, 3.6)). The statistical point is that the results of the significance tests, alone, did not adequately convey what the data were able to show about the excess synchronous firing rates in these neurons. Standard errors or confidence intervals are also necessary. □

10.4.9 One-tailed tests are sometimes used.

We summarized the logic of p-values in Eq. (10.24), and the surrounding discussion, taking q_{obs} to represent the value of a generic statistic used to test a null hypothesis. In nearly all of the special cases we have examined we have chosen q_{obs} to be the absolute value of some statistic, and then Q was the absolute value of the corresponding random variable. For example, in testing $H_0 : \mu = \mu_0$ we used either $q_{obs} = |z_{obs}|$ or $q_{obs} = |t_{obs}|$. A different choice is to remove the absolute value. This version of significance testing sometimes appears in the literature. It is called a *one-sided test* and it corresponds to testing H_0 against a *one-sided alternative hypothesis*, such as $H_A : \mu \geq \mu_0$. Let us discuss this by way of our most heavily-used example.

Example 1.4 (see p. 268) We previously posed the statistical problem of testing H_0: $p = .5$, which corresponds to saying that P.S. was guessing, and on p. 268 we obtained the exact p-value $p = .013$. We might, instead, say that we are interested *only* in the case in which P.S. might have chosen the non-burning house *more often* than half the time. In other words, we might say that we care about the possibility that her propensity to choose the non-burning house was $p > .5$ and, therefore, we would pay attention only to data for which $\hat{p} > .5$. In this case we would replace (10.13) with

$$p = P(Z \geq z_{obs})$$

which we compute (modifying the calculation on p. 268) as $P(X \geq 14) = P(X \leq 3) = .0064$, where $X \sim B(17, .5)$. This new p-value is half the size of the previous value, and thus would indicate stronger evidence against this null hypothesis than suggested previously. □

This example introduces the standard dilemma of one-sided versus two-sided testing. If one-sided testing is used, the p-value is cut in half and the evidence appears stronger. On the other hand, the alternative hypothesis has been changed. Which alternative hypothesis is more appropriate?

In order to use the one-sided hypothesis one must argue that a reverse result *would have been ignored by the data analyst*. In Example 1.4, such a claim would mean that if patient P.S. had consistently chosen the *burning* house, we would have ignored the data. This seems implausible to us. In the extreme case, if P.S. *always* chose the burning house, it might have been odd, but it surely would have provided evidence that her brain perceived the flames on the left side of the visual field. Our feeling is that the vast majority of cases are analogous to this example: the reverse result would almost always be of some interest, and it is therefore almost always preferable to use the two-sided test. Furthermore, the two-sided test is conservative in the sense of providing double the p-value (it is less likely to lead, by chance alone, to the conclusion that there is evidence against H_0) and we regard this feature as an advantage as well.[13] If a one-sided test must be used in order to claim statistical significance, the data are not conclusive and provide only weak evidence against the null hypothesis.

[13] Part of our reasoning comes from Bayesian calibration of significance tests, which is discussed briefly in Section 10.4.5 and again in Section 16.3.4.

Chapter 11
General Methods for Testing Hypotheses

In Chapter 10 we laid out the main ideas in assessing statistical significance. First, there is a null hypothesis; second, there is a statistic that defines some deviation away from a null model; third there is a p-value to judge the rarity of the observed deviation under the null hypothesis. These are the three essential ingredients of a statistical hypothesis test. We also discussed several aspects of the interpretation and evaluation of statistical tests. While Chapter 10 provided the basic notions of testing, it did so within a few simple settings. After presenting goodness-of-fit for data in categories, we considered hypotheses involving restriction of a parameter to a single value, equality of two proportions, and equality of two means. These hypotheses were chosen partly because they occur very frequently, but also because the test statistic in each case is highly intuitive. What happens when one is faced with a new problem that does not fit one of these molds? How should the statistical test be defined?

In estimation, maximum likelihood plays a unifying role and helps solve new problems: many familiar and intuitive estimators are actually maximum likelihood estimators, ML estimation may be applied in many novel situations and, it turned out, ML estimation was optimal for large samples. For testing problems there is an analogous method: the *likelihood ratio test*. This test is also quite general; it has large-sample optimality properties; and it produces as special cases familiar procedures such as the t-test. Likelihood ratio tests are the subject of Section 11.1.

ML estimation is applicable to problems involving parametric specification of statistical models. In Section 9.2.2 we discussed the parametric bootstrap, which may be applied in conjunction with ML estimation and in Section 9.2.3 we showed how the nonparametric bootstrap could be applied without the parametric specification in the statistical model—thus, its name. Similarly, there is a nonparametric bootstrap method of testing hypotheses. We discuss this, and the closely related permutation tests, in Section 11.2.

The procedures in Chapter 10 and in Sections 11.1, 11.2 treat single, isolated hypotheses. In practice one often faces many hypotheses, all of which need to be tested. This creates what is known as the *multiple testing problem*, which we treat in Section 11.3.

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_11,

11.1 Likelihood Ratio Tests

Where do statistical tests come from? Sometimes they are based on intuition. A particular discrepancy measure may seem sensible as a way to capture the relevant departure from H_0. For instance, in the case of patient P.S. in Example 1.4 it would seem reasonable to use a test based somehow on $|\hat{p} - p_0|$, and in Section 10.3.2 we suggested the ratio $(\hat{\theta} - \theta_0)/SE(\theta)$ could be used when H_0 involves only a single, scalar parameter, or a single component of a parameter vector, or a scalar function of a parameter vector. What about hypotheses that involve multiple parameters? Just as ML estimation is widely applicable to parametric estimation problems, the *likelihood ratio test* may be used in parametric testing problems. In this section we review the essential methods and results on the likelihood ratio test, but do not provide many examples. A major source of applications is the body of methods associated with generalized linear models, which provide important generalizations of linear regression including the logistic regression model we presented in Example 5.5. We discuss the way the likelihood ratio test is used with generalized linear models in Chapter 14.

11.1.1 The likelihood ratio may be used to test $H_0 : \theta = \theta_0$.

The likelihood function assigns to alternative values of θ their plausibility in light of the data $L(\theta)$. It can be used, analogously, when a particular value of θ is singled out in the form of a null hypothesis $H_0 : \theta = \theta_0$. That is, we consider the value $L(\theta_0)$ and assess whether it is nearly the same as the maximal value $L(\hat{\theta})$. Here, θ could be either a scalar or a vector. Suppose we have data $x_1, \ldots, x_n$ that are assumed to have a joint pdf $f(x_1, \ldots, x_n|\theta)$. We define the *likelihood ratio* test statistic to be

$$LR_{obs} = \frac{f(x_1, \ldots, x_n|\theta_0)}{f(x_1, \ldots, x_n|\hat{\theta})}. \tag{11.1}$$

Because the MLE maximizes the likelihood function, we have $LR_{obs} \leq 1$. If we apply the same formula to a random sample $X_1, \ldots, X_n$, we get the theoretical version of the likelihood ratio as the random variable

$$LR = \frac{f(X_1, \ldots, X_n|\theta_0)}{f(X_1, \ldots, X_n|\hat{\theta})}. \tag{11.2}$$

We now define the test procedure.

Likelihood ratio test of $H_0 : \theta = \theta_0$. For a random sample $X_1, \ldots, X_n$ with joint pdf $f(x_1, \ldots, x_n|\theta)$, the likelihood ratio test evaluates LR_{obs} defined in (11.1) and assigns the p-value

$$p = P(LR < LR_{obs}|H_0) \tag{11.3}$$

where LR is defined in (11.2).

Note that it is equivalent to examine the log of the likelihood ratio: in (11.3) we may take logs to get

$$p = P(\log \frac{f(X_1, \ldots, X_n|\theta_0)}{f(X_1, \ldots, X_n|\hat{\theta})} < \log LR_{obs}).$$

As when maximizing a likelihood function, taking logs generally simplifies the expression. In addition, the log likelihood ratio is often multiplied by -1 so that larger values produce greater evidence against H_0, i.e., we compute

$$p = P(-\log \frac{f(X_1, \ldots, X_n|\theta_0)}{f(X_1, \ldots, X_n|\hat{\theta})} \geq -\log LR_{obs}). \tag{11.4}$$

Example 1.4 (continued from p. 268) Suppose $X \sim B(n, p)$ and we wish to test $H_0 : p = p_0$. In the case of the data from P.S., we would have $p_0 = .5$ and $\hat{p} = x/n$, with $n = 17$ and $x = 14$. The pdf is

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}$$

and the observed likelihood ratio statistic is

$$\begin{aligned} LR_{obs} &= \frac{p_0^x (1-p_0)^{n-x}}{\hat{p}^x (1-\hat{p})^{n-x}} \\ &= \frac{1}{2^n (\frac{x}{n})^x (1-\frac{x}{n})^{n-x}} \\ &= \frac{1}{2^n (\frac{14}{17})^{14} (1-\frac{14}{17})^3}. \end{aligned}$$

The negative log likelihood ratio becomes

$$\begin{aligned} -\log LR_{obs} &= n \log 2 + x \log \frac{x}{n} + (n-x) \log(1-\frac{x}{n}) \\ &= 17 \log 2 + 14 \log \frac{14}{17} + 3 \log(1-\frac{14}{17}). \end{aligned}$$

□

In Chapter 10 we described several methods of testing H_0 in Example 1.4. The statistic $-\log LR_{obs}$ provides yet another approach. The conclusions reached are consistent with each other and, for sufficiently large samples, the various methods of testing $H_0 : p = .5$ for the binomial parameter will give equivalent results. The advantage of the likelihood ratio test is that it can be generalized and applied in diverse problems. Furthermore, like ML estimation, it turns out to have an important optimality property in large samples.

11.1.2 P-values for the likelihood ratio test of $H_0 : \theta = \theta_0$ may be obtained from the χ^2 distribution or by simulation.

How do we find p-values for the likelihood ratio test? One way is to use the following convenient result.

Result Under certain conditions, for large samples, if θ is m-dimensional then $-2 \log LR$, defined in (11.2), is approximately distributed as χ^2_m, so that an approximation to the p-value in (11.3) may be obtained from the chi-squared distribution with m degrees of freedom.

Example 1.4 (continued) Continuing from the calculation above, we obtain

$$-2 \log LR_{obs} = 2(17 \log 2 + 14 \log \frac{14}{17} + 3 \log(1 - \frac{14}{17})) = 7.72.$$

Here we have $m = 1$ degree of freedom for the chi-squared distribution. Writing $Y \sim \chi^2_1$ we find $P(Y \geq 7.72) = .0055$, i.e., we get $p = .0055$. This is only slightly different than the value $p = .0076$ obtained on p. 257 from the χ^2 statistic. □

We have now used several alternative methods to test H_0 in Example 1.4. The chi-squared statistic and χ^2_1 distribution gave $p = .0076$. The likelihood ratio test and χ^2_1 distribution gave $p = .0055$. The exact calculation on p. 267 gave $p = .013$. The discrepancies among these p-values are not very consequential for conclusions in this case. On the other hand, the numbers are different. This is due to the relatively small sample size. When conclusions depend on which test is used or the method of computing the p-value, the main message should be that the data are not decisive. When one must make a choice as to which p-value to report (in a publication), it is generally preferable to use an exact calculation of the p-value. The computation may be done by simulation. Specifically, under the assumption that H_0 holds, we generate a large number G of data sets and for each compute the test statistic—here, the likelihood ratio statistic—then find the proportion of such simulated test statistic that exceeds to observed value. We illustrate by returning again to the blindsight example.

Example 1.4 (continued) For the responses of patient P.S. it is actually very easy to compute the exact p-value for the likelihood ratio. By symmetry about $p = .5$, it is apparent that $-2\log LR \geq -2\log LR_{obs}$ when $X \leq 3$ or $X \geq 14$. Thus, we would simply find $P(X \leq 3 \text{ or } X \geq 14)$ under the null-hypothetical assumption $X \sim B(17, .5)$. We computed this previously by simulation on p. 267, and we also noted on p. 268 that simulation is unnecessary in this simple example. We found $p = .013$. Let us now write out the steps in the simulation based on the likelihood ratio statistic, because these would be followed in more general contexts.

We use $x[g]$ to denote element g of the vector x and we write the sum of the elements as $sum(x)$, i.e.,

$$sum(x) = \sum_{g=1}^{G} x[g].$$

1. Define a function $LLR(x)$ that evaluates the loglikelihood ratio statistic. Here
$$LLR(x) = 17\log(2) + x\log(\frac{x}{17}) + (17 - x)\log(\frac{17 - x}{17}).$$
2. Evaluate $2LLR_{obs}$ using $LLR_{obs} = LLR(14)$. Here $2LLR_{obs} = 7.72$.
3. Make x a vector of G observations from the null distribution. Here we use $G = 100{,}000$ observations from $B(17, .5)$.
4. If there are possible values of the data that make the loglikelihood ratio become undefined (because the argument of a log would become zero), fix this. Here the log likelihood ratio is not defined when $x = 0$ or $x = 17$ so: if $x[g] = 0$ set $x[g] = 1$; if $x[g] = 17$ set $x[g] = 16$.
5. Set N equal to the number of values g for which $2LLR(x[g]) \geq 2LLR_{obs}$. This may be accomplished by creating a vector y of length G; if $2LLR(x[g]) \geq 2LLR_{obs}$ set $y[g] = 1$; otherwise set $y[g] = 0$; then $N = sum(y)$.
   Here $2LLR_{obs} = 7.72$.

   A detail: The value 7.72 was actually rounded down slightly, so that we are computing $P(X \leq 3 \text{ or } X \geq 14)$ (rather than $P(X < 3 \text{ or } X > 14)$). We would rather compute $p = P(X \leq 3 \text{ or } X \geq 14)$ because it finds the probability of observing a value *at least as large as* LLR_{obs} instead of *larger than* LLR_{obs}, and is therefore more conservative in the sense of producing a larger p-value.
6. Compute $p = \frac{N}{G}$. □

11.1.3 The likelihood ratio test of H_0: $(\omega, \theta) = (\omega, \theta_0)$ plugs in the MLE of ω, obtained under H_0.

We now consider the case in which the parameter vector may be decomposed into two sub-vectors ω and θ, having respective dimensions m_1 and m_2. For example, in linear

regression we would have a parameter vector (β_0, β_1) and we might decompose it as $\omega = \beta_0$ and $\theta = \beta_2$. We consider null hypotheses of the form $H_0 : \theta = \theta_0$ which now becomes a short-hand for $H_0 : (\omega, \theta) = (\omega, \theta_0)$. In linear regression, for example, we might consider whether there is a non-zero slope by introducing $H_0 : \beta_1 = 0$. This is short for $H_0 : (\beta_0, \beta_1) = (\beta_0, 0)$, which means that H_0 does not put any restriction on $\omega = \beta_0$. A wide variety of statistical models that are submodels of larger models may be written in this form. (See for example, Kass and Vos (1997, Theorem 2.3.2).) When we focus on a sub-vector θ of a larger vector (ω, θ) the parameter vector ω is called a *nuisance parameter*.

To apply the likelihood ratio test, we must recognize that ω remains a free parameter under H_0. To evaluate the likelihood ratio we must pick a particular value of ω. We do so by maximizing the likelihood under the null-hypothetical restriction $\theta = \theta_0$. That is, we maximize $L(\omega, \theta_0)$ over ω. Let us denote the solution by $\hat{\omega}_0$. In general $\hat{\omega}_0$ may not equal the global MLE $\hat{\omega}$ (though in some particular cases they will be equal). We thus define the likelihood ratio test statistic as

$$LR_{obs} = \frac{f(x_1, \ldots, x_n|\hat{\omega}_0, \theta_0)}{f(x_1, \ldots, x_n|\hat{\omega}, \hat{\theta})}. \tag{11.5}$$

For a sample $X_1, \ldots, X_n$ with joint pdf $f(x_1, \ldots, x_n|\omega, \theta)$, the theoretical likelihood ratio becomes

$$LR = \frac{f(X_1, \ldots, X_n|\hat{\omega}_0, \theta_0)}{f(X_1, \ldots, X_n|\hat{\omega}, \hat{\theta})} \tag{11.6}$$

and from this we can define the testing procedure.

Likelihood ratio test of $H_0 : (\omega, \theta) = (\omega, \theta_0)$. For a sample $X_1, \ldots, X_n$ with joint pdf $f(x_1, \ldots, x_n|\omega, \theta)$, the likelihood ratio test evaluates LR_{obs} in (11.5) and assigns the p-value

$$p = P(LR < LR_{obs}|H_0) \tag{11.7}$$

where LR is defined in (11.6).

The nuisance parameter ω presents a substantial complication for calculation of an exact p-value by computer simulation. In principle, to compute an explicit p-value, we would not only have to assume $\theta = \theta_0$ (which we do to satisfy H_0) but we would also have to assume some value for ω: to obtain

$$p = P(\frac{f(X_1, \ldots, X_n|\hat{\omega}_0, \theta_0)}{f(X_1, \ldots, X_n|\hat{\omega}, \hat{\theta})} \geq LR_{obs})$$

we must have an explicit probability distribution. Put differently, if we were to use computer simulation to find the exact p-value, we would have to know both the parameters ω, θ in order to do the simulation.

This problem is insoluble without introducing some further restriction or principle.[1] Luckily, there are two good approximate solutions. Here is the first.

Result Under certain conditions, for large samples, if θ is a vector of length m then $-2 \log LR$, defined in (11.6), has an approximate χ^2_m distribution, so that an approximation to the p-value in (11.7) may be obtained from the chi-squared distribution with m degrees of freedom.

The second method is to use $\omega = \hat{\omega}_0$ as a "plug-in" value, under which to compute the p-value by simulation. The procedure is to set $(\omega, \theta) = (\hat{\omega}_0, \theta_0)$, generate many sets of pseudo-data $(X_1^{(g)}, \ldots, X_n^{(g)})$, and then find the proportion of them for which $LR^{(g)} < LR_{obs}$. This constitutes a *parametric bootstrap* likelihood ratio test.

11.1.4 The likelihood ratio test reproduces, exactly or approximately, many commonly-used significance tests.

The likelihood ratio test may be used to derive the t test and other standard tests used in common situations, including the F test in regression (Chapter 12) and analysis of variance (Chapter 13). For testing independence of two traits (as in Section 10.1.4), in large samples the likelihood ratio test is approximately equivalent to the χ^2 test of independence, meaning that in large samples the likelihood ratio test gives very nearly the same p-value as the χ^2 test of independence.

11.1.5 The likelihood ratio test is optimal for simple hypotheses.

Let us consider the simplest form of statistical hypothesis testing where, under both H_0 and H_A there is a distribution that is completely determined, with no free parameters. Specifically, we take H_0: $X \sim f(x)$ and $H_A : X \sim g(x)$ and consider the problem of testing H_0 versus the alternative H_A. This is often called the case of "simple versus simple" hypotheses, because a *simple hypothesis* is one with no free parameters. If T is a test statistic let us write its level and power (defined in Sections 10.4.1 and 10.4.3) as α_T and $1 - \beta_T$.

The likelihood ratio may be written

$$LR_{obs}(x) = \frac{f(x)}{g(x)}$$

[1] One idea is to find the "worst case" p-value (the largest) among all possible values of ω. However, this often remains intractable, except in large samples.

and its theoretical counterpart becomes

$$LR(X) = \frac{f(X)}{g(X)}.$$

Note that the likelihood ratio test will reject H_0 when $LR_{obs}(x)$ is sufficiently small (which is equivalent to $-\log LR(x)$ being sufficiently large). In other words, the likelihood ratio test will reject H_0 when $LR(x) < c$ for some suitable number c. The level is then

$$\alpha_{LR} = P(LR(X) < c|H_0)$$

and the power is

$$1 - \beta_{LR} = P(LR(X) < c|H_A).$$

The Neyman-Pearson Lemma Let α be a positive number less than 1 and let $c = c_\alpha$ be chosen so that

$$\alpha_{LR} = \alpha.$$

Let $T(X)$ be another test statistic having level α_T such that

$$\alpha_T \leq \alpha.$$

Then the power of these two tests satisfies

$$1 - \beta_{LR} \geq 1 - \beta_T.$$

Proof: The argument is very similar to that used in proving the theorem on optimality of Bayes classifiers in Section 4.3.4. □

In words, the Neyman-Pearson lemma says that the likelihood ratio test is the optimal test, in the sense of power, for testing H_0 versus H_A. More generally, likelihood ratio tests may be shown to be optimal for large samples (see Section 5.4.4 of Bickel and Doksum (2001), and Section 16.6 of van der Vaart 1998).

11.1.6 To evaluate alternative non-nested models the likelihood ratio statistic may be adjusted for parameter dimensionality.

The likelihood ratio LR_{obs} in (11.5) compared a statistical model having parameter vector (ω, θ) with a reduced form of the model in which the parameter was (ω, θ_0). In this case, the statistical model based on (ω, θ_0) is said to be *nested* within the larger model based on (ω, θ). For instance, the model

$$Y_i \sim N(\beta_0, \sigma^2),$$

independently, for $i = 1, \ldots, n$ is nested within the simple linear regression model

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2),$$

independently, for $i = 1, \ldots, n$. Note that LR_{obs} satisfies $LR_{obs} \leq 1$: if

$$L(\hat{\omega}, \hat{\theta}) = \max_{(\omega,\theta)} L(\omega, \theta)$$

and

$$L(\hat{\omega}_0, \theta_0) = \max_{\omega} L(\omega, \theta_0),$$

as in (11.5), then, by definition of the maximum, $L(\hat{\omega}, \hat{\theta}) \geq L(\omega, \theta)$ for any other value of (ω, θ), including $(\hat{\omega}_0, \theta_0)$. Therefore, we have

$$L(\hat{\omega}, \hat{\theta}) \geq L(\hat{\omega}_0, \theta_0). \tag{11.8}$$

The likelihood ratio test accounts for this necessity, and judges the degree to which $L(\hat{\omega}, \hat{\theta})$ exceeds $L(\hat{\omega}_0, \theta_0)$ according to (11.7).

When two models are to be compared and neither is a reduced special case of the other the models are called *non-nested*. For non-nested models the likelihood ratio test no longer applies. How should non-nested models be compared? If the two models have the same parameter dimensionality it is possible to compare their maximized loglikelihood functions. However, because of (11.8), when non-nested models of different dimensionality are to be compared, some adjustment for dimensionality of the parameter vectors must be made. The most common methods introduce a criterion that starts with the maximized loglikelihood and then subtracts a penalty for dimensionality. By convention, to match the usual form of the loglikelihood ratio statistic, these criteria are often defined to include a multiplier of -2 so that they may be written as

$$\text{criterion} = -2 \cdot \max \text{ loglikelihood} + \text{penalty}.$$

The most widely used criteria are the *Akaike information criterion*, or AIC (Akaike 1974), and the *Bayesian information criterion*, or BIC (Schwarz 1978), for which the penalties are

$$\text{AIC penalty} = 2m$$

where m is the number of parameters in the model, and

$$\text{BIC penalty} = m \log n,$$

where n is the sample size. Thus, for a random vector X following a model M having an m-dimensional parameter vector θ and pdf $f(x|\theta)$ we have

$$\text{AIC(M)} = -2\log f(x|\hat{\theta}) + 2m$$

and

$$\text{BIC(M)} = -2\log f(x|\hat{\theta}) + m\log n.$$

Many variants on these two model selection criteria have also been proposed; they begin with the same idea, and have more or less the same general form. Note that according to the definition we have just given of AIC(M) and BIC(M), smaller values indicate better models. Alternative equivalent forms, such as that obtained by omitting the multiplier -2 (so that larger values indicate better models), are also used frequently in the literature.

Example 11.1 Interspike interval distribution in resting retinalganglion cells In Section 5.4.6 we introduced the inverse Gaussian distribution as the distribution of interspike intervals for a theoretical integrate-and-fire neuron. Brown et al. (2003), following Iyengar and Liao (1997), analyzed interspike intervals from a resting retinal ganglion neuron recorded *in vitro*, and compared the fits of exponential, gamma, and inverse Gaussian distributions. They obtained AIC $=$ 8,598, 8,567, 8,174 for these three models, respectively, indicating a much better fit for the inverse Gaussian distribution than for either of the other distributions. Plots of fitted pdfs overlaid on the interspike interval histogram were consistent with this evaluation. □

The motivation for AIC begins with the Kullback-Liebler divergence defined on p. 92. Suppose we let $f(x)$ be the true pdf and we wish to obtain a model with pdf $g(x)$ that is a close as possible to $f(x)$ in the sense of minimizing $D_{KL}(f, g)$. When we minimize over $g(x)$ we are maximizing $E_f(\log(g(X)))$. Consider the special case of trying to determine the value of a single scalar parameter θ, where the true value is θ_0, based on data x. Then we are trying to find the closest pdf $g(x|\theta)$ to $f(x) = g(x|\theta_0)$. It is not too hard to show that the expectation $E_f(\log g(X|\theta))$ is maximized by $\theta = \theta_0$. Because θ_0 is unknown we might use the loglikelihood $\log g(x|\theta)$ as an estimate of $E_f(\log g(X|\theta)$, and thus might maximize to get the maximized loglikelihood $\log g(x|\hat{\theta})$. But this is, in general, a biased estimate of $E_f(\log g(X|\theta)$. Akaike proposed to subtract off an estimate of the bias, and then showed that the bias is, in general, approximately equal to the dimensionality of θ. (See Konishi and Kitagawa (2007) for full details.) Multiplying the maximized loglikelihood by -2 gives the form of AIC above.

BIC begins, instead, with the Bayesian formulation of choosing between models M_1 and M_2 based on posterior probability:

$$P(M_1|x) = \frac{f_1(x|M_1)P(M_1)}{f_1(x|M_1)P(M_1) + f_2(x|M_2)P(M_2)} \tag{11.9}$$

where $f_i(x|M_i)$ is the pdf under model M_i and $P(M_i)$ is its prior probability, for $i = 1, 2$. Equation (11.9) follows from an application of Bayes' Theorem, as in (4.38), and also (16.62). To eliminate the prior probabilities one may use the *Bayes factor*, which is the ratio of posterior odds to prior odds:

$$BF_{12} = \frac{P(M_1|x)}{P(M_2|x)} \div \frac{P(M_1)}{P(M_2)} \tag{11.10}$$

(see Section 16.3) and, because

$$\frac{P(M_1|x)}{P(M_2|x)} = \frac{f_1(x|M_1)P(M_1)}{f_2(x|M_2)P(M_2)},$$

we have

$$BF_{12} = \frac{f_1(x|M_1)}{f_2(x|M_2)} \tag{11.11}$$

(a variation on this appears again in Eq. (16.64)). It may be shown that asymptotic approximation of $\log BF$, as $n \to \infty$, leads to the form for BIC given above. Specifically, writing $BIC_{12} = BIC(M_2) - BIC(M_1)$ (so that positive values of BIC_{12} favor model M_1) we have

$$\frac{BIC_{12} - \log BF_{12}}{\log BF_{12}} \xrightarrow{P} 0. \tag{11.12}$$

See Kass and Raftery (1995) and references therein. From a theoretical perspective, BIC is consistent in the sense that, if we assume one of the models is correct (in the sense of generating the data) then, for sufficiently large samples, the probability of BIC choosing the correct model will get arbitrarily close to 1.

In practice, BIC is conservative compared to AIC in that it imposes a larger penalty for dimensionality. Thus, BIC is used, rather than AIC, when there is a strong preference for models of lower dimensionality.

11.2 Permutation and Bootstrap Tests

11.2.1 Permutation tests consider all possible permutations of the data that would be consistent with the null hypothesis.

The idea behind permutation tests is illustrated by a famous example introduced by Fisher in his book *Design of Experiments*. There was, apparently, a lady who claimed to be able to tell the difference between tea with milk added after the tea was poured, and tea with milk added before the tea was poured. Fisher asked how one might test this claim experimentally. His discussion emphasized the importance

of *randomly* allocating the two treatments (milk second versus milk first) to many cups, without the subject's knowledge, and then asking for a judgment on each. (See Section 13.4 for discussion of randomization.) He also considered the question of sample size, and the computation of a p-value. Fisher suggested using eight cups of tea, four of which would have the tea put in first and four of which would have the milk put in first. The lady had to identify tea first or milk first for each of the eight cups. The null hypothesis was that every possible combination of responses would be equally likely, which corresponds to having no ability to tell the difference. There are $\binom{8}{4} = \frac{8!}{4!4!} = 70$ ways to select four tea-first cups from the eight. Therefore, considering all these possible permutations, if the lady were randomly guessing, there would be a $1/70$ chance she would correctly identify all cups of tea as either tea first or milk first. Thus, Fisher pointed out, in the event that she correctly identified milk first or tea first for all eight cups[2] there would be evidence against H_0 with $p = \frac{1}{70} = .014$.

Example 7.2 (continued, see p. 265) We previously applied two-sample t-test to the data displayed in Fig. 7.3 obtained $t_{obs} = -3.19$ on 58 degrees of freedom, giving $p = .0023$. We now apply a permutation test analogous to that for the lady tasting tea.

In this data set there are two groups of 30 subjects. The permutation test considers all of the many ways that 60 subjects, with their learning results, could have been split into two groups of 30 and then asks, out of all those many ways of permuting the subjects, how many of them would have led to results as striking as the one actually observed? The number of ways of splitting 60 individuals into two groups of 30 is

$$\frac{60!}{30!30!} \approx 1.18 \times 10^{17}.$$

In other words, there are 10^{17} different samples of pseudo-data that would be obtained by permuting the group membership among the 60 subject values. The exact two-sample permutation test would, in principle, examine all of these 10^{17} samples and ask how many of them would produce a t-statistic at least as large in magnitude as $t_{obs} = -3.19$. This computation is possible, but it is a bit complicated and we will skip it here. However, a variant on the idea is easy and will lead us naturally to the bootstrap procedure. Instead of examining all 10^{17} permutations, we can *sample* from this distribution. In statistical software there is typically a function that does this sampling by providing random permutations. For example, a sample from the values 1, 2, 3, 4, 5 might be 1, 5, 3, 2, 4, which is a permutation of the original values. To get a relevant random permutation of the data we therefore sample the 60 data values and assign the first 30 values to the first group (SSSS) and the last 30 values to the second group (SSST). We then compute the t-statistic for this permuted data set. If we repeat the procedure a large number of times (say, 10,000 times) we can thereby generate the distribution of the t-statistic under the permutations. □

[2] Fisher also pointed out that with six cups there would be only 20 permutations and thus one would at best obtain $p = .05$; he considered this p-value too large to be useful.

Illustration: Permutation test based on two-sample *t* statistic To be clear about the procedure in Example 7.2, above, let us define the t-statistic as a function of data vectors x and y in several steps. We write the length of a vector x as $length(x)$, the mean of its components as $mean(x)$, the sample variance of its components as $var(x)$, and we make the following definitions:

$$df = length(x) + length(y) - 2$$

$$v_{pooled}(x, y) = \frac{1}{df}\left((length(x) - 1)var(x) + (length(y) - 1)var(y)\right),$$

$$s_{pooled}(x, y) = \sqrt{v_{pooled}(x, y)}$$

and

$$t(x, y) = \frac{mean(x) - mean(y)}{s_{pooled}(x, y)\sqrt{\frac{1}{length(x)} + \frac{1}{length(y)}}}. \tag{11.13}$$

We then use the following algorithm.

1. For $i = 1$ to G:
   Generate $U_1^{(g)}, \ldots, U_{n_1+n_2}^{(g)}$ by permuting the components of the data vector $(x[1], \ldots, x[n_1], y[1], \ldots, y[n_2])$.
   Set $x^{(g)} = (U_1^{(g)}, \ldots, U_{n_1}^{(g)})$ and $y^{(g)} = (U_{n_1+1}^{(g)}, \ldots, U_{n_1+n_2}^{(g)})$.
   Compute $t^{(g)} = t(x^{(g)}, y^{(g)})$.
2. Set N equal to the number of values g for which $|t^{(g)}| \geq |t_{obs}|$.
3. Compute $p = \frac{N}{G}$.

The result is a permutation-based p-value for the t-statistic defined in (10.19). The t-test defined in (10.19) is formulated as a test of $H_0 : \mu_1 = \mu_2$ under normality using (10.21), or via large-sample approximation using (10.20). The permutation test is more general in the sense that the p-value is valid even if the data are not normally distributed, and even if the CLT fails to produce approximately-normal means for the two samples. Furthermore, we may replace the t-statistic based on (10.19), which uses the pooled estimate of variance under the assumption $\sigma_1 = \sigma_2$, with (10.22). In the algorithm above we simply re-define $t(x, y)$ as

$$t(x, y) = \frac{mean(x) - mean(y)}{\sqrt{\frac{var(x)}{length(x)} + \frac{var(y)}{length(y)}}}. \tag{11.14}$$

In either case, for large samples there is generally very little difference between the p-values based on permutations and those based on the t or normal distributions.

The permutation test creates pseudo-data for which the distributions of the two samples are the same; in this sense we may write the null hypothesis as $H_0 : F_X = F_Y$, which is much more restrictive than $H_0 : \mu_1 = \mu_2$ and, therefore, in principle much

easier to reject. However, the t-statistic itself will be strongly sensitive to differences between means, and will tend to be only weakly sensitive to other distinctions between F_X and F_Y, such as differences in the variances. The permutation test based on the t-statistic is therefore generally considered to be a reliable two-sample testing procedure when the main interest is $H_0: \mu_1 = \mu_2$. □

Example 7.2 (continued) Applying the algorithm above with $G = 10{,}000$ using (11.13) we obtained $p = .0019$. Note that here the simulation standard error is $SE = \sqrt{(.0019)(.9981)/10{,}000} = .00044$. Applying the version of the algorithm based on (11.14) we found $p = .0026$. Clearly the conclusions are the same, and they are the same as those based on the ordinary t-test. □

Permutation tests can involve very complicated test procedures. We give an example in Section 11.3.2 on p. 306.

11.2.2 The bootstrap samples with replacement.

Suppose we have a vector x whose components are data values. A permutation of the components of x is a special case of sampling from that data set where (i) the sample size is equal to the length of x and (ii) the sampling is done *without replacement*, meaning that once a data value is selected it can not be selected again. An alternative type of sampling is *with replacement*. In this form, if n is the length of x, then one component of x is drawn at random repeatedly, with all components having equal probabilities of being drawn on all occasions, until a total n numbers are drawn. In this case, there may be repetitions of values. For example, when $x = (1, 2, 3, 4, 5)$ is sampled with replacement we might obtain $3, 4, 1, 4, 2$. Bootstrap tests are essentially the same as permutation tests, except that the sampling is done with replacement.

Illustration: Bootstrap test based on two-sample t statistic Using the same notation as in the illustration of the permutation test on p. 299, the bootstrap test is as follows:

1. For $i = 1$ to G:
   Generate $U_1^{(g)}, \ldots, U_{n_1+n_2}^{(g)}$ by sampling the components of the data vector $(x[1], \ldots, x[n_1], y[1], \ldots, y[n_2])$ with replacement.
   Set $x^{(g)} = (U_1^{(g)}, \ldots, U_{n_1}^{(g)})$ and $y^{(g)} = (U_{n_1+1}^{(g)}, \ldots, U_{n_1+n_2}^{(g)})$.
   Compute $t^{(g)} = t(x^{(g)}, y^{(g)})$.

2. Set N equal to the number of values g for which $|t^{(g)}| \geq t_{obs}$.
3. Compute $p = \frac{N}{G}$.

The only distinction in software implementation (e.g., in Matlab) between the bootstrap and permutation tests would be that the line involving sampling without replacement is changed to sampling with replacement. □

Example 7.2 (continued from p. 300) Applying the bootstrap procedure based on the statistic (11.14) we obtained $p = .0022$. □

11.3 Multiple Tests

11.3.1 When multiple independent data sets are used to test the same hypothesis, the p-values are easily combined.

Sometimes results for each of several subjects, or several experimental units (such as neurons), are equivocal yet all lean in the same direction. Intuitively, such consistency seems to provide additional evidence of a possible effect. Fisher (1925) suggested a simple method of combining multiple independent p-values.

Example 11.2 Precisely repeated intracellular synaptic patterns It has been suggested that precisely timed patterns of synchronous neural activity may propagate across a cortical circuit and, indeed, that such propagation is a crucial mode of information transmission in the brain (see Abeles 2009). Experimental evidence aimed at supporting this idea, which is controversial, was provided by Ikegaya et al. (2004), who recorded spontaneous intracellular activity in vitro from slices of mouse primary visual cortex and in vivo from cat primary visual cortex. Ikegaya et al. (2008) conducted additional experiments and reanalyzed the original data. The in vitro recordings produced relatively long traces of post-synaptic currents which the authors examined for repeated precise patterns. To judge whether observed patterns might be explained by chance, in one of their analyses they performed a kind of permutation test. Because the computations were very time consuming they used only 50 permutations and, when they found their observed test statistic to exceed the values obtained from all 50 sets of pseudo-data they thus achieved statistical significance $p < .02$. This was repeated across 5 neurons. In other words, for each of 5 neurons they achieved $p < .02$, which would seem to be strong statistical evidence that their null hypothesis should be rejected.[3] □

Suppose we have p-values from n independent tests. Fisher observed that under H_0 the p-value for test i would be a uniformly distributed random variable P_i, with $i = 1, \ldots, n$ (see p. 273) and, therefore, the random variable

$$X = -2 \sum_{i=1}^{n} \log P_i \tag{11.15}$$

[3] Some care is required to state correctly the null hypothesis, but roughly speaking it corresponds to time intervals between post-synaptic currents being i.i.d., which they would not be if there were repeated patterns.

would follow the distribution

$$X \sim \chi^2_\nu \tag{11.16}$$

where $\nu = 2n$.

Details: Straightforward calculation using the change of variables formula (the theorem on p. 62) shows that if $W \sim U(0, 1)$ then $-\log W \sim Exp(1)$. It follows that

$$-2 \log W \sim Exp(\frac{1}{2})$$

and the sum of n such independent random variables is distributed as $Gamma(n, \frac{1}{2})$, which is the same as χ^2_ν with $\nu = 2n$. □

Thus, we may combine the observed p-values $p_1, \ldots, p_n$ by writing

$$x_{obs} = -2 \sum_{i=1}^{n} \log p_i \tag{11.17}$$

and then, based on (11.15) and (11.16) we obtain

$$p_{\text{combined}} = P(Y > x_{obs}) \tag{11.18}$$

where $Y \sim \chi^2_\nu$ with $\nu = 2n$.

Example 11.2 (continued) To combine the 5 p-values of .02 we put $p_i = .02$ for $i = 1, 2, 3, 4, 5$, in (11.17) to get

$$x_{obs} = (-2)(5) \log(.02) = 39.$$

From (11.18) we use the χ^2_{10} distribution to obtain

$$p_{\text{combined}} = 2.5 \times 10^{-5}.$$

Because the authors reported $p < .02$ for all five neurons, the combined result is $p < 2.5 \times 10^{-5}$, which is very strong evidence against the null hypothesis. □

11.3.2 When multiple hypotheses are considered, statistical significance should be adjusted.

In Section 10.4 we tried to clarify the interpretation of significance tests. The whole discussion concerned the interpretation of a test of a *single* hypothesis. In many situations, however, multiple hypotheses must be considered within a single analysis.

Example 11.3 Adaptation in fMRI activity among autistic and control subjects Autism is characterized by difficulty in social interaction and communication. One proposal is that autism may involve a defect in the mirror neuron system, which is active in response to observation of activity by other subjects (thus the idea that an individual subject's brain may "mirror" the activity of the other subject). Several studies found the human mirror system to contain subpopulations of neurons that adapt when hand movements are observed or executed repeatedly.[4] Specifically, fMRI responses to observed or executed movements decreased when the movement occurred for a second time. Dinstein et al. (2010) studied brain response adaptation using fMRI, and found that adaptation occurred among autistic subjects as well as controls across multiple regions of interest. The authors considered this to be evidence against mirror system dysfunction in autism.

A crucial step in their argument involved the definition of each region of interest (ROI). For this they combined anatomical and functional characterizations: for each ROI they included every voxel that was both (i) located within 15 mm of an anatomically-defined region and (ii) significantly active based on a t-test of experimental condition versus baseline. Across their ROIs, however, there were thousands of voxels to be examined. In other words, the authors had to perform thousands of tests, of thousands of null hypotheses. This is very common in fMRI studies. □

To see that multiple tests require an additional calculation consider what happens when 100 tests are made. It might be tempting to declare any of the tests significant when $p < .05$. However, if each of the 100 null hypotheses were true, then we would expect about $(.05)(100) = 5$ of the p-values to satisfy $p < .05$, indicating statistical significance. Thus, we would expect several such tests (about 5) to yield spurious (false) results of evidence against the null. An additional calculation makes the situation even more worrisome. Let us suppose that we have 100 random variables T_i representing test statistics for null hypotheses $H_{0,i}$ with[5]

$$P(|T_i| > c_\alpha | H_{0,i}) = \alpha. \tag{11.19}$$

This implies

$$P(|T_i| \leq c_\alpha | H_{0,i}) = 1 - \alpha$$

for $i = 1, 2, \ldots, 100$. If all the tests are independent then we have

$$P(|T_i| \leq c_\alpha \text{ for all } i | H_{0,i} \text{ for all } i) = (1 - \alpha)^{100}$$

and, therefore,

[4] This is important to the logic of the mirror neuron argument. See Dinstein (2008).

[5] We use the absolute value form $|T_i| > c_\alpha$ for consistency with the two-sided tests emphasized in Chapter 10 but the logic is the same for all significance tests.

$$P(|T_i| > c_\alpha \text{ for at least one } i | H_{0,i} \text{for all} i) = 1 - P(|T_i| < c_\alpha \text{ for all } i | H_{0,i} \text{ for all } i)$$
$$= 1 - (1-\alpha)^{100}. \quad (11.20)$$

If we set $\alpha = .05$ we have

$$P(|T_i| > c_\alpha \text{ for at least one } i | H_{0,i} \text{ for all } i) = 1 - .95^{100} = .994.$$

In other words, there is more than a 99 % chance of obtaining at least one spurious result out of 100. Clearly there must be some re-calibration of significance in order to guard against misleading findings.

One way to re-calibrate is to consider the version of (11.20) that applies to n tests,

$$P(|T_i| > c_\alpha \text{ for at least one } i | H_{0,i} \text{ for all } i) = 1 - (1-\alpha)^n \quad (11.21)$$

and change the criterion c_α to some value c such that

$$P(|T_i| > c \text{ for at least one } i | H_{0,i} \text{ for all } i) \leq \alpha. \quad (11.22)$$

In this case we say that the *family-wise error rate* for the collection (family) of n tests is at most α. Let us refer to c_α in (11.19) and (11.21) as the *nominal* criterion for each test. The nominal criterion is the cutoff value we would use for any one test in isolation. We call the criterion c in (11.22) the *family-wise* criterion. There is a very simple way of choosing the family-wise criterion in order to satisfy (11.22).

Bonferroni Correction To test n hypotheses $H_{0,i}$, $i = 1, 2, \ldots, n$ with family-wise error rate at most α, as in (11.22), we may set

$$c = c_{\alpha/n}$$

where $c_{\alpha/n}$ is the nominal criterion for each test.

For example, if we wish to test 5 hypotheses with family-wise error rate $\alpha = .05$ we calculate $.05/5 = .01$ and use the criterion that each of the 5 tests must be significant with $p < .01$. This ensures that we would find at least one spuriously significant test no more than 5 % of the time. In the case of n two-sided t-tests, the Bonferroni correction is to use the criterion $t_\nu(1 - .025/n)$ and declare a particular test significant if $|T_{obs}| > t_\nu(1 - .025/n)$.

The Bonferroni correction is justified by the following inequality. Let A_i represent the event that the ith test is declared significant, where $i = 1, 2, \ldots, n$. If we examine 3 tests, then $n = 3$ and $P(A_1 \cup A_2 \cup A_3)$ is the probability that at least one of the tests is significant. For n tests $P(A_1 \cup A_2 \cup \cdots \cup A_n)$ is the probability that at least one test is significant.

Theorem: Bonferroni inequality For events $A_1, A_2, \ldots, A_n$ we have that

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) \leq P(A_1) + P(A_2) + \cdots + P(A_n). \tag{11.23}$$

Proof: Recall that for two events A and B we have

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{11.24}$$

This implies

$$P(A \cup B) \leq P(A) + P(B). \tag{11.25}$$

Now consider three events C, D, E. Applying the formula (11.24) with $A = C \cup D$ and $B = E$ we get

$$P(C \cup D \cup E) = P(C \cup D) + P(E) - P((C \cup D) \cap E)$$

and applying (11.24) to the right-hand side with $A = C$ and $B = D$ we obtain

$$P(C \cup D \cup E) = P(C) + P(D) - P(C \cap D) + P(E) - P((C \cup D) \cap E)$$

which gives

$$P(C \cup D \cup E) \leq P(C) + P(D) + P(E). \tag{11.26}$$

The inequalities (11.25) and (11.26) are examples of the Bonferroni inequality. We can continue the same argument to obtain (11.23). □

The Bonferroni correction is easy to apply, but it is usually quite conservative in the sense that it tends to produce relatively few statistically significant tests. This has led to development of many other ways to control the family-wise error rate, especially in the context of analysis of variance, which we comment on in Section 13.1.7. A different idea is to try to control the *proportion* of spuriously significant results, which is known as the*False Discovery Rate (FDR)*,

$$\text{FDR} = \frac{\text{number of spuriously significant tests}}{\text{total number of significant tests}}. \tag{11.27}$$

Here, the spuriously significant tests represent "false discoveries." In practice one does not know whether a particular H_0 is true or false, so one also does not know whether a particular statistically significant test is a false discovery (because its H_0 is true) or a true discovery (because its H_0 is false). Therefore, the numerator and denominator in (11.27) are not known. However, under certain general conditions it turns out to be possible to control the *expected* false discovery rate. We will use the letter q to represent the desired false discovery rate, such as $q = .05$.

FDR algorithm

1. Perform n tests using statistics T_i, for $i = 1, \ldots, n$, and obtain n p-values.
2. Put the p-values in ascending order $p_{(1)}, p_{(2)}, \ldots, p_{(n)}$ (so $p_{(1)}$ is the smallest p-value) and let $T_{(j)}$ be the test having p-value $p_{(j)}$.
3. Let r be the largest value of j such that

$$p_{(j)} \leq \frac{jq}{n}.$$

4. Consider the tests $T_{(1)}, T_{(2)}, \ldots, T_{(r)}$ to be significant with expected false discovery rate less than q.

The FDR procedure is justified by the following inequality (see Benjamini and Yekutieli 2001 and Genovese et al. 2002).

FDR inequality Under certain conditions, when tests are declared significant using the FDR algorithm we have

$$E(FDR) \leq q.$$

Example 11.3 (continued) To define their regions of interest, Dinstein et al. (2010) had to select functionally active voxels based on thousands of t tests. For this purpose they used FDR, setting the rate at $q = .05$. □

Yet another strategy for grappling with multiple hypotheses is available in some repeated-trial contexts. It is illustrated in Example 4.7.

Example 4.7 (continued from p. 100) Figure 4.4 displayed decoding accuracy based on MEG sensor recordings in an experiment on overt and imagined wrist movement. In that work, and in MEG studies generally, it is also of interest to find the brain source locations of such sensor observations. This is called the *source localization* problem (see Example 12.9). One issue is that large numbers of possible sources, typically thousands, are examined and there is the potential for false discoveries. Xu et al. (2011) described a method of finding regions of brain activity following the application of a standard source localization algorithm, and they applied a permutation test to guard against spurious results. In their scheme the sensor data from a single subject formed a 3-way array with dimensions $R \times M \times T$, where R was the number of repeated trials, M was the number of sensor signals, and T was the number of time points. A source localization algorithm produced an $N \times T$ array of source signals, where N was the number of sources. They then defined a collection of $N \times T$ likelihood ratio statistics aimed at identifying sources that contained directional hand movement information; these likelihood ratio statistics were thresholded and clustered into spatio-temporal regions that could represent important sources of activity. The finished product was nine spatial-temporal regions having directional hand movement information from a single subject. This was a complicated procedure

involving several distinct algorithms. To determine a p-value for the set of regions Xu et al. performed 100,000 permutations of the trials[6] and for each resulting set of pseudo-data they ran the *the entire procedure*. They then asked how many results based on pseudo-data were as extreme as those obtained from the data. This allowed them to report $p < 10^{-5}$ for the set of activity regions obtained from the data, which is very strong evidence that the activity regions were real as opposed to representing statistically spurious[7] results. The key idea here is that a p-value may be obtained for a procedure that searches across many spatial-temporal locations, corresponding to many null hypotheses of no directionally-related activity, by evaluating the procedure on each set of pseudo-data generated by a permutation test. □

There is a large literature on testing multiple hypotheses. See, for instance, Gordon et al. (2007) and the references therein.

[6] The permutations were done in source space; see Xu et al. (2011).

[7] The null hypothesis was that for every brain source the theoretical mean activities in all movement directions were equal.

Chapter 12
Linear Regression

Regression is the central method in the analysis of neural data. This is partly because, in all its guises, it is the most widely applied technique. But it also played a crucial historical role[1] in the development of statistical thinking, and continues to form a core conceptual foundation for a great deal of statistical analysis. We introduced linear regression in Section 1.2.1 (on p. 10) by placing it in the context of curve-fitting, reviewing the method of least squares, and providing an explicit statement of the linear regression model. This enabled us to use linear regression as a concrete example of a statistical model, so that we could emphasize a few general points, including the role of models in expressing knowledge and uncertainty via inductive reasoning. The linear regression model is important not only because many noisy relationships are adequately described as linear, but also—as we tried to explain in Section 1.2.1—because the framework gives us a way of thinking about relationships between measured variables. For this reason, we began with the more general regression model in Eq. (1.2), i.e.,

$$Y_i = f(x_i) + \epsilon_i, \tag{12.1}$$

and only later, in Eq. (1.4), specified that $f(x)$ is taken to be linear, i.e.,

$$f(x) = \beta_0 + \beta_1 x. \tag{12.2}$$

Equation (1.2), repeated here as (12.1), gave substance to the diagram in Eq. (1.1), i.e.,

$$Y \longleftarrow X. \tag{12.3}$$

To incorporate multiple explanatory variables we replace $f(x)$ in (12.1) with $f(x_1, \ldots, x_p)$, and to extend beyond the additive form of noise in (12.1) we replace the diagram in (12.3) with

[1] See the appendix of Brown and Kass (2009).

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_12,

$$Y \longleftarrow \begin{cases} \text{noise} \\ f(x_1, \ldots, x_p). \end{cases} \tag{12.4}$$

This diagram is supposed to indicate a variety of generalizations of linear regression which, together, form the class of methods known as *modern regression*.

In this chapter we provide a concise introduction to linear regression. In Sections 12.1–12.4 we treat the *simple linear regression model* given by

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{12.5}$$

for $i = 1, \ldots, n$, where ϵ_i is a random variable. The adjective "simple" refers to the single x variable on the right-hand side of (12.5). When there are two or more x variables on the right-hand side the terminology *multiple regression* is used instead. We go over some of the most fundamental aspects of multiple regression in Section 12.5. That section also lays the groundwork for modern regression. Generalizations are described in Chapters 14 and 15.

12.1 The Linear Regression Model

To help fix ideas, as we proceed we will refer to several examples.

Example 12.1 Neural correlates of reward in parietal cortex Platt and Glimcher (1999) suggested that cortical areas involved in sensory-motor processing may encode not only features of sensation and action but also key inputs to decision making. To support their claim they recorded neurons from the lateral intraparietal (LIP) region of monkeys during an eye movement task, and used linear regression to summarize the increasing trend in firing rate of intraparietal neurons with increasing expected gain in reward (volume of juice received) for successful completion of a task. Figure 12.1 shows plots of firing rate versus reward volume for a particular LIP neuron following onset of a visual cue. □

Example 2.1 (continued from p. 24) In their analysis of saccadic reaction time in hemispatial neglect, Behrmann et al. (2002) used linear regression in examining the modulation of saccadic reaction time as a function of angle to target by eye, head, or trunk orientation. We refer to this study in Section 12.5. □

In Chapter 1 we used Example 1.5 on neural conduction velocity to illustrate linear regression. Another plot of the neural conduction velocity data is provided again in Fig. 12.2.

Before we begin our discussion of statistical inference in linear regression, let us recall some of the things we said in Chapter 1 and provide a few basic formulas.

Given n data pairs (x_i, y_i), least squares finds $\hat{\beta}_0$ and $\hat{\beta}_1$ that satisfy

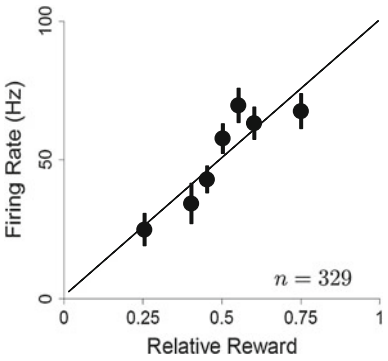

Fig. 12.1 Plots of firing rate (in spikes per second) versus reward volume (as fraction of the maximal possible reward volume). The plot represents firing rates during 200 ms following onset of a visual cue across 329 trials recorded from an LIP neuron. The 329 pairs of values have been reduced to 7 pairs, corresponding to seven distinct levels of the reward volume. Each of the 7 y_i values in the figure is a mean (among the trials with x_i as the reward volume), and error bars representing standard errors of each mean are also visible. A least-squares regression line is overlaid on the plot. Adapted from Platt and Glimcher (1999).

$$\sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2 = \min_{\beta_0^*, \beta_1^*} \sum_{i=1}^{n} \left(y_i - (\beta_0^* + \beta_1^* x_i)\right)^2 \tag{12.6}$$

where we use β_0^* and β_1^* as generic possible estimates of β_0 and β_1. The least-squares estimates (obtained by calculus) are

$$\hat{\beta}_1 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \tag{12.7}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{12.8}$$

The resulting fitted line

$$y = \hat{\beta}_0 + \hat{\beta}_1 x \tag{12.9}$$

is the *linear regression* line (and often "linear" is dropped).

Details: To be clear what we mean when we say that the least-squares estimates may be found by calculus, let us write

$$g(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 .$$

The formulas (12.8) and (12.7) may be obtained by computing the partial derivatives of $g(\beta_0, \beta_1)$ and then solving the equations

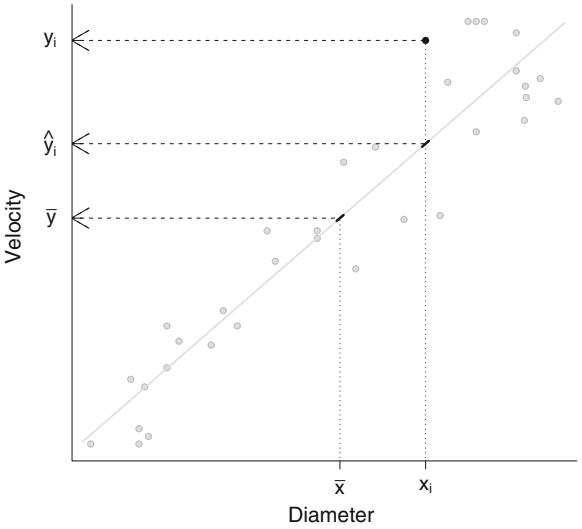

Fig. 12.2 Plot of the Hursh conduction velocity data set, for $5 < x < 15$, with data points in *gray* except for a particular point (x_i, y_i) which is shown in *black* to identify the corresponding fitted value $\hat{y}_i$. The ith residual is $y_i - \hat{y}_i$. The regression line also passes through the point $(\bar{x}, \bar{y})$, as indicated on the plot.

$$0 = \frac{\partial g}{\partial \beta_0}$$
$$0 = \frac{\partial g}{\partial \beta_1}.$$

□

The least-squares fitted values at each x_i are

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{12.10}$$

and the least-squares residuals are

$$e_i = y_i - \hat{y}_i. \tag{12.11}$$

See Fig. 12.2. If we plug (12.8) into (12.9) we get

$$y - \bar{y} = \hat{\beta}_1 (x - \bar{x}) \tag{12.12}$$

which shows that the regression line passes through the point $(\bar{x}, \bar{y})$, as may be seen in Fig. 12.2. Also, when we plug into (12.12) the (x, y) value (x_i, y_i) we get

$$y_i - \bar{y} = \hat{\beta}_1(x_i - \bar{x})$$

or

$$y_i = \bar{y} - \hat{\beta}_1(x_i - \bar{x}). \tag{12.13}$$

A few more lines of algebra show that using (12.13) in (12.11) gives

$$\sum_{i=1}^{n} e_i = 0, \tag{12.14}$$

which is useful as a math fact, and also can be important to keep in mind in data analysis: linear least squares residuals fail to satisfy (12.14) only when a numerical error has occurred.

Details: We have

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \bar{y} + \bar{y} - \hat{y}_i). \tag{12.15}$$

Because $\sum y_i = n\bar{y}$ we have

$$\sum_{i=1}^{n} (y_i - \bar{y}) = 0 \tag{12.16}$$

and, similarly,

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0. \tag{12.17}$$

Combining (12.13) with (12.17) gives

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y}) = 0. \tag{12.18}$$

Finally, using (12.16) with (12.18) in (12.15) gives (12.14). □

It is worth drawing attention to one other interesting feature of the linear regression model. While (12.1) and (12.4) emphasize potential nonlinearity in the way a variable x, or multiple variables $x_1, \ldots, x_p$ may influence y, it turns out that linear regression may be used to fit some nonlinear relationships. This is discussed in Section 12.5.4. Here is a particularly simple, yet important additional example.

Example 12.2 BOLD hemodynamic response in fMRI In Fig. 1.3 of Example 1.3 we displayed fMRI images from a single subject during a simple finger-tapping task in response to a visual stimulus. As we said there, fMRI detects changes in

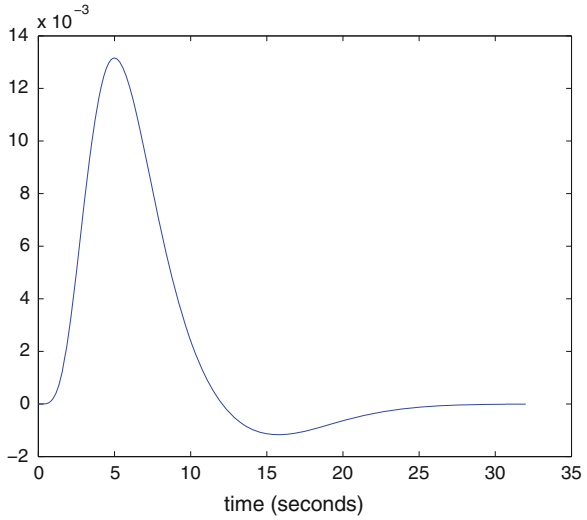

Fig. 12.3 The hemodynamic response function defined by Eq. (12.19).

blood oxygenation and the measurement is known as the BOLD signal, for Blood Oxygen-Level Dependent signal. The typical hemodynamic response that produces the signal is relatively slow, lasting roughly 20 s (seconds). Many experiments have shown, however, that it has a reasonably stable form (see Glover 1999). Software for analyzing fMRI data, such as BrainVoyager (see Goebel et al. 2006; Formisano et al. 2006), often uses a particular hemodynamic function. Figure 12.3 displays a plot of such a theoretical hemodynamic response function $h(t)$ defined by

$$h(t) = \left(\frac{t}{d_1}\right)^{a_1} \exp(-\frac{t-d_1}{b_1}) - c\left(\frac{t}{d_2}\right)^{a_2} \exp(-\frac{t-d_2}{b_2}) \tag{12.19}$$

where $a_1, b_1, d_1, a_2, b_2, d_2$ and c are parameters that have default values in the software. Using this function the fMRI data at a particular voxel (a particular small rectangular box in the brain) may be analyzed using linear regression. Let us suppose we have an on/off stimulus, as is often the case, and let $u_j = 1$ when the stimulus is on and 0 otherwise, $j = 1, \ldots, T$. The effect at time i of the stimulus being on at time j is assumed to follow the hemodynamic response function, i.e., the effect is determined by $h(t)$ where $t = i - j$ is the delay between the stimulus and the response time i. It is also assumed that the effects of multiple "on" stimuli at different times j produce additive effects at different time lags $i - j$. Therefore, the total stimulus effect at time i is[2]

$$x_i = \sum_{j<i} h(i-j)u_j. \tag{12.20}$$

[2] This expression is known as the *convolution* of the hemodynamic response function $h(t)$ with the stimulus function u_j.

The linear regression model (12.5) may then be fitted, and the coefficient β_1 represents the overall magnitude of the increased BOLD response due to the activity associated with the stimulus. □

12.1.1 *Linear regression assumes linearity of $f(x)$ and independence of the noise contributions at the various observed x values.*

The model (12.1) is *additive* in the sense that it assumes the noise, represented by ϵ_i is added to the function value $f(x_i)$ to get Y_i. This entails a *theoretical* relationship between x and y that holds except for the "errors" ϵ_i. Linear regression further specializes by taking $f(x)$ to be linear as in (12.2) so that we get the model (12.5). The ϵ_i's are assumed to satisfy

$$E(\epsilon_i) = 0$$

for all i, so that $E(Y_i) = \beta_0 + \beta_1 x_i$. In words, the linear relationship $y = \beta_0 + \beta_1 x$ is assumed to hold "on average," that is, apart from errors that are on average zero. Additivity of the errors and linearity of $E(Y_i)$ are the most fundamental assumptions of linear regression. In addition, the errors ϵ_i are assumed to be independent of each other. In Section 12.2.3 we show how lack of independence can distort statistical inferences about the regression model. The independence assumption may be violated when observations are recorded sequentially across time, in which case more elaborate *time series* methods are needed. These are discussed in Chapter 18.

Important, though less potentially problematic, additional assumptions are that the variances of the ϵ_i's are all equal, so that the variability of the errors does not change with the value of x, and that the errors are normally distributed. These latter two assumptions guarantee that the 95 % confidence intervals discussed in Section 12.3.1 have the correct probability .95 of covering the coefficients and the significance tests in Section 12.3.2 have the correct p-values. In sufficiently large samples the normality assumption becomes unnecessary, as the confidence intervals and significance tests will be valid, approximately (see (12.37)).

To summarize, the assumptions of linear regression may be enumerated, in order of importance, as follows:

(i) the linear regression model (12.5) holds;
(ii) the errors satisfy $E(\epsilon_i) = 0$ for all i;
(iii) the errors ϵ_i are independent of each other;
(iv) $V(\epsilon_i) = \sigma^2$ for all i (homogeneity of error variances), and
(v) $\epsilon_i \sim N(0, \sigma^2)$ (normality of the errors).

To repeat, the crucial assumptions are the first three: there is, on average, a linear relationship between Y and x, and the deviations from it are represented by independent errors.

12.1.2 The relative contribution of the linear signal to the total response variation is summarized by R^2.

As shown in Fig. 12.2, in Example 1.5 linear regression provides a very good representation of the relationship between x and y, with the points clustering tightly around the line. In other cases there is much more "noise" relative to "signal," meaning that the (x_i, y_i) values scatter more widely, so that the residuals tend to be much larger. In this section we describe two measures of residual deviation.

The error standard deviation σ (see item (iv) in the assumptions in Section 12.1.1) represents the average amount of deviation of each ϵ_i from zero. Thus, σ tells us how far off, on average, we would expect the line to be in predicting a value of y at any given x_i. It is estimated by $s = \sqrt{s^2}$ where

$$s^2 = \frac{1}{n-2} SSE \tag{12.21}$$

and

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{12.22}$$

is the *sum of squares for error* or the *residual sum of squares*. (Here $\hat{y}_i$ is defined by (12.10).) The variance estimate s^2 is then also called the *residual mean squared error* and we often write

$$MSE = s^2. \tag{12.23}$$

This definition of s makes it essentially the standard deviation of the residuals, except that $n-2$ is used in the denominator instead of $n-1$; here there are two parameters β_0 and β_1 being estimated so that two degrees of freedom are lost from n, rather than only one.

The other quantity, R^2, is interpreted as the fraction of the variability in Y that is attributable to the regression, as opposed to error. We begin by defining the *total sum of squares*

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2. \tag{12.24}$$

This represents the overall variability among the y_i values. We then define

$$R^2 = 1 - \frac{SSE}{SST}. \tag{12.25}$$

The fraction SSE/SST is the proportion of the variability in Y that is attributable to error, and R^2 is what's left over, which is attributable to the regression line. The value of R^2 is between 0 and 1. It is 0 when there is no linear relationship and 1 when there is a perfect linear relationship. If we define the *sum of squares due to regression* as the difference

$$SSR = SST - SSE \tag{12.26}$$

then we can re-write R^2 in the form

$$R^2 = \frac{SSR}{SST}. \tag{12.27}$$

From this version we get the interpretation of R^2 as "the proportion of variability of Y that is explained by X." In different terminology, we may think of SSR as the *signal* variability (often called "the variability due to regression") and SSE as the *noise* variability. Then $R^2 = SSR/(SSR + SSE)$ becomes the relative proportion of signal-to-noise variability. (The ratio of signal-to-noise variabilities[3] would be SSR/SSE.)

In (12.26) we defined the sum of squares due to regression by subtraction. There is a different way to define it, so that we may see how total variability (SST) is decomposed into regression (SSR) and error components (SSE). The derivation begins with the values $y_i, \hat{y}_i$, and $\bar{y}$, as shown in Fig. 12.2, where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Writing $y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y}$, we have

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

but after plugging in the definition of $\hat{y}_i$ from (12.10) some algebra shows that the cross-product term vanishes and, defining

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2, \tag{12.28}$$

we have

$$SST = SSR + SSE. \tag{12.29}$$

As we mention again in Section 12.5.3, the vanishing of the cross-product may be considered, geometrically, to be a consequence of the Pythagorean theorem. Equation (12.29) is important in understanding linear regression and analysis of variance: we think of the total variation as coming from different additive components, whose magnitudes we compare.

The estimated standard deviation s has the units of Y and is therefore interpretable —at least to the extent that the Y measurements themselves are interpretable. But R^2 is dimensionless. Unfortunately, there are no universal rules of thumb as to what constitutes a large value: in some applications one expects an R^2 of at least .99 while

[3] The signal-to-noise ratio is a term borrowed from engineering, where it refers to a ratio of the power for signal to the power for noise, and is usually reported in the log scale; under certain stochastic models it translates into a ratio of signal variance to noise variance.

in other applications an R^2 of .40 or less would be considered substantial. One gets a feeling for the size of R^2 mainly by examining, and thinking about, many specific examples.

12.1.3 *Theory shows that if the model were correct then the least-squares estimate would be likely to be accurate for large samples.*

In presenting the assumptions on p. 315 we noted that they were listed in order of importance and, in particular, normality of the errors is not essential. The following theoretical result substantiates the validity of least-squares for non-normal errors in large samples.

Theorem: Consistency of least squares estimators For the linear regression model (12.5) suppose conditions (i)–(iv) hold and let $x_1, x_2, \ldots, x_n, \ldots$ be a sequence of x values such that

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \to \infty \tag{12.30}$$

as $n \to \infty$. Then the least-squares estimator defined by (12.7) satisfies

$$\begin{aligned} \hat{\beta}_1 &\xrightarrow{P} \beta_1 \\ \hat{\beta}_0 &\xrightarrow{P} \beta_0. \end{aligned} \tag{12.31}$$

In other words, under these conditions $\hat{\beta}_1$ and $\hat{\beta}_0$ are consistent estimators of β_1 and β_0.

Proof: This is essentially a consequence of the law of large numbers in a non-i.i.d. setting, where linear combinations of the Y_i values are being used according to (12.7) and (12.8). We omit the proof and refer the interested reader to Wu (1981), which examines a more general problem but provides extensive references and discussion. □

Note that to fit a line we must have at least 2 distinct values, so that not every observation can be made at the same x value. The condition (12.30) fails when, for all sufficiently large i and j, $x_i = x_j$. In other words, it rules out degenerate cases where essentially all the observations (i.e., all but finitely many of them) are made at a single x value.[4] We may interpret this asymptotic statement as saying that for all situations in which there is any hope of fitting a line to the data, as the sample size increases the least-squares estimator of the slope will converge to the true value.

[4] In fact, the results cited in Wu (1981) show that (12.30) is necessary and sufficient for (12.31).

12.2 Checking Assumptions

12.2.1 Residuals should represent unstructured noise.

In examining single batches of data, in Chapter 2, we have seen how the data may be used not only to estimate unknown quantities (there, an unknown mean μ) but also to check assumptions (in particular, the assumption of normality). This is even more important in regression analysis and is accomplished by analyzing the residuals defined in (12.11). Sometimes the residuals are replaced by *standardized residuals*. The ith standardized residual is $e_i/SD(e_i)$, where $SD(e_i)$ is the standard deviation of e_i. Dividing by the standard deviation puts the residuals on a familiar scale: since they are supposed to be normal, about 5 % of the standardized residuals should be either larger than 2 or smaller than -2. Standardized residuals that are a lot larger than 2 in magnitude might be considered outliers.

> *A detail:* There are two different ways to standardize the residuals. We have here taken $SD(e_i)$ to be the estimated standard deviation of e_i. The formula for $SD(e_i)$ involves the x_i values. An alternative would be to compute the sample variance of the residuals
>
> $$s_e^2 = \frac{1}{n-1}\sum(e_i - \bar{e})^2$$
>
> and take its square root. The standardization using $SD(e_i)$, which allows the n residual standard deviations to be different, is often called *studentization* (by analogy with the ratio that defines Student's t distribution, see p. 129). The statistical software packages we are most familiar with use $SD(e_i)$ to standardize the residuals. □

Two kinds of plots are used. Residual versus fit plots are supposed to reveal (i) nonlinearity, (ii) inhomogeneity variances, or (iii) outliers. Plots having structure of the kind that would indicate these problems are shown in Fig. 12.4. The first plot is typical of data with no systematic variation remaining after linear regression: the pattern is "random," specifically, it is consistent with errors that are independent and normally distributed, all having the same distribution. The second plot shows departure from linearity; the third indicates more variability for large fitted values than for smaller ones. The last plot has an outlier, indicating a point that is way off the fitted line.

Histograms and Q-Q plots of the residuals are also used to assess assumptions. These are supposed to (i) reveal outliers and (ii) check whether the errors may be described, at least approximately, by a normal distribution.

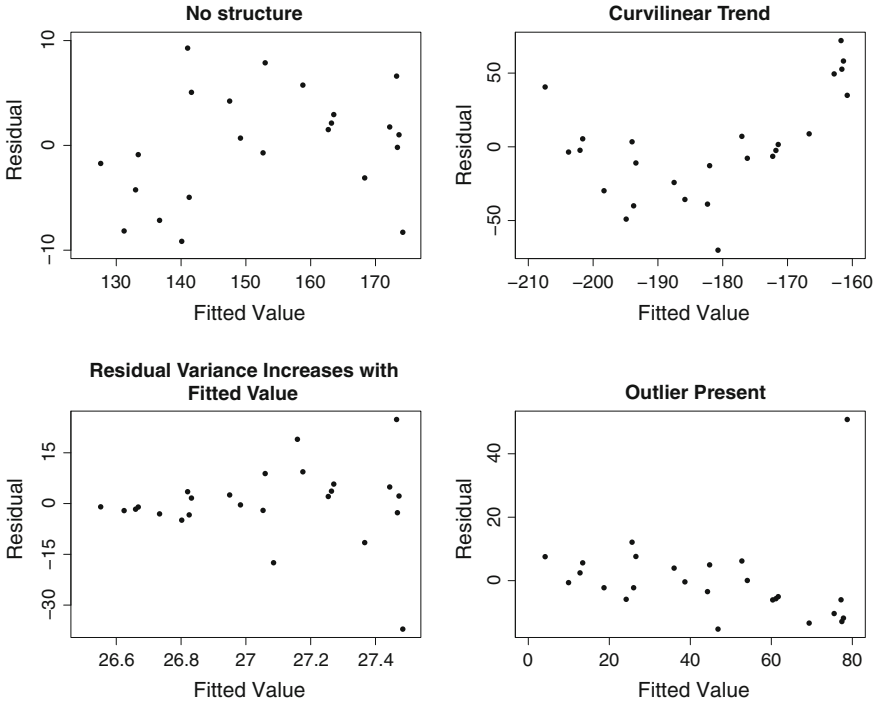

Fig. 12.4 Residual plots: the *Top Left* plot depicts unstructured noise while the latter three reveal structure, and thus deviations from the assumptions.

12.2.2 Graphical examination of (x, y) data can yield crucial information.

As we tried to emphasize in Chapters 1 and 2, it is important to examine data with exploratory methods, using visual summaries where possible. The following illustration gives a nice demonstration of how things can go wrong if one relies solely on the simplest numerical summaries of least-squares regression.

Illustration Figure 12.5 shows a striking example in which four sets of data all have the same regression equation and R^2, but only in the first case (data set 1) would the regression line appropriately summarize the relationship. In the second case (data set 2) the relationship is clearly nonlinear, in the third case there is a big outlier and removing it dramatically changes the regression. In the fourth case the slope of the line is determined entirely by the height of the point to the right of the graph; therefore, since each point is subject to some random fluctuation, one would have to be very cautious in drawing conclusions. □

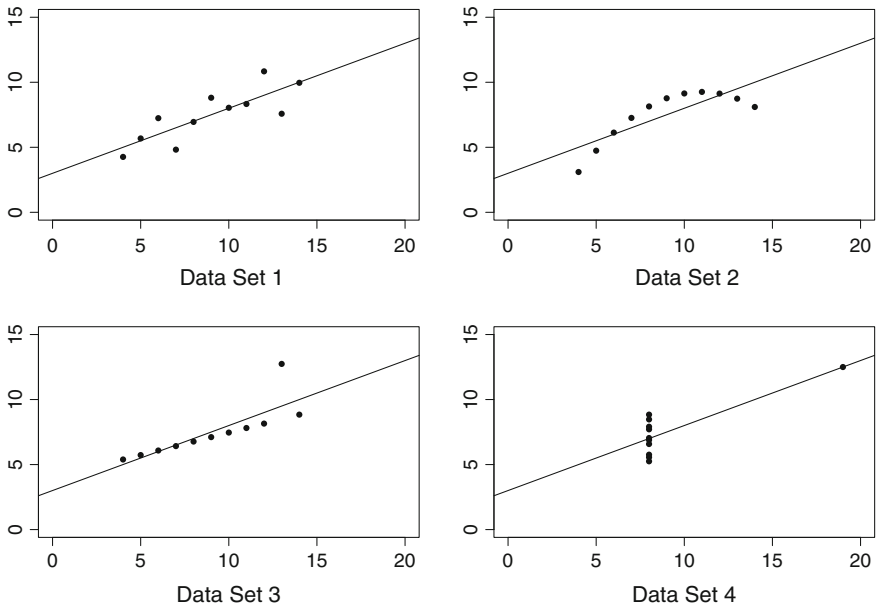

Fig. 12.5 Plots of four very different data sets all having the same fitted regression equation $Y = 3 + .5x$ and $R^2 = .667$. These were discussed in Anscombe (1973).

This illustration underscores the value of plotting the data when examining linear or curvilinear relationships.

12.2.3 Failure of independence among the errors can have substantial consequences.

In stating the assumptions of linear regression on p. 315 we stressed the importance of independence among the errors ϵ_i. To be concrete, we now consider how inference about the strength of the linear relationship between $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, as measured by R^2, can be badly misled when the data are correlated. To do this we use a simple model of serial dependence: we put

$$U_t = \rho U_{t-1} + \delta_t \tag{12.32}$$

$$W_t = \rho W_{t-1} + \eta_t \tag{12.33}$$

for $t = 2, 3, \ldots, n$ where

$$\delta_t \sim N(0, 1)$$
$$\eta_t \sim N(0, 1)$$
$$U_1 \sim N(0, 1)$$
$$W_1 \sim N(0, 1)$$

all independently of each other. Models (12.32) and (12.33) are both examples of first-order *autoregressive models*, which we discuss further in Chapter 18, with *autocorrelation coefficient* ρ. According to these models the observations U_t and W_t are likely to be close to the respective values U_{t-1} and W_{t-1}, but with noise added. The variation in experimental data observed across time may often be described well using autoregressive models. Note that U_t and W_t are independent for all t. We simulate values $u_1, \ldots, u_n$ and $w_1, \ldots, w_n$ from (12.32) and (12.33), using $n = 100$, and we then define

$$x_i = u_i$$
$$y_i = w_i$$

for $i = 1, \ldots, n$ and compute R^2 from the regression of y on x. We could say that the correct linear model in this case is

$$Y_i = \epsilon_i$$

where ϵ_i follows the autoregressive model (12.33), so that in principle we should find $R^2 = 0$. Figure 12.6 gives the results from 100 simulations (each using $n = 100$). When the autocorrelation coefficient is zero, we get values of R^2 that deviate from 0 according to the null distribution so that about 5 % of the values are above the threshold corresponding to $p < .05$ and about 1 % of the values are above the threshold corresponding to $p < .01$. However, as the magnitude of the autocorrelation coefficient increases we find many values of R^2 that are substantial, many more than would be predicted by the null distribution—thus, the p-values are no longer accurate. In fact, for magnitudes of the autocorrelation that are close to 1 it becomes highly probable to get what would look like a "significant" correlation in the data, even though the x and y data were computer-generated to be independent.

This phenomenon may be appreciated further by contrasting the variation in independent normal data with data generated from model (12.32) with $\rho = .9$. As seen in the right-hand side of Fig. 12.7, data following this autoregressive model tend to have patches of values that are all either above 0 or below 0. If we imagine two such series, there are likely be patches of time where both series are very different from 0 and this will often lead to a substantial magnitude of the correlation coefficient computed across time.

The point is that one must be very careful about the assumption of independence in linear regression. When regression or correlation analysis is to be performed on data recorded across time, where dependence among errors is likely, the standard advice is to first *pre-whiten* the data by removing temporal structure (for instance, by fitting auto-regressive models and then analyzing the residuals) as discussed in Section 18.5.2.

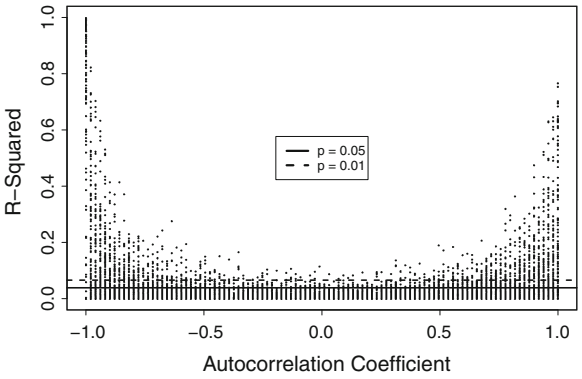

Fig. 12.6 Values of R^2 based on truly independent Y and X data that were simulated using (12.32) and (12.33), with $n = 100$. The x-axis of the plot gives the value of the autocorrelation coefficient ρ. The usual p-values, obtained from applying the t-distribution to (12.38), accurately represent the probability of deviation as large as the observed R^2 only when $\rho = 0$.

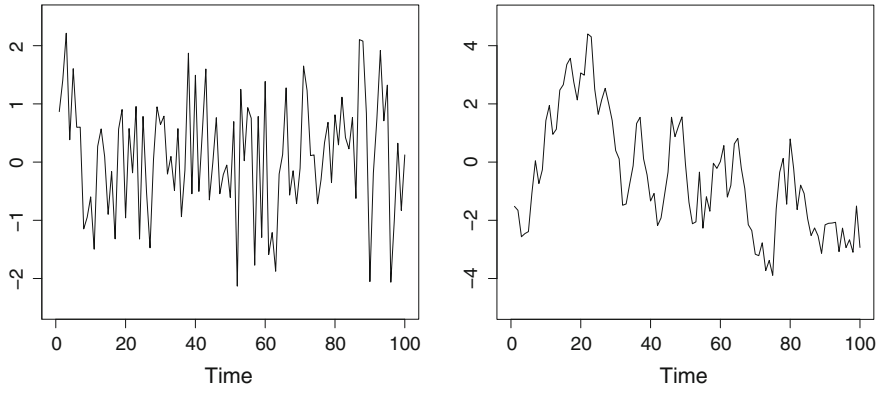

Fig. 12.7 Plots of artificial data against a variable representing time, which takes on values $1, 2, \ldots, 100$. The data values have been connected with lines. *Left* plot of 100 independent $N(0, 1)$ random values. *Right* plot of 100 values from an autoregressive model, as in (12.32) with $\rho = .9$. The independent values fluctuate without trends, while the autoregressive values show excursions of several successive values that are consistently positive or negative.

12.3 Evidence of a Linear Trend

12.3.1 Confidence intervals for slopes are based on SE, according to the general formula.

When reporting least-squares estimates, standard errors should also be supplied. That is, one reports either $\hat{\beta}_1 \pm SE(\hat{\beta}_1)$ or a confidence interval. Standard errors are given as standard output from regression software. The general formula for standard errors

in linear regression appears in Eq. (12.61). To get an approximate 95 % confidence interval for β_1 based on $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$, we again use the general form given by (7.8), i.e.,

$$\text{approx. } 95\,\%\,\text{CI} = (\hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot SE(\hat{\beta}_1)). \tag{12.34}$$

An alternative, in small samples, is analogous to the small sample procedure in (7.31) used to estimate a population mean: we substitute for 2 the value $t_{.975,\nu}$, where now $\nu = n - 2$ because we have estimated two parameters (intercept and slope) and thus have lost two degrees of freedom. Thus, we would use the formula

$$95\,\%\,\text{CI} = (\hat{\beta}_1 - t_{.025,n-2} \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + t_{.025,n-2} \cdot SE(\hat{\beta}_1)). \tag{12.35}$$

Example 1.5 (continued, see p. 11) Using least squares regression we found $\hat{\beta}_1 = 6.07$ and $SE(\hat{\beta}_1) = .14$. We would report this by saying that, on average, action potential velocity increases by $6.07 \pm .14$ m/s for every micron increase in diameter of a neuron. Applying (12.34), an approximate 95 % CI for the slope of the regression line is $6.07 \pm 2(.14)$ or (5.79, 6.35). For these data there were $n = 67$ observations, so we have $\nu = 65$ and $t_{.975,n-1} = 2.0$. Thus, the CI based on (12.35) is the same as that based on (12.34). □

Formula (12.34) may be justified by an extension of the theorem on the consistency of $\hat{\beta}_1$ in (12.31), which we present next.

Theorem: Asymptotic normality of least squares estimators For the linear regression model (12.5) suppose conditions (i)–(iv) hold and let $x_1, x_2, \ldots, x_n, \ldots$ be a sequence of x values such that

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \to c \tag{12.36}$$

for some positive constant c, as $n \to \infty$. Then the least-squares estimator defined by (12.7) satisfies

$$\begin{aligned} \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} &\xrightarrow{D} N(0, 1) \\ \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} &\xrightarrow{D} N(0, 1) \end{aligned} \tag{12.37}$$

where $SE(\hat{\beta}_1)$ and $SE(\hat{\beta}_0)$ are the standard errors given by (12.61).

Proof: This is a consequence of the CLT, but requires some algebraic manipulation. We omit the proof and again refer the interested reader to Wu (1981) for references. □

The condition (12.36) implies (12.30). It would be satisfied if we were drawing x_i values from a fixed probability distribution.[5] In the context of a particular set of data, the x_i values, even when selected by an experimenter, are somehow spread out and thus could be conceived as coming from some probability distribution (one that is not concentrated on a single value). On the other hand, the Anscombe example in Section 12.2.2 is a reminder that sensible interpretations require the fitted line to represent well the relationship between the x_i and y_i values. In the theoretical world this is expressed by saying that the model assumptions (i)–(iv) are satisfied. In practice we would interpret the theorems guaranteeing consistency and asymptotic normality of least-squares estimators, according to (12.31) and (12.37), as saying that if the regression model does a good job in describing the variation in the data, and the sample size is not too small, then the approximate confidence interval in (12.34) will produce appropriate inferences. We typically do not need normality of the errors, as specified in assumption (v). What we need is normality of the estimator, as in (12.37).

12.3.2 Evidence in favor of a linear trend can be obtained from a t-test concerning the slope.

In Examples 1.5 and 12.1 it is obvious that there are linear trends in the data. This kind of increasing or decreasing tendency is sometimes a central issue in an analysis. Indeed, in Example 12.1 the quantitative relationship, meaning the number of additional spikes per second per additional drop of juice, is not essential. Rather, the main conclusion involved the qualitative finding of increasing firing rate with increasing reward. In problems such as this, it makes sense to assume that y is roughly linear in x but to consider the possibility that in fact the slope of the line is zero—meaning that y is actually constant, on average, as x changes; that is, that y is really not related to x at all. We formalize this possibility as the null hypothesis H_0: $\beta_1 = 0$ and we test it by applying the z-test discussed in Section 10.3.2. In the one-sample problem of testing H_0: $\mu = \mu_0$, considered in Section 10.3.3, the z-test is customarily replaced by a t-test, which inflates the p-value somewhat for small samples and is justified under the assumption of normality of the data. Similarly, in linear regression, the z-test may be replaced by a t-test under the assumption of normality of errors (assumption (v) on p. 315). The test statistic becomes the t-ratio,

$$t\text{-ratio} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}. \tag{12.38}$$

For large samples, under H_0, this statistic has a $N(0, 1)$ distribution, but for small samples, if assumption (v) is satisfied, under H_0 the t-ratio has a t distribution on $\nu = n - 2$ degrees of freedom. This is the basis for the p-value reported by most

[5] Beyond (12.30), condition (12.36) says that the x_i values do not diverge extremely quickly, which would make $\hat{\beta}_1$ converge faster than $1/\sqrt{n}$.

statistical software. Here, the degrees of freedom are $n - 2$ because two parameters β_1 and β_0 from n freely ranging data values y_i. Generally speaking, when the magnitude of the t-ratio is much larger than 2 the p-value will be small (much less than .05, perhaps less than .01) and there will be clear evidence against H_0: $\beta_1 = 0$ and in favor of the existence of a linear trend.

Example 1.5 (continued, see page 11) For the conduction velocity data, testing H_0: $\beta_1 = 0$ with (12.38) we obtained $p < 10^{-15}$. Keeping in mind that very extreme tail probabilities are not very meaningful (they are sensitive to small departures from normality of the estimator) we would report this result as very highly statistically significant with $p << .0001$, where the notation $<<$ is used to signify "much less than." □

Example 12.1 (continued from p. 310) For the data shown in Fig. 12.1 the authors reported $p < .0001$. □

In the data reported in Fig. 12.1 there are only 7 distinct values of x_i, with many firing rates (across many trials) corresponding to each reward level. Thus, the 329 data pairs have been aggregated to 7 pairs with the mean value of y_i reported for each x_i. It turns out that the fitted line based on means is the same as the fitted line based on all 329 values considered separately. However, depending on the details of the way the computation based on the means is carried out, the standard error may or may not agree with the standard error obtained by analyzing all 329 values. To capture the full regularity and variation in the data, the hypothesis test should be based on all 329 values.

12.3.3 The fitted relationship may not be accurate outside the range of the observed data.

We have so far ignored an interesting issue that arises in Example 1.5. There, the fitted line does not go through the origin $(0, 0)$. In fact, according to the fitted line, when the diameter of the nerve is 0, the conduction velocity becomes negative! Should we try to fix this?

It is possible to force the line through $(0, 0)$ by omitting the intercept in the fitting process. Regression software typically provides an option for leaving out the intercept. However, for this data set, and for many others, omission of the intercept may be unwise. The reason is that the relationship may well be nonlinear near the origin, and there are no data to determine the fitted relationship in that region. Instead, we would view the fitted relationship as accurate only for diameters that are within the range of values examined in the data. Put differently, when the linear regression model does a good job of representing the regularity and variability in the data it allows us to interpolate (predict values within the range of the data) but may not be trustworthy if we try to extrapolate (predict values outside the range of the data).

12.4 Correlation and Regression

Sometimes the "explanatory variable" x is observed, rather than fixed by the experimenter. In this case the pair (x, y) is observed and we may model this by considering a pair of random variables X and Y and their *joint* distribution. Recall (from Section 4.2.1) that the *correlation coefficient* ρ is a measure of linear association between X and Y. As we discussed in Section 4.2.1, the best linear predictor $\beta_0 + \beta_1 X$ of Y satisfies

$$\beta_1 = \frac{\sigma_Y}{\sigma_X} \cdot \rho \tag{12.39}$$

as in Eq. (4.9). Also, the theoretical regression of Y on X is defined (see Section 4.2.4) to be $E(Y|X = x)$, which is a function of x, and it may happen that this function is linear:

$$E(Y|X = x) = \beta_0 + \beta_1 x.$$

In Chapter 4 we noted that the regression is, in fact, linear when (X, Y) has a bivariate normal distribution and then (12.39) holds. This linearity, and its interpretation, was illustrated in Fig. 4.3. However, the right-hand plot in Fig. 4.3 concerns data, rather than a theoretical distribution, and there is an analogous formula and interpretation using the sample correlation r, which was defined in (4.7). Under the assumption of bivariate normality, it may be shown that the sample correlation r is the MLE of ρ.

The sample correlation is related to the relative proportion of signal-to-noise variability R^2 by $R^2 = r^2$. Important properties are the following:

- $-1 \le r \le 1$ with $r = 1$ when the points fall exactly on a line with positive slope and $r = -1$ when the points fall exactly on a line with negative slope;
- the value of r is unitless and does not change when either or both of the two variables are linearly rescaled (e.g., when x is replaced by $ax + b$);
- just as ρ measures linear association between random variables X and Y, so too may r be considered a measure of *linear* association.

As we said in discussing R^2, there are no general guidelines as to what constitutes a "large" value of the correlation coefficient. Interpretation depends on the application.

12.4.1 The correlation coefficient is determined by the regression coefficient and the standard deviations of x and y.

Equation (12.39) gives the relationship of the theoretical slope β_1 to the theoretical correlation coefficient ρ. For data pairs (x_i, y_i) we have the analogous formula

$$\hat{\beta}_1 = \frac{s_Y}{s_X} \cdot r.$$

As a consequence, if x and y have about the same variability, the fitted regression slope becomes approximately equal to the sample correlation. In some contexts it is useful to standardize x and y by dividing each variable by its standard deviation. When that is done, the regression slope will equal the sample correlation.

12.4.2 Association is not causation.

There are numerous examples of two variables having a high correlation while no one would seriously suggest that high values of one causes high values of the other. For instance, one author (Brownlee 1965) looked at data from many different countries and pointed out that the number of telephones per capita had a strong correlation with the death rate due to heart disease. In such situations there are confounding factors that, presumably, have an effect on both variables and thus create a "spurious" correlation. Only in well-performed experiments, often using randomization,[6] can one be confident there are no confounding factors. Indeed, discussion sections of articles typically include arguments as to why possible confounding factors are unlikely to explain reported results.

12.4.3 Confidence intervals for ρ may be based on a transformation of r.

The sample correlation coefficient r may be considered an estimate of the theoretical correlation ρ and, as we mentioned on p. 327, under the assumption of bivariate normality r is the MLE of ρ. To get approximate confidence intervals the large-sample theory of Section 8.4.3 may be applied.[7] If we have a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ we may compute its sample correlation R_n, which is itself a random variable (so that when $X_1 = x_1, Y_1 = y_1, \ldots, X_n = x_n, Y_n = y_n$ we compute the sample correlation $R_n = r$ based on $(x_1, y_1), \ldots, (x_n, y_n)$). Now, if we consider a sequence of such samples from a bivariate normal distribution with correlation ρ it may be shown that

$$\frac{\sqrt{n}(R_n - \rho)}{(1 - \rho^2)} \xrightarrow{D} N(0, 1)$$

[6] Randomization refers to the random assignment of treatments to subjects, and to the process of randomly ordering treatment conditions; we discuss this further in Section 13.4.

[7] The usual derivation of the limiting normal distribution of r begins with an analytic calculation of the covariance matrix of (V_x, V_y, C) where $V_x = V(X)$, $V_y = V(Y)$, and $C = Cov(X, Y)$, in which (X, Y) is bivariate normal. That calculation provides an explicit formula for the covariance matrix in the limiting joint normal distribution of (V_x, V_y, C), and then propagation of uncertainty is applied as in Section 9.1.2.

as $n \to \infty$. This limiting normal distribution could be used to find confidence intervals. However, Fisher (1924) showed that a transformation of the correlation $R_n = r$ improves the limiting normal approximation. This is known as *Fisher's z transformation* (z because it creates a nearly $N(0, 1)$ distribution) defined by

$$z_r = \frac{1}{2} \log \left(\frac{1+r}{1-r} \right). \tag{12.40}$$

For the theoretical statement we again consider a sequence of bivariate normal random samples with sample correlations R_n and define

$$Z_R = \frac{1}{2} \log \left(\frac{1+R_n}{1-R_n} \right)$$

and

$$\zeta = \frac{1}{2} \log \left(\frac{1+\rho}{1-\rho} \right)$$

to get

$$\sqrt{n-3}(Z_R - \zeta) \xrightarrow{D} N(0, 1) \tag{12.41}$$

as $n \to \infty$ (see[8] p. 43 in DasGupta 2008). Consequently, we can define the lower and the upper bounds of an approximate 95 % confidence interval for the theoretical quantity ζ by

$$\begin{aligned} L_z &= z_r - 2\sqrt{\frac{1}{n-3}} \\ U_z &= z_r + 2\sqrt{\frac{1}{n-3}}. \end{aligned} \tag{12.42}$$

To get an approximate 95 % confidence interval for ρ we apply the inverse transformation

$$\rho = \frac{\exp(2\zeta) - 1}{\exp(2\zeta) + 1}$$

to L and U in (12.42) to get

$$\begin{aligned} L &= \frac{\exp(2L_z) - 1}{\exp(2L_z) + 1} \\ U &= \frac{\exp(2U_z) - 1}{\exp(2U_z) + 1}. \end{aligned} \tag{12.43}$$

[8] The z-transformation may be derived as a variance-stabilizing transformation, as on p. 232, beginning with the limiting result mentioned in footnote 7. More general results are given by Hawkins (1989).

Confidence interval for ρ
Suppose we have a random sample from a bivariate normal distribution with correlation ρ and $R_n = r$ is the sample correlation. Then an approximate 95 % confidence interval for ρ is given by (L, U) where L and U are defined by (12.43), (12.42), and (12.40).

The result (12.41) may also be used to test H_0:$\rho = 0$, which holds if and only if H_0: $\beta_1 = 0$. The procedure is to apply the z-test in Section 10.3.2 using

$$z_{obs} = \sqrt{n-3}z_r,$$

which is z_r divided by its large-sample standard deviation $1/\sqrt{n-3}$, and is thus a z-ratio.

12.4.4 When noise is added to two variables, their correlation diminishes.

When measurements are corrupted by noise, the magnitude of their correlation decreases. The precise statement is given in the theorem below, where we begin with two random variables U and W and then add noise to each, in the form of variables ϵ and δ. The noise-corrupted variables are then $X = U + \epsilon$ and $Y = W + \delta$.

Theorem: Attenuation of Correlation Suppose U and W are random variables having correlation ρ_{UW} and ϵ and δ are independent random variables that are also independent of U and V. Define $X = U + \epsilon$ and $Y = W + \delta$, and let ρ_{XY} be the correlation between X and Y. If $\rho_{UW} > 0$ then

$$0 < \rho_{XY} < \rho_{UW}.$$

If $\rho_{UW} < 0$ then

$$\rho_{UW} < \rho_{XY} < 0.$$

Proof details: We assume that $V(\epsilon) > 0$ and $V(\delta) > 0$ and we begin by writing

$$\begin{aligned} Cov(X, Y) &= Cov(U + \epsilon, W + \delta) \\ &= Cov(U, W) + Cov(U, \delta) + Cov(W, \epsilon) + Cov(\epsilon, \delta). \end{aligned}$$

Because of independence the last 3 terms above are 0. Therefore, $Cov(X, Y) = Cov(U, W)$, which shows that ρ_{XY} and ρ_{UW} have the same sign. Suppose $\rho_{UW} > 0$, so that $Cov(U, W) > 0$. Then we have

$$\begin{aligned}\rho_{XY} &= Cor(U+\epsilon, W+\delta)\\ &= \frac{Cov(U,W)}{\sqrt{V(U+\epsilon)V(W+\delta)}}\\ &= \frac{Cov(U,W)}{\sqrt{(V(U)+V(\epsilon))(V(W)+V(\delta))}}\\ &< \frac{Cov(U,W)}{\sqrt{Var(U)Var(W)}}\\ &= \rho_{UW}.\end{aligned}$$

If $\rho_{UW} < 0$ then $Cov(U, W) < 0$ and the inequality above is reversed. □

The theorem above indicates that when measurements are subject to substantial noise a measured correlation will underestimate the strength of the actual correlation between two variables. In the notation above, we wish to find ρ_{UW} but the corrupted measurements we observe would be $(X_1, Y_1), \ldots, (X_n, Y_n)$, and if we compute the sample correlation r based on these observations it will tend to be smaller than ρ_{UW} even for large samples. Thus, it is often the case that the sample correlation will underestimate an underlying correlation between two variables. However, if the *likely magnitude* of the noise is known it becomes possible to correct the estimate. Such corrections for attenuation of the correlation can be consequential.

Example 12.3 Correction for attenuation of the correlation in SEF selectivity indices Behseta et al. (2009) reported analysis of data from an experiment on neural mechanisms of serial order performance. Monkeys were trained to perform eye movements in a given order signaled by a cue. For example, one cue carried the instruction: look up, then right, then left. Based on recordings of neural activity in frontal cortex (the supplementary eye field, SEF) during task performance, Behseta et al. reported that many neurons fire at different rates during different stages of the task, with some firing at the highest rate during the first, some during the second and some during the third stage. These rank-selective neurons might genuinely be sensitive to the monkey's stage in the sequence. Alternatively, they might be sensitive to some correlated factor. One such factor is expectation of reward. Reward (a drop of juice) was delivered only after all three movements had been completed. Thus as the stage of the trial progressed from one to three, the expectation of reward might have increased.

To see whether rank-selective neurons were sensitive to the size of the anticipated reward, the same monkeys were trained to perform a task in which a visual cue presented at the beginning of the trial signaled to the monkey whether he would receive one drop or three drops of juice after a fixed interval. The idea was that neuronal activity related to expectation of reward would be greater after the promise of three drops than after the promise of one. Spike counts from 54 neurons were collected during the performance of both the serial order task and the variable reward task, and selectivity indices for rank in the serial order task and size of the anticipated reward in the variable reward task were computed. The rank selectivity index was $I_{\text{rank}} =$

$\frac{(f_3-f_1)}{(f_3+f_1)}$, where f_1 and f_3 were the mean firing rates measured at the times of the first and third saccades respectively, the mean being taken across trials. Similarly, the reward selectivity index was $I_{\text{reward}} = \frac{(f_b-f_s)}{(f_b+f_s)}$ where f_b and f_s were the mean firing rates during the post-cue delay period on big-reward and small-reward trials respectively. The selectivity indices I_{rank} and I_{reward} turned out to be positively correlated, but the effect was smaller than expected, with $r = .49$. The correlation between the rank and reward indices was expected to be larger because, from previous research, it was known that (a) the expectation of reward increases over the course of a serial order trial and (b) neuronal activity in the SEF is affected by the expectation of reward. Behseta et al. speculated that the correlation between the two indices had been attenuated by noise arising from trial-to-trial variations in neural activity, and they applied a correction for attenuation discussed in Chapter 16. This gave a dramatically increased correlation, with the new estimate of correlation becoming .83. Results given by Behseta et al. showed that the new estimate may be considered much more reliable than the original $r = .49$. □

12.5 Multiple Linear Regression

The simple linear regression model (12.5) states that the response variable Y arises when a linear function of a single predictive variable x is subjected to additive noise ϵ. The idea is easily extended to two or more predictive variables. Let us write the ith observation of the jth predictive variable as x_{ji}. Then, for p predictive variables the linear regression model becomes

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i \tag{12.44}$$

where the ϵ_i's have the same assumptions as in (12.5).

Let us start with the case $p = 2$. Just as $y = \beta_0 + \beta_1 x_1$ describes a line, the equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ describes a plane. When only a single variable x_1 is involved, the coefficient β_1 is the slope: $\beta_1 = \Delta y / \Delta x$. For example, if we increase x by $\Delta x = 2$ then we increase y by $\Delta y = 2\beta_1$. In the case of the equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, if we increase x_1 by $\Delta x_1 = 2$ and ask what happens to y, the answer will depend on how we change x_2. However, if we hold x_2 fixed while we increase x_1 by $\Delta x_1 = 2$ then we will increase y by $\Delta y = 2\beta_1$. When $p = 2$, β_1 is interpreted as the change in y for a one-unit change in x_1 *when* x_2 *is held fixed*. If $p > 2$ then β_1 becomes the change in y for a one-unit change in x_1 *when* $x_2, \ldots, x_p$ *are all held fixed*. Thus, linear regression is often used as a way of assessing what *might* happen if we *were* to hold one or more variables fixed while allowing a different variable to fluctuate. Put differently, regression allows us to assess the relationship between x_1 and y *after adjusting for* the variables $x_2, \ldots, x_p$. In this context $x_2, \ldots, x_p$ are often called *covariates*, because[9] they co-vary with x_1 and y.

[9] See also "analysis of covariance," mentioned in the footnote on p. 379.

Example 12.4 Developmental change in working memory from fMRI Many studies have documented the way visuospatial working memory (VSWM) changes during development. Kwon et al. (2002) used fMRI to examine neural correlates of these changes. These authors studied 34 children and young adults, ranging in age from 7 to 22. Each subject was given a VSWM task while being imaged. The task consisted of 12 alternating 36-s working memory (WM) and control epochs during which subjects viewed items on a screen. During both the WM and control versions of the task the subjects viewed the letter "O" once every 2 s at one of nine distinct locations on the screen. In the WM task the subjects responded when the current location was the same as it was when the symbol was presented two stimuli back. This required the subjects to engage their working memory. In the control condition the subjects responded when the "O" was in the center of the screen.

One of the y variables used in this study was the maximal BOLD activation (as a difference between WM and control) among voxels within the right prefrontal cortex. They were interested in the relationship of this variable with age (x_1). However, it is possible that Y would increase due to better performance of the task, and that this would increase with age. Therefore, in principle, the authors wanted to "hold fixed" the performance of task while age varied. This is, of course, impossible. What they did instead was to introduce two measures of task performance: the subjects' accuracy in performing the task (x_2) and their mean reaction time (x_3). □

Example 12.1 (continued, see p. 310) The firing rates in Fig. 12.1 appear clearly to increase with size of reward, and the analysis the authors reported (see p. 326) substantiated this impression. Platt and Glimcher also considered whether other variables might be contributing to firing rate by fitting a multiple regression model using, in addition to the normalized reward size, amplitude of each eye saccade, average velocity of saccade, and latency of saccade. This allowed them to check whether firing rate tended to increase with normalized reward size after accounting for these eye saccade variables. □

Equation (12.6) defined the least squares fit of a line. Let us rewrite it in the form

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \min_{\beta^*} \sum_{i=1}^{n}(y_i - y_i^*)^2 \tag{12.45}$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $y_i^* = \beta_0^* + \beta_1^* x_i$ and $\beta^* = (\beta_0^*, \beta_1^*)$. If we now re-define y_i^* as

$$y_i^* = \beta_0^* + \beta_1^* x_{1i} + \cdots + \beta_p^* x_{pi}$$

with $\beta^* = (\beta_0^*, \beta_1^*, \ldots, \beta_p^*)$, Eq. (12.45) defines the least-squares multiple regression problem. We write the solution in vector form as

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p), \tag{12.46}$$

where the components satisfy (12.45) with the fitted values being

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_p x_{pi}. \tag{12.47}$$

We interpret the multiple regression equation in Section 12.5.1 and discuss the decomposition of sums of squares in Section 12.5.2. In Section 12.5.3 we show how the multiple regression model may be written in matrix form, which helps in demonstrating how it includes ANOVA models as special cases, and in Section 12.5.4 we show that multiple regression also may be used to analyze certain nonlinear relationships. In Section 12.5.5 we issue an important caveat concerning correlated explanatory variables; in Section 12.5.6 we describe the way interaction effects are fitted by multiple regression; and in Section 12.5.7 we provide a brief overview of the way multiple regression is used when there are substantial numbers of alternative explanatory variables. We close our discussion of multiple regression in Section 12.5.8 with a few words of warning.

12.5.1 Multiple regression estimates the linear relationship of the response with each explanatory variable, while adjusting for the other explanatory variables.

To demonstrate multiple regression in action we consider a simple example.

Example 12.5 Toxicity as a function of dose and weight In many studies of toxicity, including neurotoxicity (Makris et al. 2009) a drug or other agent is given to an animal and toxicity is examined as a function of dose and animal weight. A relatively early example was the study of sodium arsenate (arsenic) in silkworm larvae (Bliss 1936). We reanalyzed data reported there. The response variable (y) was $\log(w/1{,}000)$ where w was minutes survived, and the two predictive variables were log weight, in log grams, and log dose, given in 1.5 plus log milligrams. A plot of log survival versus log dose is given in Fig. 12.8. Because there were two potential outliers that might affect the slope of the line fitted to the plotted data we have provided in the plot the fitted regression lines with and without those two data pairs. The results we discuss were based on the complete set of data.

The linear regression of log survival on log dose gave the fitted line

$$\text{log survival} = .140(\pm.057) - .704(\pm.078)\text{log dose}$$

which says that survival decreased roughly $.704(\pm.078)$ log 1,000 min for every log milligram increase in dose. The regression was very highly significant ($p = 10^{-12}$), consistently with the obvious downward trend.

The linear regression of log survival on both log dose and log weight gave the fitted line

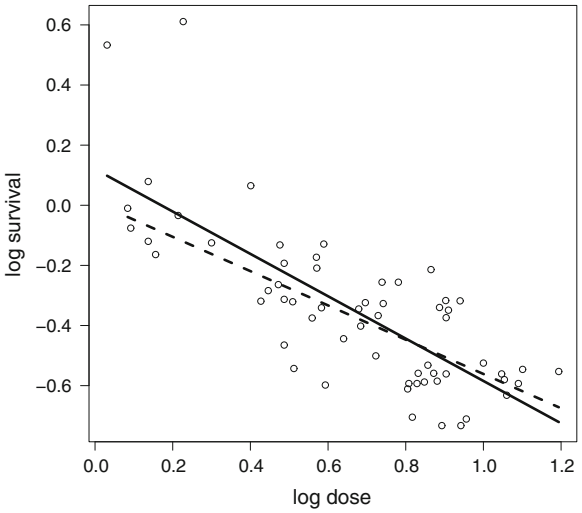

Fig. 12.8 Plot of log survival time (log(w/1,000) where w was minutes survived) versus log dose (1.5 plus log milligrams) of sodium arsenate in silkworm larvae; data from Bliss (1936). Lines are fits based on linear regression: *solid line* used the original data shown in plot; *dashed line* after removing the two high values of survival at low dose.

$$\text{log survival} = .140(\pm.057) - .734(\pm.058)\text{log dose} + 1.07(\pm.16)\text{log weight}.$$

In this case, including weight in the regression does not change very much the relationship between dose and survival: the slope is nearly the same in both cases. □

12.5.2 Response variation may be decomposed into signal and noise sums of squares.

As in simple linear regression we define the sums of squares *SSE* and *SSR*, again using (12.22) and (12.28) except that now $\hat{y}_i$ is defined by (12.47). If we continue to define the total sum of squares as in (12.24) we may again decompose it as

$$SST = SSR + SSE$$

and we may again define R^2 as in (12.25) or, equivalently, (12.27). In the multiple regression context R^2 is interpreted as a measure of the strength of the linear relationship between y and the multiple explanatory variables.

With p variables we may again use the sum of squares of the residuals to estimate the noise variation σ^2 but we must change the degrees of freedom appearing in (12.21). Because we again start with $n-1$ degrees of freedom in total, we subtract p to get $n-1-p$ degrees of freedom for error, and we have

$$s^2 = \frac{1}{n-1-p} SSE \tag{12.48}$$

where *SSE* is defined by (12.22). In multiple regression the hypothesis of no linear relationship between y and the x variables is H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$. To test this hypothesis we define and compare suitable versions of *MSR* and *MSE*, the idea being that under H_0, with no linear relationship at all, *MSR* and *MSE* should be about the same size because both represent fluctuation due to noise. With p explanatory variables there are p degrees of freedom for regression. We therefore define the mean squared error for regression

$$MSR = \frac{SSR}{p}.$$

We use (12.48) in (12.23) for the mean squared error. We then form[10] the F-ratio

$$F = \frac{MSR}{MSE}. \tag{12.49}$$

In words, F is the ratio of the mean squared errors for regression and error, which are obtained by dividing the respective sums of squares by the appropriate degrees of freedom. Under the standard assumptions, if H_0 holds this F-ratio follows an F distribution, which will be centered near 1.

To state the result formally we must define a theoretical counterpart to (12.49). Let $\hat{Y}_i$ be the random variable representing the least-squares fit under the linear regression assumptions on p. 315, i.e., it is the theoretical counterpart of (12.47). We define

$$U_{MSE} = \frac{1}{p} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{12.50}$$

and

$$U_{MSR} = \frac{1}{n-1-p} \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_i)^2. \tag{12.51}$$

[10] The letter F was chosen (by George Snedecor in 1934) to honor Fisher, who had first suggested a log-transformed normalized ratio of sums of squares, and derived its distribution, in the context of ANOVA, which we discuss in Chapter 13.

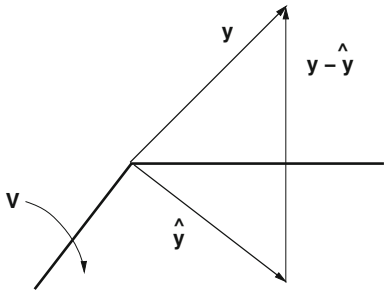

Fig. 12.9 Orthogonal projection of the vector y onto the vector subspace V resulting in the vector $\hat{y}$ in V. The residual vector $y - \hat{y}$ is orthogonal to $\hat{y}$, which gives the Pythagorean relationship (12.57). This corresponds to the total sum of squares (the squared length of y) equaling the sum of the regression sum of squares (the squared length of $\hat{y}$) and the error sum of squares (the squared length of $y - \hat{y}$).

Result: F-Test for Regression
Under the linear regression assumptions on p. 315, with (12.44) replacing (12.5), if H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ holds then the F-statistic

$$F = \frac{U_{MSR}}{U_{MSE}} \tag{12.52}$$

follows an F_{ν_1,ν_2} distribution, where $\nu_1 = p$ and $\nu_2 = n - 1 - p$.

Proof outline: If H_0 is true, it may be shown that

$$\sum(\hat{Y}_i - \bar{Y})^2 \sim \chi^2_{\nu_1}$$

and

$$\sum(Y_i - \hat{Y}_i)^2 \sim \chi^2_{\nu_2}$$

where $\nu_2 = n - 1 - p$ is the degrees of freedom for error, and it may be shown that these are independent. Therefore, the random variable F defined by (12.52) is a ratio of independent chi-squared random variables divided by their degrees of freedom, which, by the definition on p. 129 has an F_{ν_1,ν_2} distribution. □

We provide a geometrical interpretation of the sum of squares decomposition below, in Fig. 12.9 and Eq. (12.57).

In simple linear regression, where there is only one explanatory variable, $\nu_1 = 1$ and F is equal to the square of the t-ratio. Because the square of a t_ν distributed random variable has an $F_{1,\nu}$ distribution, it follows that the t-test and the F-test of H_0: $\beta_1 = 0$ are identical. In multiple regression, hypotheses may also be tested about the individual coefficients, e.g., $H_0: \beta_2 = 0$, using t-tests.

Table 12.1 Simple linear regression results for Example 12.5.

Variable	Coefficient	SE	t_{obs}	p-value
(Intercept)	.120	.057	2.1	.038
Log dose	−.704	.078	−9.1	10^{-12}

Table 12.2 Multiple regression results for Example 12.5.

Variable	Coefficient	SE	t_{obs}	p-value
(Intercept)	−.140	.057	−2.49	.017
Log dose	−.734	.058	−12.6	2×10^{-16}
Log weight	1.07	.16	6.8	6×10^{-9}

Example 12.5 (continued) Returning to the toxicity data, the results for the regression of log survival on log dose are given in Table 12.1. We also obtained $s = .17$ and $R^2 = .59$. The F-statistic was $F = 82$ on 1 and 58 degrees of freedom, with $p = 10^{-12}$ in agreement with the p-value for the t-test in Table 12.1. The results for the regression of log survival on both log dose and log weight are in Table 12.2 and here $s = .13$ and $R^2 = .77$, which is a much better fit. The F-statistic was $F = 97$ on 2 and 57 degrees of freedom, with $p = 2 \times 10^{-16}$.

We would interpret the t ratios and F-statistics as follows: there is very strong evidence of a linear relationship between log survival and a linear combination of log dose and log weight ($F = 97, p << 10^{-5}$); given that log weight is included in the regression model, there is very strong evidence ($t = -12.6, p << 10^{-5}$) that log survival has a decreasing linear trend with log dose; similarly, given that log dose is in the model, there is very strong evidence ($t = 6.8, p << 10^{-5}$) that survival has an increasing linear trend with log weight. □

Example 12.4 (continued from p. 333) Recall that in one of their analyses Kwon et al. defined Y to be the maximal BOLD activation (as a difference between WM and control) among voxels within the right prefrontal cortex, and they considered its linear relationship with age (X_1), accuracy (X_2) and reaction time (X_3). They then performed multiple linear regression and found $R^2 = .53$ with $\beta_1 = .75(\pm.20)$, $p < .001$, $\beta_2 = -.21(\pm.19)$, $p = .28$, and $\beta_3 = -.15(\pm.17)$, $p = .37$. They interpreted the results as showing that the right PFC tends to become much more strongly activated in the VSWM task as the subjects' age increases, and that this is not due solely to improvement in performance of the task. □

Example 12.1 (continued from p. 333) Platt and Glimcher fit a multiple regression model to the firing rate data using as explanatory variables normalized reward size, amplitude of each eye saccade, average velocity of saccade, and latency of saccade. They reported the results of the t-test for the normalized reward size coefficient as $p < .05$, which indicates that firing rate tended to increase with normalized

reward size even after accounting for these eye saccade variables. A plot showing the coefficient with *SE* makes it appear that actually $p << .05$, which is much more convincing. □

The distributional results for the statistic F in (12.52) are based on the assumption of normality of the errors. For sufficiently large samples the p-values for the F-statistic, and the t-based p-values and confidence intervals, will be approximately correct even if the errors are non-normal. This is due to the theorems on consistency (p. 318) and approximate normality (p. 324), which extend to multiple regression (p. 344). However, the independence assumption is crucial. The standard errors and other distributional results generally may be trusted for reasonably large samples when the errors are independent, but they require correction otherwise. The assumptions should be examined using residual plots, as in simple linear regression.

12.5.3 Multiple regression may be formulated concisely using matrices.

Mathematical manipulations in multiple regression could get very complicated. A great simplification is to collect multiple equations together and write them as single equations in matrix form. We start by writing the n random variables Y_i as an $n \times 1$ random vector

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

and then similarly write

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

and

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}.$$

The linear regression model may then be written in the form

$$Y = X\beta + \epsilon \tag{12.53}$$

where it is quickly checked that both left-hand side and right-hand side are $n \times 1$ vectors. The usual assumptions may also be stated in matrix form. For example, we have

$$\epsilon \sim N_n(0, \sigma^2 \cdot I_n) \tag{12.54}$$

which says that ϵ has a multivariate normal distribution of dimension n, with mean equal to the zero vector and variance matrix equal to σ^2 times the n-dimensional identity matrix, i.e.,

$$V(\epsilon) = \sigma^2 \cdot \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & & & \ddots & 0 \\ 0 & \cdots & 0 & & 1 \end{pmatrix}.$$

Equation (12.53), together with the assumptions, is often called the *general linear model*. It accommodates not only multiple regression but also a large variety of models[11] that compare experimental conditions, which arise in analysis of variance (Chapter 13). For example, a standard approach to the analysis of fMRI data is based on a suitable linear model.

Example 12.2 (continued from p. 313) In Eq. (12.20) we defined a variable x_i that could be used with simple linear regression to analyze the BOLD response due to activity associated with a particular stimulus, according to an assumed form for the hemodynamic response function.[12] Suppose there are two stimuli with $u_j = 1$ corresponding to the first stimulus being on, with $u_j = 0$ otherwise, and $v_j = 1$ corresponding to the second stimulus being on, with $v_j = 0$ otherwise. We then define

[11] Sometimes when someone refers to the general linear model they may also allow the variance matrix to be different, or they may allow for non-normal errors.

[12] Before regression is applied various pre-processing steps are usually followed to make the assumptions of linear regression a reasonable representation of the variation in the fMRI data.

$$x_{i1} = \sum_{j<i} h(i-j)u_j$$
$$x_{i2} = \sum_{j<i} h(i-j)v_j$$

and set the X matrix equal to

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix}.$$

If we apply (12.53) with $\beta = (\beta_0, \beta_1, \beta_2)^T$ the coefficient β_1 will represent the magnitude of the effect of the first stimulus on the BOLD response, the coefficient β_2 will represent the magnitude of the effect of the second stimulus on the BOLD response, and the coefficient β_0 will represent the baseline BOLD response. □

Because X often reflects the design of an experiment, as in Example 12.2 above, it is called the *design matrix*. The assumptions associated with (12.53) are essentially the same as those enumerated (i)–(v) for simple linear regression, where (i) becomes the validity of Eq. (12.53) and (ii)–(v) refer to the components of ϵ.

In matrix form we may write the least-squares fit as $\hat{y}$ according to

$$||y - \hat{y}||^2 = \min_{\beta^*} ||y - y^*||^2$$
$$y^* = X\beta^*$$

where $||w||$ is used to indicate the length of the vector w. We assume here that $X^T X$ is nonsingular (see the Appendix for a definition). The solution is found by solving the equations

$$X^T X\beta = X^T y \tag{12.55}$$

numerically (by numerically stable methods) and the solution may be written in the form[13]

$$\hat{\beta} = (X^T X)^{-1} X^T y. \tag{12.56}$$

Formula (12.56) may be obtained by a simple geometrical argument. We begin by thinking of y as a vector in n-dimensional space and we consider the subspace V consisting of all linear combinations of the columns of X. We say that V is the linear subspace spanned by the columns of X, which is the set of all vectors that may be written in the form $X\beta^*$ for some β^*, i.e.,

[13] The equations are *not* solved merely by inverting the matrix $X^T X$; this can lead to grossly incorrect answers due to seemingly innocuous round-off error. See Section 12.5.5.

$$V = \{X\beta^*,\ \beta^* \in R^{p+1}\}$$

(see the Appendix). The subspace V is the space of all possible fitted vectors. The problem of least squares, then, is to find the closest vector in V to the data vector y, i.e., the problem is to minimize the Euclidean distance between y and V. The solution to this minimization problem is the fitted vector $\hat{y} = X\hat{\beta}$. See Fig. 12.9. This geometry also gives us the Pythagorean relationship

$$||y||^2 = ||\hat{y}||^2 + ||y - \hat{y}||^2 \tag{12.57}$$

which is the basis for the decomposition $SST = SSR + SSE$.

Details: Euclidean geometry says that $\hat{y}$ must be obtained by orthogonal projection of y onto the subspace spanned by the columns of X and, as a result, the residual $y - \hat{y}$ must be orthogonal to the subspace spanned by the columns of X, which means that $y - \hat{y}$ must be orthogonal to $X\beta$ for every β. This, in turn, may be written in the following form: for all β,

$$\langle X\beta, y - \hat{y}\rangle = 0 \tag{12.58}$$

where $\langle u, v\rangle = u^T v$ is the inner product of u and v. Substituting $\hat{y} = X\hat{\beta}$ we have

$$\langle X\beta, y - X\hat{\beta}\rangle = 0$$

for all β, and rewriting this we find that

$$\beta^T X^T y = \beta^T X^T X\hat{\beta}$$

for all β, which gives us Eq. (12.55). Equation (12.55) is sometimes called the set of *normal equations* (presumably using "normal" in the sense of "orthogonal"; and plural because (12.55) is a vector equation and therefore a set of scalar equations). Because (12.58) holds for all β, it holds in particular for $\beta = \hat{\beta}$, i.e.,

$$\langle \hat{y}, y - \hat{y}\rangle = 0$$

which, as illustrated in Fig. 12.9, gives (12.57).
For the SST decomposition we introduce the $n \times 1$ vector having all of its elements equal to 1, which we write $1_{vec} = (1, 1, \ldots, 1)^T$. In the argument above we replace y by the residual following projection of y onto 1_{vec},

$$\begin{aligned}\tilde{y} &= y - \frac{< y, 1_{vec} >}{< 1_{vec}, 1_{vec} >} 1_{vec} \\ &= y - \bar{y} 1_{vec}\end{aligned}$$

(which is the vector of residuals found by regressing y on 1_{vec}) and similarly for all $j = 2, \ldots, p+1$ replace the j column of X by its residual following projection onto 1_{vec} (which produces the vectors of residuals found by regressing each x variable on 1_{vec}). When we repeat the argument with these new variables we get a new fitted vector $\hat{\tilde{y}}$ and everything goes through as before. We then obtain the version of (12.57) needed for the decomposition:

$$||\tilde{y}||^2 = ||\hat{\tilde{y}}||^2 + ||y - \hat{y}||^2.$$

It may be verified that this is the same as $SST = SSR + SSE$. For example, $||\tilde{y}||^2 = \sum(y_i - \bar{y})^2$. □

The variance matrix of the least-squares estimator is easy to calculate using a generalization of Eq. (4.26): with a little algebra it may be shown that if W is a $p \times 1$ random vector with variance matrix $V(W) = \Sigma$ and A is a $k \times p$ matrix, then the variance matrix of AW is

$$V(AW) = A\Sigma A^T. \tag{12.59}$$

Using (12.59) we obtain

$$\begin{aligned} V(\hat{\beta}) &= ((X^TX)^{-1}X^T)\sigma^2 I_n((X^TX)^{-1}X^T)^T \\ &= \sigma^2 \cdot (X^TX)^{-1}X^TX(X^TX)^{-1} \\ &= \sigma^2 \cdot (X^TX)^{-1}. \end{aligned}$$

This variance matrix summarizes the variability of $\hat{\beta}$. For instance, we have

$$V(\hat{\beta}_k) = \sigma^2 \cdot (X^TX)^{-1}_{kk},$$

which is the kth diagonal element of the variance matrix. To use such formulas with data, however, we must substitute s for σ. We then have the estimated variance matrix

$$\hat{V}(\hat{\beta}) = s^2 \cdot (X^TX)^{-1} \tag{12.60}$$

and the standard errors are given by

$$SE(\hat{\beta}_k) = \sqrt{s^2 \cdot (X^TX)^{-1}_{kk}}. \tag{12.61}$$

For example, (12.61) is the formula that was used to produce the standard errors in Table 12.2, and to get the standard errors and t-ratios, and thus the p-values, in Example 12.4 reported on p. 338. For problems involving propagation of uncertainty (Section 9.1) to a function of $\hat{\beta}$, the variance matrix in (12.60) would be used.

The estimator (12.60), and resulting inferences, may be justified by the analogue to (12.37).

Theorem: Asymptotic normality of least squares estimators For the linear regression model (12.53) suppose conditions (i)–(iv) hold and let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of design matrices such that

$$\frac{1}{n} X^T X \to C \tag{12.62}$$

for some positive definite matrix C, as $n \to \infty$. Then the least-squares estimator defined by (12.56) satisfies

$$\frac{1}{s}(X_n^T X_n)^{1/2}(\hat{\beta} - \beta) \xrightarrow{D} N_{p+1}(0, I_{p+1}). \tag{12.63}$$

Proof: See Wu (1981) for references. □

> *A Detail:* It is also possible to use the bootstrap in regression, but this requires some care because under the assumptions (i)–(iv) the random variables Y_i have distinct expected values,
>
> $$E(Y_i) = (1, x_{i1}, \ldots, x_{ip})\beta$$
>
> and so are not i.i.d. The usual approach is to resample the studentized residuals (see p. 319), which are approximately i.i.d. See Davison and Hinkley (1997, page 275). Alternatively, when each vector $x_i = (x_{i1}, \ldots, x_{ip})$ is observed, rather than chosen by the experimenter, it is possible to treat x_i as an observation from an unknown multivariate probability distribution, and thus (x_i, y_i) becomes an observation from an unknown distribution, and the data vectors $((x_1, y_1), \ldots, (x_n, y_n))$ may be resampled.[14] This was the bootstrap procedure mentioned in Example 8.2 on p. 241. For additional discussion see Davison and Hinkley (1997). □

There are many conveniences of the matrix formulation of multiple regression in (12.53) together with (12.54). One is that the independence and homogeneity assumptions in (12.54) may be replaced. Those assumptions imply

$$V(\epsilon) = \sigma^2 I_n,$$

as in (12.54). The analysis remains straightforward if we instead assume

$$V(\epsilon) = R \tag{12.64}$$

[14] Here, Eq. (9.27) becomes

$$\hat{F}_n(x, y) \xrightarrow{P} F_{(X,Y)}(x, y)$$

where $\hat{F}_n$ is the empirical cdf computed from the random vectors $((X_1, Y_1), \ldots, (X_n, Y_n))$.

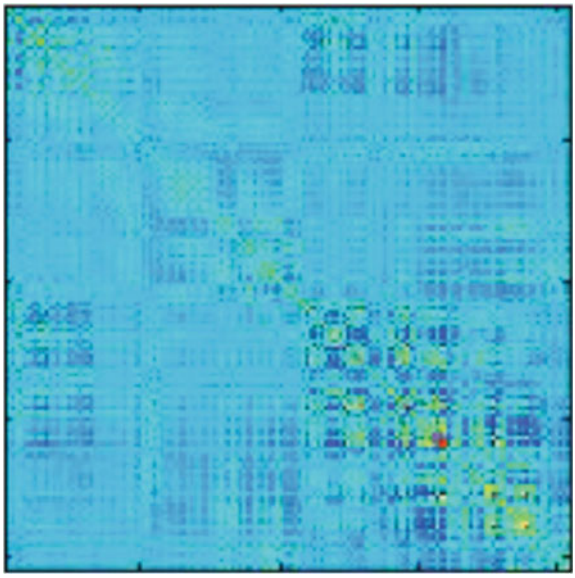

Fig. 12.10 MEG gradiometer background noise covariance matrix. The *light blue* corresponds to zero elements and *darker blue*, *yellow*, and *red* indicate non-zero elements. (Figure furnished by Gustavo Sudre.)

where R can be any $n \times n$ variance matrix (i.e., a positive definite symmetric matrix). **Example 1.2 (continued from p. 5)** We previously noted that MEG imaging requires sensor data to be obtained first from background scanner noise, meaning the sensor data must be obtained with nothing in the scanner. We displayed on p. 54 a histogram of such data, from a single sensor, as an example of a normal distribution. The separate sensor readings are not independent but are, instead, correlated. Figure 12.10 displays a representation of the background noise variance matrix from 204 gradiometer sensors in a MEG scanner. MEG analysis is based on (12.53) together with (12.64), with R being based on the background noise variance matrix. □

Given a matrix R in (12.64), and assuming it is positive definite, the least-squares problem may be reformulated. Letting $U = R^{-1/2}Y$ and $W = R^{-1/2}X$ we have

$$R^{-1/2}(Y - X\beta) = R^{-1/2}\epsilon \sim N_n(0, I_n),$$

so that the new model

$$U = W\beta + \delta,$$

where $\delta = R^{-1/2}\epsilon$, satisfies the usual assumptions in (12.53) together with (12.54). Therefore, to fit the model (12.53) with (12.64) we may first transform Y and X by premultiplying with $R^{-1/2}$ and then can apply ordinary least squares to the transformed variables. This is called *weighted least squares* and it arises in various extensions of

multiple regression. On p. 212 we showed that the least-squares estimator was also the MLE under the standard assumptions of regression, including normality of the errors. More generally, the weighted least squares estimator of β is the MLE under (12.53) with (12.64).

Example 1.2, above, provides a case in which the non-independence of the components of ϵ is due to the spatial layout of the sensors, and the resulting dependence among the magnetic field readings at different sensors. Neuroimaging also typically generates temporal correlation in the measurements, i.e., the measurements are time series with some dependence across time. Using auto-regressive time series models described in Section 18.2.3 the variance matrix may be determined from the data and this furnishes an R matrix in (12.64). The model (12.53) with (12.64) then leads to *regression with time series errors*.

12.5.4 The linear regression model applies to polynomial regression and cosine regression.

In many data sets the relationship of y and x is mildly nonlinear, and a quadratic in x may offer better results than a line. Even though a quadratic is nonlinear, a neat trick allows us to fit quadratic regression via multiple linear regression. The trick is to set $w_1 = x$ and to define a new variable $w_2 = x^2$. Then, when y is regressed on both w_1 and w_2 this amounts to fitting a general quadratic of the form $y = a + bx + cx^2$, where now $a = \beta_0$, $b = \beta_1$ and $c = \beta_2$. To be clear, we define the vector w_1 as

$$w_1 = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \tag{12.65}$$

and the vector w_2 as

$$w_2 = \begin{pmatrix} x_1^2 \\ x_2^2 \\ \vdots \\ x_n^2 \end{pmatrix} \tag{12.66}$$

and then we regress $y = (y_1, \ldots, y_n)$ on w_1 and w_2.

In quadratic regression there are several possibilities. First, there may be evidence of a linear association between y and x (from the simple linear regression), but the relationship appears nonlinear and there is also evidence of a linear association between y and both x and x^2 combined. This latter evidence would come from the combined regression output of (i) a statistically significant F-ratio and (ii) a

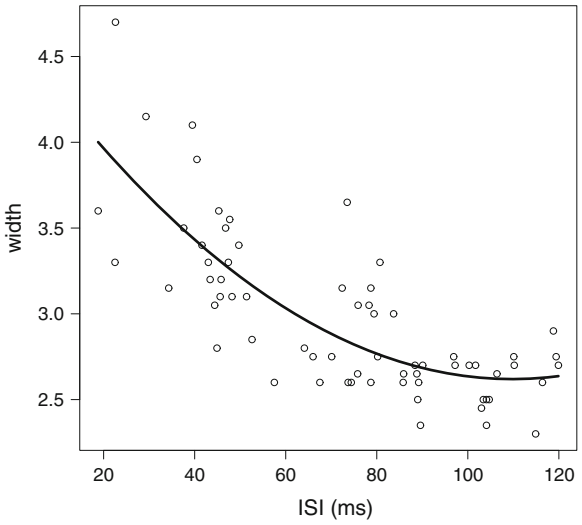

Fig. 12.11 Plot of action potential width against length of previous ISI, together with quadratic fitted by linear regression.

significant t-ratio for the coefficient of x^2. This case is illustrated below. Note that it is possible for the coefficient of x in the combined regression to be non-significant. This should not necessarily be taken to mean that there is no linear component to the relationship: it is generally preferable to use the general form $y = a + bx + cx^2$, which requires the bx term and thus the x variable. Actually, it is possible for the coefficients of *both* x and x^2 to be non-significant while the F-ratio is significant; this occurs when the two variables are themselves so highly correlated that neither adds anything to the regression when the other is already used.

As a second possibility, there may be evidence of a linear association between y and x (from the simple linear regression), but there is no evidence of a quadratic relationship. The latter would be apparent from (i) an OK (not curved) residual plot in the simple linear regression and (ii) a non-significant t-ratio for the coefficient of x^2. The third possibility is that there may be no evidence of a relationship between y and *either* x by itself or x combined with x^2. This would be evident from an insignificant t-ratio in the simple linear regression and an insignificant F-ratio in the combined regression.

Let us now turn to an example.

Example 8.2 (continued from p. 193) On p. 193 we examined spike train data recorded from a barrel cortex neuron in slice preparation, which was part of a study on the effects of seizure-induced neural activity. Figure 8.5 displayed the decreasing width of action potentials with increasing length of the interspike interval. Figure 12.11 shows a plot of many action potential widths against preceding interspike interval (ISI), where the data have been selected to include only ISIs of length

less than 120 ms. In the plot, the downward trend begins to level off near 100 ms, and a quadratic curve fitted by linear regression is able to capture the leveling off reasonably well within this range of ISI values. In this case the linear and quadratic regression coefficients were both highly significant ($p = 6 \times 10^{-6}$ and $p = .0017$, respectively, with the overall F-statistic giving $p = 8 \times 10^{-14}$) and $R^2 = .61$. □

In quadratic regression, illustrated in Example 8.2 above, we defined $w_1 = x$ and $w_2 = x^2$. To fit cubic and higher-order polynomials we may continue the process with $w_3 = x^3$, etc. An important caveat, however, is that the variables x_1, x_2, and x_3 defined in this way are likely to be highly correlated, which may cause difficulties in interpretation and, in extreme cases, may cause the matrix $X^T X$ to be singular (non-invertible), in which case least-squares software will fail to return a useful result. We discuss this issue further in Section 12.5.5.

A second nonlinear function that may be fitted with linear regression is the cosine.

Example 12.6 Directional Tuning in Motor Cortex In a well-known set of experiments, Georgopoulos, Schwartz and colleagues showed that motor cortex neurons are directionally "tuned." Figure 12.12 shows a set of raster plots for a "center-out" reaching task: the monkey reached to one of eight points on a circular image, and this neuron was much more active for reaches in some directions than for others. The bottom part of Fig. 12.12 shows a cosine function that has been fitted to the mean firing rate as a function of the angle around the circle, which indicates the direction of reach. For example (and as is also shown in the raster plots), reaches at angles near 180° from the x-axis produced high firing rates, while those at angles close to 0° (movement to the right) produced much lower firing rates. The angle at which the maximum firing rate occurs is called the "preferred direction" of the cell. It is obtained from the cosine function.

To fit a cosine to a set of spike counts, multiple linear regression is used. Let $v = (v_1, v_2)$ be the vector specifying the direction of movement and let $d = (d_1, d_2)$ be the preferred direction for the neuron. Both v and d are unit vectors. Assuming cosine tuning, the firing depends only on $\cos\theta$, where θ is the angle between v and d. We have

$$\cos\theta = v \cdot d = v_1 d_1 + v_2 d_2.$$

Letting $\mu(v)$ be the mean firing rate in a given interval of time when the movement is in direction v, if we let the minimal firing rate be B_{min} and the maximal firing rate be B_{max}, then cosine tuning may be written as the requirement that

$$\mu(v) = B_{min} + \frac{B_{max} - B_{min}}{2} + \frac{B_{max} - B_{min}}{2} \cos\theta.$$

(Recall that the minimal value of the cosine is -1, and its maximal value is 1.) If we now define $\beta_1 = \frac{B_{max} - B_{min}}{2} d_1$, $\beta_2 = \frac{B_{max} - B_{min}}{2} d_2$, and $\beta_0 = B_{min} + \frac{B_{max} - B_{min}}{2}$ we obtain the linear form

$$\mu(v) = \beta_0 + \beta_1 v_1 + \beta_2 v_2. \tag{12.67}$$

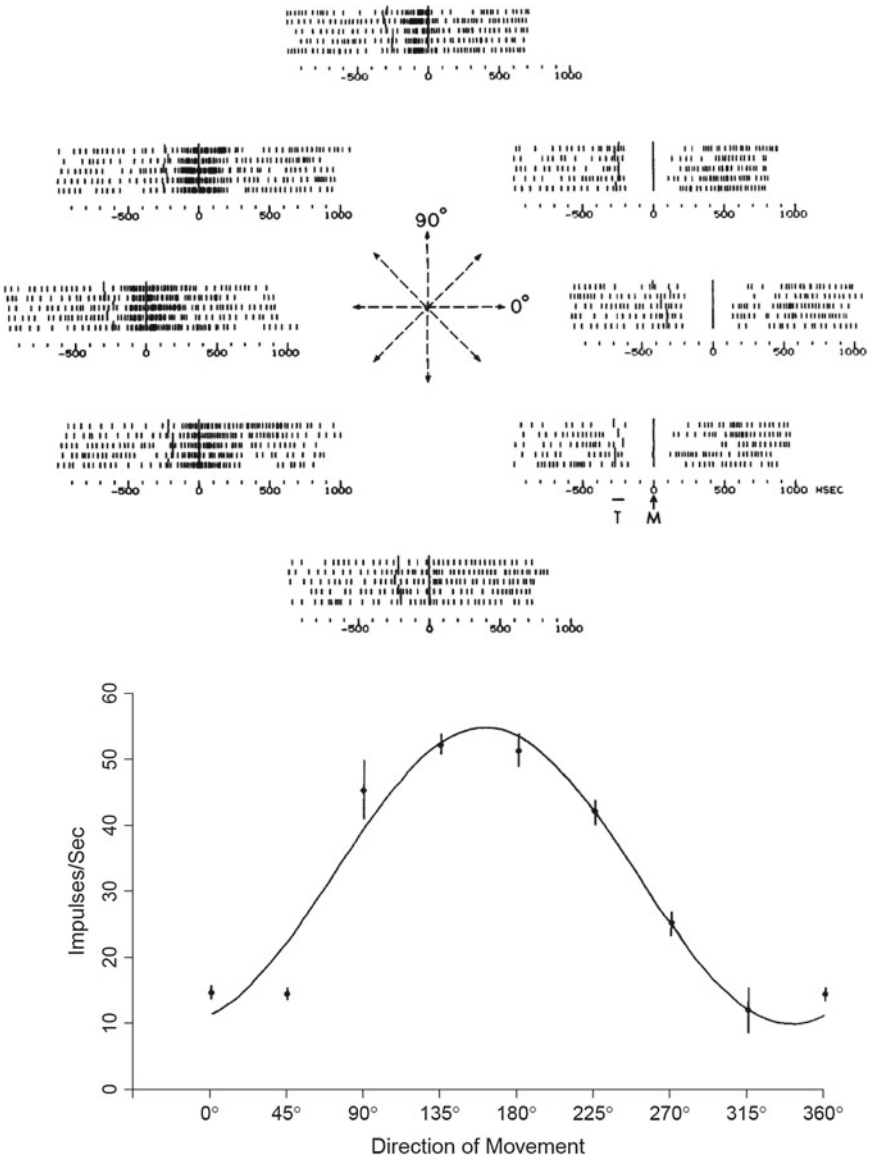

Fig. 12.12 Directional tuning of motor cortex neurons (adapted from Georgopoulos et al. 1982). *Top* displays raster plots (spike trains across five trials) for each of eight reaching directions. *Bottom* displays corresponding mean firing rates.

Taking $C_i(v)$ to be the spike count for the ith trial in direction v across a time interval of length T, the observed spike count per unit time is

$$Y_i(v) = \frac{1}{T} C_i(v).$$

and we have

$$Y_i(v) = \mu(v) + \epsilon_i(v). \tag{12.68}$$

Together, Eqs. (12.68) and (12.67) define a two-variable multiple linear regression model from which the tuning parameters may be obtained. □

12.5.5 Effects of correlated explanatory variables cannot be interpreted separately.

On p. 347 we used Example 8.2 to illustrate quadratic regression, and we then issued a note of caution that x and x^2 are often highly correlated. High correlation among explanatory variables may cause numerical and inferential difficulties. Let us first describe the numerical issue.

The least-squares solution (12.56) to Equation (12.55) results from multiplying both sides of Equation (12.55) by $(X^T X)^{-1}$, under the assumption that $X^T X$ is non-singular, i.e., that its inverse exists, which occurs when the columns of X are linearly independent (see the Appendix). Linear independence fails when it is possible to write some column of X as a linear combination of the other columns; in this case a regression of that dependent column on the other columns would produce $R^2 = 1$, i.e., perfect multiple correlation. When the columns of X are very highly correlated, even if they are mathematically linearly independent, they may be numerically essentially dependent; for example, a regression of any one column on all the others might produce R^2 that is very nearly equal to 1 (e.g., $R^2 = .999$). Because of this and related considerations the details of the methods used to compute the least-squares solution are important, as indicated in the footnote on p. 341. In the quadratic regression of Example 8.2 on p. 347, for instance, the correlation between *ISI* and its square was $r = .98$. An easy way to reduce correlation is to subtract the mean of the x variable before squaring, i.e., take $w_1 = x$ and $w_2 = (x - \bar{x})^2$. With w_1 and w_2 defined in this way for $x = ISI$ in Example 8.2 we obtained $r = -.08$. Good numerical methods use general procedures that effectively transform the x variables to reduce their correlations.

A deeper issue involves interpretation of results. The potential confusion caused by correlated explanatory variables may be appreciated from the following concocted illustration.

Illustration: Quadratic regression To demonstrate the interpretive subtlety when explanatory variables are correlated we set $x = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$ and then defined

$$y_i = x_i + u_i$$

Table 12.3 Quadratic regression results for the artificial data in the illustration.

Variable	Coefficient	SE	t_{obs}	p-value
(Intercept)	−2.4	2.5	−.95	.37
x	1.86	1.04	1.8	.12
x^2	−.067	.092	−.73	.487

where $u_i \sim N(0, 4)$. We then defined w_1 and w_2 using (12.65) and (12.66) and regressed $y = (y_1, \ldots, y_n)$ on both w_1 (representing x) and w_2 (representing x^2). We obtained the results shown in Table 12.3, with $R^2 = .77$, $s = 2.1$ and $F = 11.9$ on 2 and 7 degrees of freedom, yielding $p = .0056$. From Table 12.3 alone this regression might appear to provide no evidence that y was linearly related to either x or x^2. However, regressing y on either x or x^2 alone produces a highly significant linear regression. Furthermore, the F-statistic from the regression on both variables together is highly significant. These potentially puzzling results come from the high correlation of explanatory variables: the correlation between x and x^2 is $r = .975$. Keep in mind that the t-statistic for x^2 in Table 12.3 reflects the contribution of x^2 *after* the variable x has been used to explain y and likewise the t-statistic for x reflects the contribution of x after the variable x^2 has been used to explain y. □

Let us consider this phenomenon further. Suppose we want to use linear regression to say something about the degree to which a particular variable, say x_1, explains y (meaning the degree to which the variation in y is matched by the variation in the fit of x to y) but we are also considering other variables $x_2, \ldots, x_p$. We can regress y on x_1 by itself. Let us denote the resulting regression coefficient by b. Alternatively we can regress y on $x_1, \ldots, x_p$ and, after applying Eq. (12.56), the relevant regression coefficient would be $\hat{\beta}_1$, the first component of $\hat{\beta}$. When the explanatory variables are correlated, it is not generally true that $b = \hat{\beta}_1$ and, similarly, the quantities that determine the proportion of variability explained by x_1, the squared magnitudes of the fitted vectors, are not generally equal. Thus, when the explanatory variables are correlated, as is usually the case, it is impossible to supply a unique notion of the extent to which a particular variable explains the response—one must instead be careful to say which other variables were also included in the linear regression.

This lack of uniqueness in explanatory power of a particular variable may be considered a consequence of the geometry of least squares.

> *Details:* Let us return to the geometry depicted in Fig. 12.9. As in that figure we take V to be the linear subspace spanned by the columns of X. Because the columns of X are vectors, let us write them in the form $v_1, \ldots, v_p$, and let us ignore the intercept (effectively assuming it to be zero, as we did when we related the SST decomposition to the Pythagorean theorem). The observations on the first explanatory variable x_1 then make up the vector v_1. The extent to which x_1 "explains" the response vector y now becomes the proportion of y that

lies in the direction v_1. This is the length of the projection of y onto v_1 divided by the length of y. However, length of the projection of y onto v_1 depends on whether we do the calculation using v_1 by itself or together with $v_2, \ldots, v_p$. Let us write the projection as cv_1 for some constant c. If we consider v_1 in isolation, we find

$$c = \frac{\langle v_1, y \rangle}{\langle v_1, v_1 \rangle} = b. \tag{12.69}$$

If we consider v_1 together with $v_2, \ldots, v_p$, we must first project y onto V, and then find the component in the direction v_1. The result is $c = \hat{\beta}_1$. The exception to this bothersome reality occurs when v_1 is orthogonal to the span of $v_2, \ldots, v_p$ (i.e., $\langle v_1, v \rangle = 0$ for every vector v that is a linear combination of $v_2, \ldots, v_p$). In this special case of orthogonality we have $b = \hat{\beta}$, and we regain the interpretation that there is a proportion of y that lies in the direction of v_1. Specifically, in this orthogonal case we may write the projection of y onto V as $\hat{y} = c_1 v_1 + v$ for some v in the span of $v_2, \ldots, v_p$. We then have

$$\langle v_1, \hat{y} \rangle = \langle v_1, c_1 v_1 + v \rangle = c_1 \langle v_1, v_1 \rangle$$

so that the projection is $c_1 v_1$ where

$$c_1 = \frac{\langle v_1, \hat{y} \rangle}{\langle v_1, v_1 \rangle}.$$

On the other hand, we may reconsider the value c in (12.69). Because $y - \hat{y}$ is orthogonal to V when we write

$$\langle v_1, y \rangle = \langle v_1, \hat{y} + (y - \hat{y}) \rangle$$

we have $\langle v_1, y - \hat{y} \rangle = 0$. Therefore,

$$\langle v_1, \hat{y} \rangle = \langle v_1, y \rangle$$

so, in this case, $c = c_1$. Thus, in this orthogonal case, $b = \hat{\beta}_1$. □

12.5.6 In multiple linear regression interaction effects are often important.

We saw earlier that it is possible to fit a quadratic in a variable x using linear regression by defining a new variable x^2 and then performing multiple linear regression on x and x^2 simultaneously. Now suppose we have variables x_1 and x_2. The general quadratic

in these two variables would have the form

$$y = a + bx_1 + cx_2 + dx_1^2 + ex_1x_2 + fx_2^2.$$

Thus, we may again use multiple linear regression to fit a quadratic in these two variables if, in addition to defining new variables x_1^2 and x_2^2 we also define the new variable $x_1 \cdot x_2$. This latter variable is often called the *interaction* between x_1 and x_2. To see its effect consider the simpler equation

$$y = a + bx_1 + cx_2 + dx_1x_2. \tag{12.70}$$

Here, for instance, we have $\Delta y/\Delta x_1 = b + dx_2$. That is, the slope for the linear relationship between y and x_1 depends on the value of x_2 (and similarly the slope for x_2 depends on x_1). When $d = 0$ and we graph y versus x_1 for two different values of x_2 we get two parallel lines, but when $d \neq 0$ the two lines are no longer parallel.

Interaction effects are especially important in analysis of variance models, which we discuss in Chapter 13.

12.5.7 Regression models with many explanatory variables often can be simplified.

When one considers multiple explanatory variables it is possible that some of them will have very little predictive benefit beyond what the others offer. In that eventuality one typically removes from consideration the variables that seem redundant or irrelevant, and then proceeds to fit a model using only the variables that help predict the response. When the number of variables p is small it is not difficult to sort through such possibilities quickly, but sometimes there are much larger numbers of variables, particularly if combinations of them, defining interactions as described in Section 12.5.6, are considered. In this case choosing a suitable collection of variables to fit is called the problem of *model selection*, and is based on *model comparison* procedures such as those discussed in Section 11.1.6.

Example 12.7 Prediction of burden of disease in multiple sclerosis Li et al. (2006) investigated the relationship between a measure of severity of multiple sclerosis, known as burden of disease (BOD), and many clinical assessments. The response variable, BOD, was based on MRI scans, and 18 different clinical measurements were used as potential explanatory predictors, including such things as disease duration, age at onset, and symptom types, as well as an important variable of interest the Expanded Disability Status Scale (EDSS). One of their main analyses examined data from an initial set of 1,312 patients who had been entered into 11 clinical trials in multiple centers. The problem they faced was to determine the variables to use as predictors from among the 18, together with possible interactions. Note that there are $\binom{18}{2} = 153$ possible pairwise interaction terms. □

There is a huge literature on model selection in multiple regression. We very briefly describe the ideas behind a few of the major methods, and then offer some words of caution.

Let us begin with variables $x_1, x_2, \ldots, x_p$ and the aim of selecting some subset that predicts the response y well. Here, some of the x variables could be defined as interaction terms. For example, if we had variables $x_1, \ldots, x_k$ and wanted to consider all possible interaction effects, as defined in Section 12.5.6, then we would end up with $p = \binom{k}{2}$ variables in total. A very simple variable-selection algorithm is as follows:

1. Regress y on each single variable x_i and find the variable x_a that gives the best prediction (using R^2).
2. Regress y on all two-variable models that include x_a as one of the variables and find the variable x_b such that x_a together with x_b gives the best prediction.
3. Continue in this way: for $k \geq 3$ and some set of variables we label $x_{a_1}, x_{a_2}, \ldots, x_{a_{k-1}}$ that have already been selected in previous steps, consider all regression models that include, in addition, each of the remaining variables; find x_j such that (1) $x_{a_1}, x_{a_2}, \ldots, x_{a_{k-1}}, x_j$ gives the best prediction and (2) the coefficient of x_j is statistically significant.

   Note that criterion (2) provides a way of stopping the process with $k < p$.

This algorithm is an example of *forward selection*. It is also called a *greedy* algorithm (because at every step in the process it is taking an apparently best next step). In the form given above it is not yet completely specified because the level of significance, or the value of the t-ratio, must be chosen; this will determine the number of variables k that are selected. It is also possible to reverse the process by starting with a regression based on all variables $x_1, \ldots, x_p$ and then choosing, analogously to step 1 above, one variable to drop, and then repeatedly finding variables to drop until a satisfactory model is found in which all variables are statistically significant. This is called *backward elimination*. An algorithm that alternates between forward and backward steps is called *stepwise regression*.

Within model selection algorithms, including forward selection, backward elimination, or stepwise regression, it is also possible to use criteria such as AIC and BIC (see Section 11.1.6) to evaluate alternative regression models. (In regression, AIC is very similar to another popular criterion known as *Mallow's* C_p.) In principle, one would examine all possible models and pick the one that is optimal with respect to the chosen criterion, such as AIC. However, because each variable may be either included in a model, or excluded from the model, there are 2^p possible models and it quickly becomes prohibitive to examine all possible models as p grows. Model selection algorithms, therefore, provide search strategies but can not guarantee that the optimal model is found.

Example 12.7 (continued) In their study, Li et al. used a stepwise procedure based on AIC to select variables for predicting BOD. □

An additional, widely-used criterion for model selection is *cross-validation*. The idea begins by considering the prediction of y by each model. Let us define an observation from all the variables $x_1, \ldots, x_p$ to be a vector x. Then we are predicting y by some function $f(x)$. In the case of linear regression,

$$f(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

where each model fixes some of the coefficients β_j to be 0 (these are the coefficients corresponding to variables excluded from the model). The corresponding theoretical problem is to predict Y by some function $f(x)$ of a random vector X, and we may evaluate the prediction using mean squared error (MSE), $E((Y - f(X))^2)$. According to the prediction theorem on p. 89 the MSE is minimized by the conditional expectation $E(Y|X = x)$, and we would, in principle, find this conditional expectation through model selection and fitting. One possibility would be to attempt to choose the model that gives the smallest MSE. However, because the MSE will depend on unknown values of the coefficients, we must estimate it from the data. If we use the same data both to fit models and to evaluate how well the models fit, we necessarily obtain an overly optimistic answer for the MSE: we will have optimized the fit for the particular data values at hand; if we were to get new data we probably would not do as well. In other words, the estimated MSE will tend to be too small; it will be downwardly biased. Furthermore, the amount of downward bias in the estimated MSE will vary with the model, so the estimated MSE will not be a reliable model comparison procedure.

Cross-validation attempts to get around the problem of optimistic MSE assessment by splitting the n observations y_i into a set of K groups, each group having the same number of observations, or nearly the same number. Let us label the kth group G_k. Then, for $k = 1, \ldots, K$, we pick group G_k and call its observations "test data" and the remainder of the observations "training data." We use the training data to fit models and we use the test data to evaluate the fits. Specifically, an observation $y_i \in G_k$ is predicted by the fit from the training data in the $K - 1$ groups containing all $y_i \notin G_k$. Letting $\hat{y}_{i,CV}$ denote the fit of y_i based on the training data that excludes group G_k, the cross-validated estimate of MSE is

$$\widehat{MSE} = \frac{1}{n} \sum_{k=1}^{K} \sum_{y_i \in G_k} (y_i - \hat{y}_{i,CV})^2.$$

This represents the quality of "out of sample" fit; conceptually, MSE is the average squared error we would expect, theoretically, if we were to apply the fit on entirely new data collected under precisely the same conditions. The model with the best cross-validation performance $\widehat{MSE}$ is the model selected by *K-fold cross-validation*. Cross-validation should, in principle, provide good estimates of MSE as K gets large (so that the estimates of MSE will have good statistical properties). For any given

sample size n the largest possible value of K is $K = n$. This results in *leave-one-out cross validation*, a method recommended by Frederick Mosteller and John Tukey in an influential book (Mosteller and Tukey 1968). Here is an example.

Example 12.8 Prediction of fMRI face selectivity using anatomical connectivity Saygin et al. (2011) used anatomical connectivities established from diffusion-weighted imaging to predict differential responses to faces and objects in fMRI. It is highly intuitive that functional activity in the brain, as measured by fMRI, should depend on anatomical structure. Saygin et al. examined fMRI responses in the fusiform face area of the temporal lobe, an area known to respond more strongly when a subject is shown pictures of faces than when the same subject is shown pictures of objects. They considered the response to pictures of faces, and to objects, at every voxel in the fusiform face area and took as their y_i variable in regression analyses the normalized ratio of face response to object response for voxel i. The x_i vector of variables was made up of connectivities to 84 brain regions, which were found using diffusion weighted imaging. This constituted their "connectivity" model. Leave-one-out cross-validation was used across 23 subjects to compare this model with two other models that did not involve connectivity information. One model defined the x_i variables to be physical distances to the 84 brain regions. This was the "distance" model. The other used the group average among all the other subjects, as a single predictor x_i. This was the "group average" model. For each subject the authors fit these models to the other 22 subjects, then used the fits to predict the fMRI responses among all the voxels for each subject. These authors used mean absolute error instead of MSE. (We comment on this below.) Thus, they computed the sample mean absolute error across all voxels for each subject. The cross-validated estimate of mean absolute error was the sample mean[15] of these 23 values. The results were as follows: connectivity model, .65; distance model, 1.06; group average model, .78. This provided evidence that the connectivity model predicts fMRI activity better than either physical distances or group averaged responses. □

In some problems it is computationally expensive to obtain n distinct fits, one for each of the n training data sets needed for leave-one-out cross-validation. In such cases, K is chosen to be much smaller, so that only K fits need to be computed. The most popular value in this context is $K = 10$.

Cross-validation has been studied extensively (see Efron 2004; Arlot and Celisse 2010; and references therein). The argument that cross-validation should provide a correction for a downwardly biased estimate of MSE is reminiscent of the motivation for AIC given in Section 11.1.6. There, AIC was introduced to correct the bias in estimating the Kullback-Liebler discrepancy between fitted model and true model. In

[15] In K-fold cross-validation it is tempting to regard the average of the n MSE estimates as an ordinary mean, and to apply the usual standard error formula (7.17). This does not work correctly, however, because the n separate evaluations are not independent. Instead, the square of the standard error in (7.17) is an underestimate of the variance. In fact, it is not possible to provide a simple evaluation of the uncertainty attached to the cross-validation estimate of MSE, or risk (see Bengio and Granvalet 2004).

regression, minimizing the Kullback-Liebler discrepancy corresponds to minimizing MSE and, for large samples, AIC and leave-one-out cross-validation agree (Stone 1974). The great advantage of cross-validation is that it furnishes an estimate of MSE even if the relationship between Y and X does not follow the assumed linear model. On the other hand, if the linear model assumptions are roughly correct then AIC tends to outperform cross-validation (Efron 2004).

Let us make two additional remarks. First, we phrased our comments above in terms of MSE but, more generally, cross-validation provides an estimate of risk (see p. 102) using loss functions other than that defined by squared error. In Example 12.8 absolute error was used. Second, cross-validation is not a substitute for replication of experiments. Experimental replication provides much stronger evidence than any statistical manipulation can create: new data will inevitably involve both small and, sometimes, substantial changes in details of experimental design and data collection; to be trustworthy, findings should be robust to such modifications and should therefore be confirmed in subsequent investigations.

There is a different approach to the problem of using multiple regression in the presence of a large number of possible predictor variables. Instead of thinking that some variables are irrelevant, and trying to identify and remove them, one might say that the coefficients are noisy and, therefore, on aggregate, likely to be too large in magnitude. This suggests reducing the overall magnitude of the coefficients, a process usually called *shrinkage*. We replace the least squares criterion (12.45) with

$$\sum_{i=1}^{n}(y_i - \hat{y}_{i,p})^2 = \min_{\beta^*}\left(\sum_{i=1}^{n}(y_i - y_i^*)^2 + \lambda\,\text{magnitude}(\beta^*)\right) \tag{12.71}$$

where magnitude(β) is some measure of the overall size of β and is called a *penalty*. The number λ is an adjustable constant and is chosen based on the data, often by cross-validation (or, for some penalties, AIC or BIC). The criterion to be minimized in (12.71) is *penalized least squares* and the solution $\hat{y}_{i,p}$ is called *penalized regression*. The two most common penalties are

$$\text{magnitude}(\beta) = \sum_{j=1}^{p}\beta_j^2 \tag{12.72}$$

and

$$\text{magnitude}(\beta) = \sum_{j=1}^{p}|\beta_j|. \tag{12.73}$$

These penalties are also called, respectively, $L2$ and $L1$ penalties.[16] In the statistics literature $L2$-penalized regression is often called[17] *ridge regression* and $L1$-penalized regression is called the *LASSO* (see Tibshirani 2011, and references therein). We give a Bayesian interpretation of penalized regression in Section 16.2.3.

Example 12.9 MEG source localization In Example 1.2 we described, briefly, the way MEG signals are generated and detected, and we discussed an application in Example 4.7. There are 306 sensors and the sensor data may be analyzed directly or, alternatively, an attempt may be made to identify the brain sources that produce the sensor signals, a process known as *source localization*. One class of methods overlays a large grid of possible sources on a representation of the cortex, and then applies Maxwell's equations in what is known as a "forward solution" that predicts the sensor signals for any particular set of source activities. This results in a linear model of the form (12.53) where X is determined by Maxwell's equations and β represents the source activity. A typical number of sources might be 5,000, so this becomes a large problem. Furthermore, because $n = 306$ we have $p > n$ which makes the matrix $X^T X$ singular (non-invertible) and some alternative to least squares must be used. The most common solutions involve $L2$ and $L1$ penalized least squares,[18] which are used in the *minimum norm estimate* MNE and *minimum current estimate* MCE methods of source localization in MEG. □

12.5.8 Multiple regression can be treacherous.

Multiple linear regression is a wonderful technique, of wide-ranging applicability. It is important to bear in mind, however, the cautions we raised in the context of simple linear regression, especially in our discussion of Fig. 12.5. With many explanatory variables, the inadequacies of the linear model illustrated in Fig. 12.5 could be present for any of the y versus x_j relationships, for $j = 1, \ldots, p$, and there are similar but more complex possibilities when we use the multiple variables simultaneously. Furthermore, it is no longer possible to plot the data in the form y versus x when $x = (x_1, x_2, \ldots, x_p)$ and $p > 2$. The assumption of linearity of the relationship between y and x is crucial, and with multiple variables it is difficult to check.

An additional issue involves one of the most useful features of multiple regression, that it allows an investigator to examine the relationship of y versus x while adjusting for another variable u. This was discussed in Section 12.5.1 and its use in the interpretation of neural data was described in Examples 12.4 and 12.1. In this context, however, the phenomenon of attenuation of correlation, discussed in Section 12.4.4,

[16] The penalty in (12.72) may also be written $\text{magnitude}(\beta) = ||\beta||^2$ and in mathematical analysis the Euclidean length is called an $L2$ norm. The penalty (12.73) is called an $L1$ penalty because it is based, analogously, on the $L1$ norm.

[17] Strictly speaking ridge regression refers to $L2$-penalized regression after the x variables are normalized.

[18] Actually, the penalty is applied to weighted least squares as described on p. 345.

must be considered. In Example 12.4, for instance, the authors wanted to examine the effect of age on BOLD activity while adjusting for task performance. The variables used for adjustment were accuracy (x_2) and mean reaction time (x_3). For each subject, the numbers x_2 and x_3 obtained for these variables were based on limited data and therefore represent accuracy and reaction time with some uncertainty, which could be summarized by standard errors. These standard errors were not reported by the authors, and probably were small, but suppose, hypothetically, that the x_2 and x_3 measurements had large standard errors. In this case, according to the result in Section 12.5.1, the correlation of these noisy variables with BOLD activity would be less than it would have been if accuracy and reaction time had been measured perfectly. Therefore, the adjustment made with x_2 and x_3 would also be less than the adjustment that *would have been made* in the absence of noise.

A similar concern arises when the measured variables capture imperfectly the key features of the phenomenon they are supposed to represent. In Example 12.1, the authors wanted to adjust the effect of reward size on firing rate for relevant features of each eye saccade. They did this by introducing eye saccade amplitude, velocity, and latency. If, however, a different feature of eye saccades was crucial in determining firing rate (e.g., acceleration), then these measurements would only be correlated with the key feature and would represent it imperfectly. In this sense, the measured variables would again be noisy representations of the ideal variables. The fundamental issue for adjustment is whether the measured variables used in a regression analysis correctly represent the possible additional explanatory factors, which are often called *confounding* variables. We discuss confounding variables further in Section 13.4. The general problem of mismeasured explanatory variables is discussed in the statistics and epidemiology literature under the rubric of *errors in variables*. When multiple regression is used to provide statistical adjustments, the accuracy of explanatory variables should be considered.

Finally, in Section 12.5.7 we noted the many alternative regression models that present themselves when there are multiple possible explanatory variables, and we described very briefly some of the methods used for grappling with the problem of model determination. These approaches can be very successful in certain circumstances. However, there is often enormous uncertainty concerning the model that best represents the data. A careful analyst will consider whether interpretations are consistent across all plausible models. Furthermore, in assessing the relationship between the response y and one of the explanatory variables x_j, the process of model selection can spuriously inflate the magnitude of an estimated coefficient $\hat{\beta}_j$. See Kriegeskorte et al. (2010) for discussion.

Chapter 13
Analysis of Variance

Many experiments examine the effects of multiple experimental conditions. When each measured response from a subject is a single-number, the data are usually analyzed with *analysis of variance* (*ANOVA*). The name has a certain logic because, as we will see, the technique rests on a breakdown of sums of squares (assessing variation), but the null hypothesis typically takes the theoretical means to be equal among the experimental conditions, specifying no treatment effect, so that one may think of the methodology as an investigation of means. The general ideas developed in Chapters 10 and 11 carry over to ANOVA. One additional, very important notion involves the structure of the experiment. This is spelled out in Section 13.1. In Section 13.2 we indicate the way standard ANOVA models may be considered special cases of linear regression, as treated in Section 12.5. This is important conceptually and computationally. In Section 13.3 we take up nonparametric methods in ANOVA and in Section 13.4 we discuss causality and the role of randomization, which is especially relevant in clinical studies.

13.1 One-Way and Two-Way ANOVA

ANOVA can take many forms, depending on the design of the experiment and the resulting structure of the data. We consider here only the two simplest kinds of ANOVA and introduce them with a pair of examples.

Example 13.1 Stimulation and development of motor control Zelazo et al. (1972) conducted a study to see whether stimulation of infants during the first eight weeks of life could make them walk earlier. The stimulation involved a simulation of walking in which a parent held the baby in a manner that would make it respond reflexively with walking-type leg movements. The data in Table 13.1 are ages in months at which 24 infants were judged to begin walking.[1] Each 1-week-old infant

[1] For pedagogical simplicity, we wanted the number of subjects per group to be equal. This is not required for ANOVA; it merely makes things a bit easier to discuss. In the original data there were

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_13,

Table 13.1 Data from motor control experiment of Zelazo et al. (1972).

Active-exercise Group	Passive-exercise Group	No-exercise Group	8-Week control Group
9.00	11.00	11.50	13.25
9.50	10.00	12.00	11.50
9.75	10.00	9.00	12.00
10.00	11.75	11.50	13.50
13.00	10.50	13.25	11.50
9.50	15.00	13.00	12.35

Entries are ages at which each of 24 infants began walking. The treatment group is "active-exercise" and the other three groups served as controls

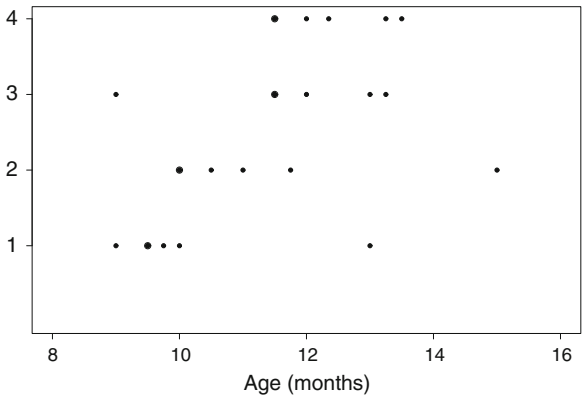

Fig. 13.1 Display of data from Table 13.1. The age of walking is shown for each of the *four* conditions, with *1* being active exercise, *2* being passive exercise, *3* being no exercise, and *4* being the 8-week control. Each larger plotted *dot* indicates the presence of 2 identical values of age within a given condition (so that for each condition there are 6 observations at 5 locations on the graph).

was assigned to one of four groups, namely, an experimental group (active-exercise) and three control groups (passive-exercise, no-exercise, 8-week control).[2] The issue is whether the active-exercise group walked earlier than the controls. From Fig. 13.1 it may be seen that the active-exercise group infants had somewhat earlier reported

(Footnote 1 continued)
only 5 subjects in the 8-week control group. We therefore added the 12.35 value to the 8-week control group.

[2] Infants in the active-exercise group received stimulation of the walking and placing reflexes during four 3-minute sessions that were held each day from the beginning of the second week until the end of the eighth week. The infants in the passive-exercise group received equal amounts of gross motor and social stimulation as those who received active-exercise, but unlike the active-exercise group, these infants had neither the walking nor placing reflex exercised. Infants in the no-exercise group did not receive any special training, but were tested along with the active-exercise and passive-exercise subjects. The 8-week control group was tested only when they were 8 weeks of age; this group served as a control for the possible helpful effects of repeated examination.

Table 13.2 Data from finger tapping experiment of Scott and Chen (1944).

Drug	Subject No.			
	1	2	3	4
Pl	11	56	15	6
Th	26	83	34	13
Ca	20	71	41	32

Entries are tapping rates. Each of 4 subjects received all 3 treatments (drugs): placebo, theobromine, and caffeine

ages of walking than those in the three control groups. However, there is quite a bit of variability, with one of the 6 infants in the active group being relatively late (13.0) and one in the no-exercise group being quite early (9.0). Thus, it's hard to tell whether there is a consistent pattern. □

Notice the layout of the data in the example above: it makes sense to display them in columns, with each column identified with a different treatment. The next example is different.

Example 13.2 Finger tapping in response to stimulants Scott and Chen (1944) conducted an experiment on finger tapping in response to orally-administered stimulants. Four subjects were each given three different treatments and then their finger-tapping rates were analyzed. The treatments were caffeine (Ca); 1-ethyltheobromine (Th: the stimulant in chocolate, similar to caffeine); and a placebo (Pl). The tapping rates (rate minus 440, with "rate" not defined but possibly taps per minute) are shown in Table 13.2.

In this case we would be interested in comparing the three treatments. The mean tapping rates for Pl, Th, and Ca are 22, 39, and 41. Is this evidence that theobromine and caffeine led to increased tapping rates? □

An important distinction between the two experiments above is that in the finger tapping experiment in Example 13.2 each subject received *all* of the treatments. Thus, the 12 data values were produced by only 4 subjects in the experiment, not 12. In the motor control experiment of Example 13.1, each subject received only one treatment, and the 24 data values came from 24 subjects. The two situations require related but different statistical methods. Table 13.1 is sometimes called a *one-way* table and is treated by *one-way ANOVA* while Table 13.2 is called a *two-way* table and is treated by *two-way ANOVA*.

13.1.1 ANOVA is based on a linear model.

The one-way ANOVA model is

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \tag{13.1}$$

where Y_{ij} is the jth observation in the ith group, $\mu + \alpha_i$ is the mean for the ith group and ϵ_{ij} is the error for the jth observation in the ith group (the discrepancy between Y_{ij} and $\mu + \alpha_i$). Here, μ is the overall mean (the "grand mean") and α_i is the increment added to that overall mean in obtaining the mean for the ith group, so that

$$\frac{1}{I}\sum_{i=1}^{I} \mu + \alpha_i = \mu$$

and this implies

$$\sum_{i=1}^{I} \alpha_i = 0. \tag{13.2}$$

We take the number of groups to be I, so that $i = 1, 2, \ldots, I$, and write the number of observations in group i as n_i. In some places we also write the ith group mean as

$$\mu_i = \mu + \alpha_i.$$

The one-way ANOVA assumptions are

(i) the ANOVA model (13.1) holds;
(ii) the errors satisfy $E(\epsilon_i) = 0$ for all i;
(iii) the errors ϵ_i are independent of each other;
(iv) $V(\epsilon_i) = \sigma^2$ for all i (homogeneity of error variances), and
(v) $\epsilon_i \sim N(0, \sigma^2)$ (normality of the errors).

Note that these are the same assumptions as those used in linear regression (apart from the replacement of (12.5) with (13.1); see p. 315). As a result, residual analysis may be used in very much the same way as in regression. Indeed, mathematically, analysis of variance may be considered a special case of linear regression. We return to this in Section 13.2.

The purpose of this model is to provide a basis for statistical comparison of the group means $\mu + \alpha_i$. That is, we ask whether there is evidence that the means are different and, if so, we can estimate how different they are. Formally, we want to test the null hypothesis that the groups means are equal:

$$\mu + \alpha_1 = \mu + \alpha_2 = \cdots = \mu + \alpha_I.$$

The usual way the hypothesis is stated is as follows:

$$\mathrm{H}_0 : \alpha_i = 0 \tag{13.3}$$

for all i, which implies that the group means are equal. It also satisfies the condition that the grand mean μ remains the expectation of Y_{ij} under H_0.

13.1.2 One-way ANOVA decomposes total variability into average group variability and average individual variability, which would be roughly equal under the null hypothesis.

At the beginning of Section 12.5.2 we wrote the basic signal and noise decomposition for regression,

$$SST = SSR + SSE.$$

In ANOVA we decompose the variability in the data similarly into two pieces, replacing *SSR* with a treatment or "group" sum of squares SS_{group}. To test H_0 defined by (13.3) we compute a measure of the *average* amount of variability due to the groups, and an *average* amount of variability due to error, then compare these. Under the null hypothesis that the group means are equal, there should be no systematic variability due to groups, so that the variability we see in our "average variability due to groups" is the result of background variability in the measurements themselves, that is, the error variability. In other words, the average variability due to groups should be about the same size as the average variability due to error. Thus, to test H_0 we use a ratio of these measures of average variability and when the ratio is much larger than 1 there is evidence against H_0, in favor of there being differences among the groups. We first specify and illustrate the procedure and then indicate its motivation as a likelihood ratio test.

We begin with the total sum of squares

$$SST = \sum_{i,j} (y_{ij} - \bar{y}_{..})^2$$

where the double dots in the subscript on $y_{..}$ indicate that the mean is being taken over all the values of y, averaging across both rows and columns. In the infant exercise example we average across all 24 values. We also define the error (residual) sum of squares to be

$$SSE = \sum_{i,j} (y_{ij} - \bar{y}_{i.})^2$$

where the single dot in the subscript on $y_{i.}$ indicates that the mean is being taken *within* the ith group. In the infant exercise example there would be 4 means $\bar{y}_{i.}$ for $i = 1, 2, 3, 4$ and each would be an average across all 6 values in the appropriate column. The group sum of squares is then

$$SS_{group} = SST - SSE.$$

We next obtain averages of the group and error sums of squares by dividing by their respective degrees of freedom, df_{group} and df_{error}. Because of the constraint (13.2) we have $df_{group} = I - 1$ and, with n being the total number of observations, this leaves $n - 1 - (I - 1) = n - I$ degrees of freedom for error, i.e., $df_{error} = n - I$.

Table 13.3 Group means and standard deviations for the data in Example 13.1.

Group	N	Mean	St. Dev.
Active exercise	6	10.1	1.5
Passive exercise	6	11.3	1.9
No exercise	6	11.7	1.5
8-week control	6	12.35	.86

Table 13.4 Analysis of Variance table for data in Example 13.1.

Source	DF	SS	MS	F	p-value
Groups	3	15.74	5.25	2.40	0.098
Error	20	43.69	2.18		
Total	23	59.43			

The table lists each source of variability, the degrees of freedom for that source, and the sum of squares. For the groups and errors sources the mean squares (given by (13.4)) are also shown, and the F-statistic (given by (13.5)) and p-value are shown on the groups line

The resulting averages, called the *group mean square* and the *mean squared error*, are defined by

$$\begin{aligned} MS_{group} &= SS_{group}/df_{group} \\ MSE &= SSE/df_{error}. \end{aligned} \tag{13.4}$$

Finally, we obtain from these the F-ratio

$$F = MS_{group}/MSE. \tag{13.5}$$

Under the null hypothesis this ratio follows an F_{ν_1,ν_2} distribution, where $\nu_1 = df_{group}$ and $\nu_2 = df_{error}$ which is used to compute the p-value. Equations (13.4) and (13.5) should be compared with Eq. (12.49).

Note that in a certain sense "analysis of variance" is a misnomer. We are really analyzing several means, and determining whether there's evidence that they are different. However, the basic tool for doing so is a comparison of sums of squares, that is, a comparison of different sources of variability, which explains the terminology.

Example 13.1 (continued from p. 361) The means and standard deviations for the 4 groups are shown in Table 13.3, and the basic ANOVA breakdown is given in Table 13.4. The pooled standard deviation is $s = \sqrt{2.18} = 1.48$. Because $F = 2.40$ on 3 and 20 d.f. with $p = .098$ there is no evidence of any differences among the means. Although from the sample means it may appear that the mean age of walking is somewhat smaller for the first group than those for the control groups, according to the ANOVA F-test there is enough variability in the data that any differences among the means are consistent with chance fluctuation. As we mentioned on p. 361, there

are a couple of points visible in Fig. 13.1 that increase the variability and, thus, the denominator of the F-ratio. We will analyze these data further on p. 368. □

We now indicate how the F-test in (13.4) and (13.5) arises as a likelihood ratio test by considering the simpler ANOVA problem in which σ is known. Let us write the group means in the form $\mu_i = \mu + \alpha_i$. The pdf for observation y_{ij} is

$$f(y_{ij}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(y_{ij}-\mu_i)^2}{\sigma^2}}$$

and from the joint pdf

$$f(y_{11}, y_{12}, \ldots, y_{In_I}) = \prod_{ij} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(y_{ij}-\mu_i)^2}{\sigma^2}}$$

the loglikelihood function (after dropping the constant involving $\sqrt{2\pi}\sigma$) is

$$\ell(\mu_1, \ldots, \mu_I) = -\frac{1}{2\sigma^2} \sum_{i,j} (y_{ij} - \mu_i)^2. \tag{13.6}$$

Under H_0 we have $\mu_i = \mu$, for $i = 1, \ldots, I$ and the loglikelihood function becomes

$$\ell(\mu) = -\frac{1}{2\sigma^2} \sum_{i,j} (y_{ij} - \mu)^2. \tag{13.7}$$

When we maximize the loglikelihood in (13.6) we get

$$\hat{\mu}_i = \bar{y}_{i.}$$

and

$$\begin{aligned} \ell(\hat{\mu}_1, \ldots, \hat{\mu}_I) &= -\frac{1}{2\sigma^2} \sum_{i,j} (y_{ij} - \bar{y}_{i.})^2 \\ &= -\frac{1}{2\sigma^2} SSE. \end{aligned}$$

When we maximize the loglikelihood in (13.7) we get

$$\hat{\mu}_i = \bar{y}_{..}$$

and

$$\ell(\hat{\mu}) = -\frac{1}{2\sigma^2}\sum_{i,j}(y_{ij} - \bar{y}_{..})^2$$
$$= -\frac{1}{2\sigma^2}SST.$$

The log of the likelihood ratio LR in (11.6) is

$$\log LR = \ell(\hat{\mu}) - \ell(\hat{\mu}_1, \ldots, \hat{\mu}_I)$$

and multiplying this by -2, and combining with (13.7) and (13.6) after inserting the MLEs we get

$$-2\log LR = \frac{1}{\sigma^2}SST - \frac{1}{\sigma^2}SSE$$
$$= \frac{SS_{group}}{\sigma^2}. \tag{13.8}$$

From (13.8), the likelihood ratio test will reject H_0 when SS_{group} is sufficiently large relative to σ^2.

The ANOVA F-statistic (13.5) arises from[3] (13.8) when we estimate σ^2 by MSE and normalize SS_{group} by its degrees of freedom, which is done for mathematical convenience (the ratio of MS_{group} to MSE follows an F_{ν_1,ν_2} distribution).

13.1.3 When there are only two groups, the ANOVA F-test reduces to a t-test.

In the special case of only two groups with two means μ_1 and μ_2, the null hypothesis H_0: $\mu_1 = \mu_2$ may be tested with a t-test. This turns out to be equivalent to the ANOVA F test and, in fact, the square of the t-statistic is equal to the F-statistic (compare the similar statements about regression on p. 337).

Example 13.1 (continued from p. 366) From the pooled standard deviation $s = 1.48$ reported on p. 366 we get the standard error of each mean $SE = s/\sqrt{6} = .60$. Comparing the active exercise group mean with the eight-week control we have a difference of $12.35 - 10.1 = 2.25$. Using the pooled estimate s, this difference has a standard error of $SE(\bar{X}_4 - \bar{X}_1) = s\sqrt{\frac{1}{6} + \frac{1}{6}} = .853$ and the t ratio is

$$t_{obs} = 2.25/.853 = 2.6$$

[3] When σ is unknown the derivation is slightly different because σ must be included among the parameters in the loglikelihood function, so its MLE must be found and the likelihood ratio is different; but the end result is equivalent to the F-test.

analogously with Eq. (10.19). Here, however, we are using *all* the data from the 4 groups to compute s, rather than only the data from two groups we are currently comparing. Therefore, we have 20 degrees of freedom going into s and thus 20 degrees of freedom for the t-test (rather than 10 degrees of freedom if we were using only the 2 groups). We obtain $p = .017$.

An alternative analysis compares the active exercise group with the other three groups, all of which could be considered controls. In this case, we would combine the data from the 3 control groups and thereby end up with two groups: the active exercise group and a single control group, the latter now having 18 observations. We would then use the "two-sample t" analysis, as in (10.21). Carrying this out, we obtain (i) a test of the null hypothesis that the means for these two groups are equal, which we may write as H_0: $\mu_{active} - \mu_{controls} = 0$, and (ii) a 95 % CI for the difference between the means $\mu_{active} - \mu_{controls}$.

First, we find the two means and standard errors to be 10.12 ± 0.59 and $11.81 \pm .34$, which gives a t-ratio of 2.46 on 22 degrees of freedom and $p = .022$. Second, applying the formula for the 95 % CI in Eq. (7.31) we find our 95 % CI for the decrease in mean age of walking for the active group compared with controls to be (.26, 3.1) months.

The conclusions from this analysis are different from those on p. 366, based on the F-test. We summarize on p. 374. □

13.1.4 Two-way ANOVA assesses the effects of one factor while adjusting for the other factor.

On p. 363 we described the distinction between one-way and two-way tables by contrasting Examples 13.1 and 13.2. To introduce the two-way analysis let us first look further at the data in Example 13.2.

Example 13.2 (continued from p. 363) Figure 13.2 displays the tapping rates for the three drugs across the four subjects. We can see that the subjects have very different tapping rates, but for all four of them the placebo rate is noticeably lower than that obtained with theobromine or caffeine. Also, the comparison of rates for theobromine and caffeine is inconsistent across subjects. The quantitative analysis, below, will support these qualitative observations. □

The two-way ANOVA model is

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij},$$

where Y_{ij} is the observation for the ith treatment on the jth subject, $\mu + \alpha_i + \beta_j$ is its mean, and ϵ_{ij} is the error for the ith treatment and jth subject. Here, α_i is the increment added to the overall mean μ in obtaining the mean for the ith treatment while β_j is the increment added to overall mean in obtaining the mean for the jth subject. We say that α_i is the *effect* for the ith treatment and β_j the effect for the

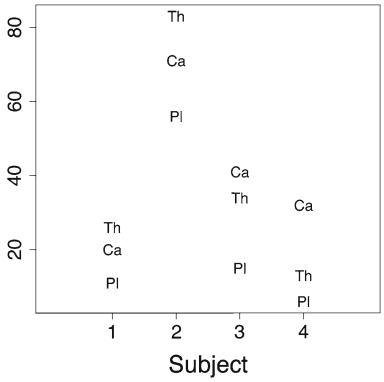

Fig. 13.2 Tapping rates displayed with identifiers "Pl" for placebo, "Ca" for caffeine, and "Th" for theobromine.

Table 13.5 Analysis of Variance table for data in Example 13.2.

Source	DF	SS	MS	F	*p*-value
Drugs	2	872	436	7.88	.021
Subjects	3	5478	1826	33	.0004
Error	6	332	55.3		
Total	11	6682			

The form of the table is similar to that in Table 13.4, except there are now F-ratios and p-values for both drugs and subjects

jth subject. A common terminology replaces the subjects with *blocks*, so that one would say β_j is the effect for the jth block. This terminology comes from the origin of ANOVA in agricultural field trials, where it referred to a block of land in a field.

As in one-way ANOVA, in two-way models the null hypothesis of interest is H_0: $\alpha_i = 0$ for all i. In the two-way case it is also possible to formulate the hypothesis that all the β_j's are zero, as well. This is not usually an object of investigation in experiments on multiple subjects because it would typically not be plausible for the subjects all to react the same way to the various treatments. However, statistics packages print out F-statistics and p-values for both hypotheses, so it's important to keep them straight (Table 13.5).

Example 13.2 (continued from p. 369) In the ANOVA for the finger tapping data there are two "factors" to be considered, drugs and subjects. Here, $F = 7.88$ on 2 and 6 d.f. with $p = .021$ indicates some evidence that the treatment means are different. There is also an F-ratio for subjects, which in fact is much larger and has a considerably smaller p-value: in this example, there is a very substantial difference among the subjects. In particular, the second subject has a much higher tapping rate than the others. The variability among subjects might be important to the conclusions one would wish to draw.

We may say something about the means, as well. For the three groups the mean tapping rates are, respectively, 22, 39, and 41. Standard errors are found by plugging in an estimate s of σ and again applying $SE = s/\sqrt{n}$. We have $s = \sqrt{MSE} = \sqrt{55.3} = 7.44$. Since there are 4 observations per treatment group, we use $n = 4$ and get 22 ± 3.7, 39 ± 3.7 and 41 ± 3.7. Clearly, the caffeine and theobromine groups have tapping rates substantially above that for the placebo group. □

13.1.5 When the variances are inhomogeneous across conditions a likelihood ratio test may be used.

The ANOVA F-test remains accurate for modest deviations from the homogeneity of variance assumption, which is assumption (iv) on p. 364. A rough rule of thumb is that as long as each ratio of pairs of standard deviations for two different groups is less than 3, the F-test should be accurate. However, in extreme cases where group i has a standard deviation σ_i that is much larger than the standard deviation σ_k for group k, there will be much more information in an observation y_{ij} about μ_i than in y_{kj} about μ_k. In such situations the usual F-statistic fails to take account of the differing contributions of data from different groups to the assessment of H_0 and it no longer has an F distribution. The problem may be fixed by re-deriving the likelihood ratio statistic and applying a permutation or bootstrap test. See Behseta et al. (2007) and references therein.

Example 4.7 (continued from p. 306) In examining directional information at each MEG brain source Wang et al. (2010) found grossly different standard deviations for the 4 different movement directions. They therefore applied the procedure of Behseta et al. (2007) to get likelihood ratio test statistics at every source and every time point. This was also used by Xu et al. (2011) within the permutation test described briefly on p. 306. □

13.1.6 More complicated experimental designs may be accommodated by ANOVA.

We have reviewed the fundamental ideas in ANOVA but have specified the procedures only in the two simplest cases involving one or two experimental factors. In many studies, especially involving human subjects, the designs can be more complicated. Sometimes they involve *multiple factors*, e.g., when there are 3 factors the analysis involves 3-way ANOVA. In Example 13.2 each subject's tapping rate was measured repeatedly, across 3 conditions. This is a special case of a *repeated measures* design. In many situations each subject is measured for all treatment conditions, but there is another factor, such as gender, that applies to groups of subjects. Such repeated-measures designs require specialized ANOVA methods. An additional possibility is that subjects, or other factors, may be considered themselves to provide an interesting

source of variation. In this case their effects may be modeled as random variables. This generates *random-effects models* and they too require specialized techniques. We discuss random-effects models briefly in Chapter 16.

13.1.7 Additional analyses, involving multiple comparisons, may require adjustments to p-values.

Because ANOVA involves comparison of several means, many possible hypotheses may be of interest.

Example 13.1 (continued from p. 368) We have already looked at the data on development of motor control in two different ways. On p. 366 we used ANOVA to test the hypothesis of no differences among the mean age of walking, H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$. Then, on p. 368, we reported two further analyses. The first used a t-test to test the null hypothesis of no difference between the active exercise group and the eight-week control group mean ages of walking, H_0: $\mu_1 = \mu_4$ with a t-test. The second used a t-test to test the null hypothesis of no difference between the mean age of walking in the active exercise group and that in the three control groups combined, H_0: $\mu_1 = \frac{1}{3}(\mu_2 + \mu_3 + \mu_4)$. We also could have singled out the other control groups and tested H_0: $\mu_1 = \mu_2$ and H_0: $\mu_1 = \mu_3$. Furthermore, because the p-value quantifies the rarity, or surprise, of the results, we ought to ask what other results *might have been* as surprising as those we actually observed. What if the passive exercise group had produced apparent earlier walking, similar to the active exercise group, by comparison with the eight-week control group? Wouldn't that have been a result we would have found interesting? Once we admit that this, too, would have been reported as a finding, then we realize that we were, effectively, testing many possible null hypotheses. The problem of testing multiple hypotheses was discussed in Section 11.3. □

As illustrated in Example 13.1, above, ANOVA often generates many plausible null hypotheses and, in this context, the problem of multiple hypothesis testing is also called the problem of *multiple comparisons*. In Section 11.3 we presented the Bonferroni correction, which can be applied when the number of comparisons (null hypotheses) is easily enumerated. We commented that the Bonferroni method is conservative, in the sense of yielding adjusted p-values that sometimes seem unnecessarily large, making it relatively difficult to obtain statistically significant results. This has spawned a large literature on multiple comparison procedures, most of which aim to provide smaller p-values under specific circumstances, so that it becomes easier to declare statistical significance. For example, a method due to Dunnett assumes there is a single control group with mean μ_c and considers all null hypotheses of the form H_0: $\mu_i = \mu_c$, for $i \neq c$. When there are I means, there are $I - 1$ such null hypotheses and, under the standard ANOVA assumptions it is possible to find an exact p-value for this case. Similarly, when there is no single control group, a method due to Tukey examines all pairs of means, i.e., all null hypotheses of the form H_0: $\mu_i = \mu_j$ for

distinct i and j. When there are I means, this narrows the number of hypotheses down to $\binom{I}{2}$ and, again, an exact p-value can be obtained.

We have two general comments on the problem of multiple comparisons in ANOVA. First, permutation tests discussed in Chapter 11 can be used to obtain p-values that take account of multiple testing procedures, as illustrated in Example 4.7 on p. 306. In Example 13.1, for instance, we might want to compare each of the 3 control groups to the active exercise group, using 3 t-tests. We then might focus on the t-test having the largest t-value. To obtain a p-value for this comparison we could create permutation pseudo-data and for each set of pseudo-data we could test all 3 null hypotheses of equality between mean of the active exercise group and the mean of each of the three control groups and we could store the largest of the 3 t-statistics based on the pseudo-data. A comparison of the largest t-statistic computed from the real data with those computed from the pseudo-data would give us a p-value, as in the cases examined in Section 11.2.1.

A second point is that multiple comparisons procedures in ANOVA are different than those arising in the neuroimaging of Example 11.3, which was used to motivate the multiple testing procedures discussed in Section 11.3.2. In neuroimaging there are typically thousands of null hypotheses, while in ANOVA, even when considering many possible combinations, the number is usually much smaller. The adjustments in ANOVA, including the Bonferroni correction, are therefore less severe. Importantly, when different multiple comparison methods lead to inconsistent conclusions it is an indication that the results are equivocal. In fact, in many ANOVA settings a very workable way to proceed is to begin by relying on the F test. If one obtains a significant F-statistic there is evidence for a difference among the means, and it therefore makes sense to go ahead and examine whichever means happen to look interesting, without worrying much about the process of selecting them. In other words, a widely-advocated method, sometimes called the *protected least-significant difference*, is to require a significant F statistic and then to report results from the many t tests, or any of them that seem to be of interest.

Details: A *contrast* among the means is a linear combination $\sum_i c_i \mu_i$ for which $\sum c_i = 0$. For example, when $I = 4$, the contrast vector $c = (1, -1, 0, 0)$ would define the contrast $\mu_1 - \mu_2$. Corresponding to any contrast we have the null hypothesis that the contrast is zero, i.e.,

$$H_0\colon \sum_{i=1}^{I} c_i \mu_i = 0. \tag{13.9}$$

It is possible to define a test of this null hypothesis with a p-value that adjusts for examining all possible contrasts. In other words, the null hypothesis being tested is that H_0 in (13.9) holds for all contrast vectors c. This is usually called the *Scheffé* test. In terms of linear combinations of the means, this is a maximally protective procedure: it guards against spurious results from examining all possible linear

comparisons. Under the standard assumptions, it may be shown that the F-test is significant at level α if and only if there exists a linear contrast for which a test of H_0 defined by (13.9) is significant at level α according to the Scheffé test. □

Example 13.1 (continued from p. 372) Where does all this leave us in this example? We may summarize by saying that there is some evidence, but not strong evidence, that the active group mean age of walking is a bit younger than that for the control groups. The marginal nature of this evidence becomes clear when we ignore the special feature that the latter three groups are all controls and look for differences among all four groups: we find no evidence for this, according to the F-test. Given that it may be difficult to determine exactly when a given child walks, and it is not clear that the parents made this determination in the absence of knowledge about what to expect based on the experimental hypothesis, some skepticism would seem appropriate.[4] □

13.2 ANOVA as Regression

13.2.1 The general linear model includes both regression and ANOVA models.

We now return to the matrix formulation of multiple regression, discussed in Section 12.5.3, and show how linear regression may be used to solve problems of analysis of variance. The points are, first, it can be helpful conceptually to re-frame ANOVA as regression and, second, statistical software typically does this.

ANOVA concerns the comparison of means among several groups, corresponding to experimental conditions. Let us consider two simple examples. Suppose X is the $n \times 1$ vector of 1s

$$X = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

We then compute $X^T X = n$ and $X^T Y = \sum y_i$ and find

$$(X^T X)^{-1} X^T y = \bar{y}.$$

Therefore, the sample mean may be found by applying regression with this very special version of the design matrix X.

[4] On the other hand, the paper by Zelazo et al. presented an additional measure where the results were more striking. On this subject, see Adolph (2002).

Next, consider two groups of m values $y_{11}, \ldots, y_{1m}$ and $y_{21}, \ldots, y_{2m}$, corresponding to two experimental conditions, having sample means $\bar{y}_1$ and $\bar{y}_2$. We define

$$y = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1m} \\ y_{21} \\ \vdots \\ y_{2m} \end{pmatrix} \tag{13.10}$$

and

$$X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \tag{13.11}$$

where the first column contains m rows of 1s followed by m rows of 0s and the second column contains m rows of 0s followed by m rows of 1s. The first column of X is an *indicator variable*, indicating membership in the first group, i.e., the ith element of the first column of X is 1 if the ith element of y is in the first group and is 0 otherwise. The second column of X is an indicator variable indicating membership in the second group. We compute

$$X^T X = \begin{pmatrix} m & 0 \\ 0 & m \end{pmatrix}$$

$$X^T y = \begin{pmatrix} \sum y_{1i} \\ \sum y_{2i} \end{pmatrix}$$

and

$$(X^T X)^{-1} X^T y = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \end{pmatrix}.$$

Thus, the sample means are obtained from multiple regression based on the design matrix in (13.11). In a similar manner we may use linear regression to compute means across several experimental conditions: for each condition we introduce an additional indicator variable as an additional column of the design matrix. The ANOVA from this regression becomes the same as the ANOVA table used in 1-way ANOVA. In

this case of two conditions, the regression results would be equivalent to those from a t-test, as described in Section 13.1.3.

Before leaving the subject of indicator variables, let us make the further point that there are typically many reasonable choices of the way to code the columns of the X matrix. For example, if we reconsider two groups of m values $y_{11}, \ldots, y_{1m}$ and $y_{21}, \ldots, y_{2m}$, we could take

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ \vdots & \vdots \\ 1 & -1 \end{pmatrix}. \tag{13.12}$$

In this case, X is no longer made up of indicator variables, but its columns span the same space as that spanned by the indicator variables given in (13.11). That is, a vector $\boldsymbol{v}$ is a linear combination of the columns of X using (13.12) if and only if it is a linear combination of the columns of X using (13.11), though the coefficients of the linear combinations will be different in the two cases. Another way to say this is that the space of fitted values $V = \{X\beta^*,\ \beta^* \in R^2\}$, defined in Section 12.5.3, is the same regardless of whether the design matrix X takes the form of (13.11) or (13.12). Using (13.12) we obtain

$$X^T X = \begin{pmatrix} 2m & 0 \\ 0 & 2m \end{pmatrix}$$

$$X^T Y = \begin{pmatrix} \sum y_{1i} + \sum y_{2i} \\ \sum y_{1i} - \sum y_{2i} \end{pmatrix}$$

and

$$(X^T X)^{-1} X^T Y = \begin{pmatrix} \bar{y} \\ (\bar{y}_1 - \bar{y}_2)/2 \end{pmatrix}$$

where $\bar{y}$ is the overall mean. The second component $(\bar{y}_1 - \bar{y}_2)/2$ is often called a *contrast*, because it is "contrasting" the means of the groups. Generally speaking, a contrast vector (leading to a contrast estimate) is one whose components add to zero; see the discussion surrounding (13.9). In ANOVA settings, where there are multiple groups, it is often of interest to define an X matrix made up of contrast vectors, together with the vector 1_{vec} whose components are all equal to 1.[5]

A different way to represent ANOVA data is also useful, especially with statistical software. The input to software is typically a vector of data, such as represented

[5] It is also convenient to require the vectors to be orthogonal to one another, in which case they are called *orthogonal contrasts*. For orthogonal contrasts, each estimate is independent of the others. This is a topic discussed in many books on regression analysis and experimental design.

in (13.10), and the software must be informed which observations correspond to different groups. In conjunction with the data in (13.10) we define

$$L = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 2 \\ 2 \\ \vdots \\ 2 \end{pmatrix} \tag{13.13}$$

where the first m rows are 1s and the last m rows are 2s. The values 1 and 2 in the vector L in (13.13) are called the *levels* of the conditions or factor. In the case of the finger tapping data in Example 13.2 we could define $y = (11, 26, 15, 6, 26, 83, 34, 13, 20, 71, 41, 32)^T$ and then set

$$L = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 2 & 1 \\ 2 & 2 \\ 2 & 3 \\ 2 & 4 \\ 3 & 1 \\ 3 & 2 \\ 3 & 3 \\ 3 & 4 \end{pmatrix} \tag{13.14}$$

so that the first column of the level matrix L represents the "levels" of the drugs (1 for Placebo, 2 for Theobromine, 3 for Caffeine) and the second column represents "levels" of the subjects (1 for first subject, etc.). Statistical software used for 1-way or 2-way ANOVA requires some identifier of group structure, such as (13.13) and (13.14). It is possible to produce a design matrix X from a level matrix L, and vice-versa. ANOVA software often provides functions for this purpose.

13.2.2 In multi-way ANOVA, interactions are often of interest.

In Section 12.5.6 we described the way interactions between explanatory variables arise in multiple regression. Interactions play an important role in many ANOVA settings. Here we consider the simplest case of interactions between two conditions that each have two levels and then connect the ANOVA and regression contexts.

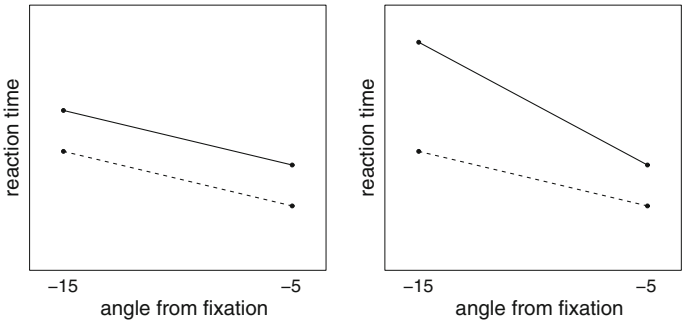

Fig. 13.3 Hypothetical plots of mean saccadic reaction time when angular distance from fixation to target is either −15 or −5 degrees, i.e., when the eyes fixate either 15 or 5 degrees to the *right* of the target. *Solid lines* correspond to patients; *dashed* correspond to controls. In the *left plot* the lines are parallel, indicating the reaction time is longer among patients by the same amount for both angular distances; there is no interaction between angular distance and subject classification. In the *right plot* the increase reaction time among patients is greater at −15 degrees than at −5 degrees, so the lines are no longer parallel; this represents an interaction between angular distance and subject classification.

Example 2.1 (continued) In the experiment on saccadic reaction time, Behrmann et al. (2002) sought to characterize the way eye saccades differed among patients with hemispatial neglect compared with control subjects.[6] We use this context to illustrate presence and absence of interaction. Let Y be saccadic reaction time, x_1 represent the distance from eye fixation to target, measured in degrees of angle to the right. When the target was on the left side of fixation, which was the neglected side for the patients, the angle was negative. We let $x_1 = 1$ when the target was at -15 degrees (15 degrees to the left of fixation) and $x_1 = 0$ when the target was at -5 degrees. We also let x_2 be an indicator variable indicating patients, i.e., $x_2 = 1$ for patients and $x_2 = 0$ for control subjects. These variables define 4 mean saccadic reaction times: μ_{11} is the mean reaction time among patients when the target was at -15 degrees; μ_{10} is the mean reaction time among controls when the target was at -15 degrees; μ_{01} is the mean reaction time among patients when the target was at -5 degrees; and μ_{00} is the mean reaction time among controls when the target was at -5 degrees. If patients and controls reacted similarly, except that patients had a fixed latency of response, then the means would satisfy

$$H_0\colon \mu_{11} - \mu_{10} = \mu_{01} - \mu_{00} \tag{13.15}$$

which is the null hypothesis of no interaction. The left side of Fig. 13.3 displays a possible set of four means satisfying H_0 in (13.15). On the other hand, if the patients also moved their eyes more slowly then their mean response would be even longer at -15 than at -5, and we would have

[6] The purpose of the study was to distinguish responses based on eye-centered coordinates, head-centered coordinates, and trunk-centered coordinates.

$$\mu_{11} - \mu_{10} > \mu_{01} - \mu_{00},$$

as shown on the right side of Fig. 13.3. The second case, but not the first, corresponds to the presence of an interaction effect between x_1 and x_2. Statistical evidence of an interaction effect would be found by obtaining a statistically significant interaction of x_1 and x_2. □

In Section 12.5.6 we said that in regression based on explanatory variables x_1 and x_2 the variable defined as the product x_1x_2 represents the interaction between these variables. In the equation

$$y = a + bx_1 + cx_2 + dx_1x_2, \tag{13.16}$$

which was Eq. (12.70), we noted that when $d = 0$ the graphs of y versus x_1 for two different values of x_2 produce two parallel lines, but when $d \neq 0$ the two lines are no longer parallel. Figure 13.3 displays an example of this phenomenon. In ANOVA the variables correspond to the experimental design, as outlined briefly in Section 13.2.1, and interaction effects are found via least-squares regression.[7] We omit details. Here is a neuroimaging example.

Example 13.3 Neural correlates of delay of gratification Successful decision making often requires an ability to forgo immediate gain in favor of increased future reward. Casey et al. (2011) reported fMRI results for group of individuals who had been studied 40 years earlier, as preschool children, for their ability to delay gratification. Previously it had been shown that performance on a delay-of-gratification task during childhood predicted ability to perform on a go/no-go task as adults. The authors imaged their subjects during go/no-go tasks. One of their findings involved the inferior prefrontal gyrus, an area thought to be involved in impulse control during similar tasks. Based on the childhood results, the authors categorized the subjects has either "low" or "high" childhood ability to delay gratification. The question was whether the two groups had different neural activity in the inferior prefrontal gyrus 40 years later, and the experimental prediction was that in the low ability group neural activity in the inferior prefrontal gyrus would be similar on go and no-go trials, but for the high ability group there would be much stronger activity on no-go trials (when impulse control is operative) than on go trials. This corresponds to an interaction between trial type ("go" vs. "no-go") and subject group (low or high childhood ability). Let us write the means of the neural activity in go and no-go trials[8] for the low and high ability groups as $\mu_{\text{go}}^{\text{low}}$, $\mu_{\text{nogo}}^{\text{low}}$, $\mu_{\text{go}}^{\text{high}}$, $\mu_{\text{nogo}}^{\text{high}}$. The null hypothesis of no interaction would be

[7] ANOVA may also be applied, as a special case of regression, when one explanatory variable is quantitative and another variable is an ANOVA indicator variable. This is usually called *analysis of covariance* or ANCOVA. Its purpose is to adjust the ANOVA for effects of the quantitative variable. See p. 332.

[8] We are here simplifying by ignoring some aspects of the experimental design.

$$H_0\colon \mu_{\text{nogo}}^{\text{low}} - \mu_{\text{go}}^{\text{low}} = \mu_{\text{nogo}}^{\text{high}} - \mu_{\text{go}}^{\text{high}}.$$

Casey et al. found evidence against H_0, reporting a statistically significant interaction ($p = .014$) between trial type and subject group. □

In Example 13.3 it was hypothesized that for one group of subjects (the low ability group) the means under the two conditions (μ_{go} and μ_{nogo}) would be very close in magnitude while for the other group (the high ability group) they would be quite different. It would be tempting to test $H_0 : \mu_{\text{go}} = \mu_{\text{nogo}}$ for each of the two groups: if the test were significant for the second group but not for the first group one might then conclude that the two groups were different with regard to the two conditions. In fact, such reasoning is common in neuroscience and psychology (see Nieuwenhuis et al. 2011). Unfortunately, it is not correct. As pointed out in Section 10.4.8, a non-significant test does not itself provide evidence for H_0. Thus, in particular, a non-significant test of $H_0 : \mu_{\text{go}} = \mu_{\text{nogo}}$ does not provide evidence that the two means are approximately the same. Instead, a confidence interval or test for the interaction effect should be reported, as in Example 13.3.

13.2.3 ANOVA comparisons may be adjusted using analysis of covariance.

In comparing results under two or more experimental conditions it often happens that the subjects (or other experimental units) are not comparable with respect to some background variable, often called a *covariate*. For instance, suppose we have data under two conditions as in (13.10). As indicated in Section 13.2.1, the two means $\bar{y}_1$ and $\bar{y}_2$ may be compared by performing the regression of y on the X matrix given by (13.11), producing results that are equivalent to a t-test (and a t-based confidence interval). Now suppose we have an additional covariate u with values given by

$$u = \begin{pmatrix} u_{11} \\ \vdots \\ u_{1m} \\ u_{21} \\ \vdots \\ u_{2m} \end{pmatrix}. \tag{13.17}$$

If we regress y on both X and u we will obtain a comparison between the means under the two conditions *after adjusting for* the covariate u. As explained at the beginning of Section 12.5, this is a consequence of the regression formulation.

Example 13.4 Improving Working Memory in Children with ADHD Deficits in working memory (WM) are associated with ADHD. Klingsberg et al. (2005)

reported results of a randomized, controlled double-blind trial aimed at assessing the possible benefits of a computerized training program aimed at improving WM. (The virtues of randomized, double-blind trials are discussed briefly in Section 13.4.) The training program consisted of at least 25 sessions, each lasting roughly 40 minutes, in which subjects completed WM tasks. In the experimental condition the difficulty of the WM tasks was automatically adjusted to match the current assessment of the subject's WM. In the control condition difficulty remained at an initial low level. A total of 42 children with ADHD (ages 7–12) were randomly allocated to one of the two conditions and completed the entire protocol.

The key outcome was "span-board" task performance, a standard assessment of visuospatial WM. This was assessed at the subject's initial visit and then twice after training had been completed: both 5–6 weeks after the initial visit and, again, 3 months subsequent to this. Baseline score at the initial visit was used as a covariate, together with age and number of days of training. The authors reported a highly significant difference between span-board task performances under the experimental and control conditions, after adjusting for the covariates, with $p = .001$ at 5–6 weeks post initial visit and $p = .002$ at the second visit 3 months later. This constitutes strong evidence that WM can be improved by training among ADHD children. □

The use of covariates to adjust comparisons in the context of ANOVA is usually called *analysis of covariance*.

13.3 Nonparametric Methods

ANOVA assumption (v) on p. 364, normality, is often suspect. Because ANOVA is a special case of regression and, under weak conditions, the least-squares estimates are asymptotically normal according to (12.63), the ordinary ANOVA procedures work well with large samples even for non-normal data. Sometimes, however, the sample size may be modest while the data appear grossly non-normal. In the next two subsections we discuss two approaches to ANOVA for non-normal data. The first, in Section 13.3.1, is based on *ranks*, and the idea is to replace each data value by its rank within the whole data set. Rank-based procedures remove the assumption of a specific distributional form. The second approach involves permutation and bootstrap tests, as discussed in Sections 11.2.1 and 11.2.2. We describe these very briefly in Section 13.3.2.

The body of ANOVA methods under the assumption of normality are called *parametric*, meaning that they are based on probability models characterized by a small number of parameters. The methods in Sections 13.3.1 and 13.3.2 are *nonparametric*. Please note, however, that all these procedures continue to make the more consequential assumptions of additivity and independence of the errors.

Table 13.6 Data from Frezza et al. (1990) on first-pass alcohol metabolism.

Alcoholic Women	Non-alcoholic Women	Alcoholic Men	Non-alcoholic Men
0.6	0.4	1.5	0.3
0.6	0.1	1.9	2.5
1.5	0.2	2.7	2.7
	0.3	3.0	3.0
	0.3	3.7	4.0
	0.4		4.5
	1.0		6.1
	1.1		9.5
	1.2		12.3
	1.3		
	1.6		
	1.8		
	2.0		
	2.5		
	2.9		

13.3.1 Distribution-free nonparametric tests may be obtained by replacing data values with their ranks.

To describe rank-based ANOVA we begin with an example.

Example 13.5 Alcohol metabolism among men and women Women seem to have a lower tolerance for alcohol than men, and are more prone to develop alcohol-related diseases. When men and women of the same size and history of drinking consume equal amounts of alcohol, the alcohol in the bloodstream of the women tends to be higher. In research by Frezza et al. (1990), the "first-pass" metabolism of alcohol in the stomach was studied. The data shown in Table 13.6 come from 18 women and 14 men who volunteered to be studied. Each subject was given two doses of .3 g ethanol per kilogram of body weight, one orally and one intravenously on two different days. The difference in concentrations of alcohol in the blood (at some fixed time after administration), between the intravenous dose and the oral dose, provides a measure of first-pass metabolism in the digestive system and liver; this defines the response variable in the table, with units in mmols per liter per hour. If first-pass metabolism were more effective in men than women, the difference in levels following intravenous and oral administration would tend to be higher among men.

We begin by ignoring the distinction between alcoholic and non-alcoholic subjects. This reduces the data to two groups: women and men. The data in Table 13.6 are strikingly skewed toward high values. One possibility would be transform the data and apply the usual t-test. Instead, we describe a rank-based analysis.

Table 13.7 Data from Table 13.6 together with corresponding ranks, where the smallest observation has rank 1 and the largest has rank $n = 32$.

Case	Difference	Female	Rank
1	0.6	1	8.5
2	0.6	1	8.5
3	1.5	1	14.5
4	0.4	1	6.5
5	0.1	1	1.0
6	0.2	1	2.0
7	0.3	1	4.0
8	0.3	1	4.0
9	0.4	1	6.5
10	1.0	1	10.0
11	1.1	1	11.0
12	1.2	1	12.0
13	1.3	1	13.0
14	1.6	1	16.0
15	1.8	1	17.0
16	2.0	1	19.0
17	2.5	1	20.5
18	2.9	1	24.0
19	1.5	0	14.5
20	1.9	0	18.0
21	2.7	0	22.5
22	3.0	0	25.5
23	3.7	0	27.0
24	0.3	0	4.0
25	2.5	0	20.5
26	2.7	0	22.5
27	3.0	0	25.5
28	4.0	0	28.0
29	4.5	0	29.0
30	6.1	0	30.0
31	9.5	0	31.0
32	12.3	0	32.0

The data are printed out again in Table 13.7, with each rank listed at the end. The rank goes from 1 up to 32, with the smallest value getting the rank 1 and the largest value getting the rank 32. Ranks ending in .5 represent ties, i.e., cases in which some data value appears twice. The women in the study have a 1 in the "females" column. □

Rank-sum methods compare the ranks of the two groups. That is, if one group has values of its ranks that are sufficiently much larger than those of the other group, there will be evidence that the means of the two groups are different. More specifically, we

Table 13.8 Four observations from Table 13.7.

Case	Difference	Female	Rank
1	0.6	1	1
18	2.9	1	3
19	1.5	0	2
32	12.3	0	4

may find the sum of the ranks from one of the groups and see whether it is either much larger or much smaller than would be expected if, in fact, the two groups followed the same distribution. Based on the null hypothesis that the probability distributions for the two groups are the same, we can get a p-value. The test statistic W is the sum of the ranks from one of the two groups. This is the *rank-sum test*. It is sometimes called the Wilcoxon rank-sum test, and it is also often called the Mann-Whitney test. Let us write the distribution functions for males and females as $F_{\text{males}}(x)$ and $F_{\text{females}}(x)$. The rank-sum test tests the null hypothesis

$$H_0\colon F_{\text{males}}(x) = F_{\text{females}}(x)$$

for all x.

To be specific about the procedure, suppose the alcohol metabolism data consisted only of the four observations in Table 13.8. In this case we would rank the data as 1, 3, 2, 4 (0.6 is the smallest, 2.9 is the third smallest, 1.5 is the second smallest, and 12.3 is the fourth smallest). Then we would add up the values of the ranks for the females to get the statistic $W = 1 + 3 = 4$.

Example 13.5 (continued) For the data in Table 13.7 we obtained the rank-sum test statistic $W_{obs} = 330$ with $p = .0002$. This may be compared with the usual t-based method gave $T_{obs} = 3.41$ with $p = .0042$. In this case, we get similar conclusions and are reassured that the assumption of normality is not crucial. In fact, if we first transform the data by taking logs, the usual t-test gives $p = .0002$. □

An analogous procedure for several groups is called the *Kruskal-Wallis test*. It may be used in place of the usual F-statistic from an ANOVA.

Example 13.5 (continued) When all four groups are used and the data are transformed by logs we find $p = .003$ from the usual ANOVA F-test. In fact, the residual analysis for the log-transformed data looks pretty good and we would find little reason to worry about the assumption of normality. However, using the Kruskal-Wallis test we get $p = .002$, which again corroborates the conclusion.

In using this example to describe rank-based methods we have concentrated on technique, but a more basic concern lurks here: we must wonder about the extent to which the volunteers represent the population as a whole, and whether the particular men and women in the study might for some reason self-select in a manner that was related to their alcohol metabolism. We return to such considerations in Section 13.4. □

13.3.2 Permutation and bootstrap tests may be used to test ANOVA hypotheses.

In Section 11.2 we described how permutation and bootstrap tests may be used as alternatives to the t-distribution for computing a p-value in order to test H_0: $\mu_1 = \mu_2$ based on data involving sample sizes n_1 and n_2. The essential method was to (i) merge the data, then (ii) repeatedly resample the $n_1 + n_2$ data values, putting them arbitrarily into groups of size n_1 and n_2 to create pseudo-data, (iii) to each pseudo-data pair of samples apply the t-statistic, and finally (iv) see what proportion of the pseudo-data give t-statistic values greater than that observed in the real data. When the sampling is done without replacement the method is a permutation test, and with replacement it becomes a bootstrap test.

For one-way ANOVA the procedure is exactly analogous. For instance, with 3 conditions we would have data with sample sizes n_1, n_2, and n_3; we would follow step (i) then in (ii) resample the $n_1 + n_2 + n_3$ data values and put them into groups of sizes n_1, n_2, n_3; in (iii) we would get the F-statistic, and likewise in (iv) we would see what proportion of the pseudo-data F values exceed the F obtained for the real data.

Two-way ANOVA is more complicated because the two-way structure must be respected, but the concept is the same. See Manly (2007).

13.4 Causation, Randomization, and Observational Studies

13.4.1 Randomization eliminates effects of confounding factors.

Most studies aim to provide causal explanations of observed phenomena. To claim causality, investigators must argue that alternative explanations of an observed relationship are implausible.

Example 13.6 IQ and breast milk Lucas et al. (1992) obtained IQ test scores from 300 children who had been premature infants and initially fed milk by a tube. The children were 8 years old when they took the IQ test. The milk they had been fed by tube was either breast milk or prepared formula, or some combination of the two. Of interest was the relationship between IQ test scores and the proportion of milk the infants received that was breast milk. The amount of breast milk a baby had drunk was determined by whether or not the mother wished to feed the infant by breast milk, and how much milk the mother was able to express. □

In Example 13.6, immediately we must be aware of possible *confounding factors*. The decision to administer the treatment, i.e., to use breast milk or not, was the mother's; whatever might determine that decision *and also be related to subsequent IQ* would affect the observed relationship between IQ and consumption of breast

Table 13.9 Regression results from Lucas et al. (1992).

Explanatory variable	Estimated coefficient	p-Value
Social class	−3.5	.0004
Mother's education	2.0	.01
Female or not	4.2	.01
Days of ventilation	−2.6	.02
Received breast milk or not	8.3	<0.0001

The increase in IQ after adjusting for the other variables was 8.3 points (with $p < 0.0001$)

milk. If, for example, mothers who chose to breast feed were also more likely to provide intellectual stimulation to their young children, then the decision to breast feed could appear to raise IQ even though it was the increased stimulation that had the greater impact. The study would be free of these concerns if babies instead received a randomly-determined percentage of breast milk, but few mothers would give up this decision in order to be part of a scientific investigation.

Example 13.6 (continued) In an attempt to control confounding factors, and to reduce variability and make the comparisons more sensitive, the researchers performed a regression that included characteristics of both the mothers and the babies: social class (ordered from 1 to 5 with 5 being highest), mother's education (ordered from 1 to 5 with 5 being highest), whether or not the child was a female (1 if female, 0 if male), the number of days of ventilation of the baby after birth, and whether or not the baby received any breast milk (1 if yes, 0 if no). The results of the regression are shown in Table 13.9.

Let us begin by interpreting the main finding. If we hold fixed social class, mother's education, sex of the baby, and days of ventilation, there is a highly significant effect of whether or not the baby received breast milk, with breast milk increasing subsequent IQ, on average, by 8.3 points. This is quite a large effect. If it were felt appropriate to generalize from these data to the population at large, this effect would certainly be something the pediatric professions would pay attention to.

Should we believe that early consumption of breast milk would tend to increase IQ in the general population? □

To analyze the possibility of confounding factors it is useful to introduce some terminology and list some basic points.

In both experiments and observational studies, we are typically interested in effects of some explanatory variable or treatment on a response variable. A study is called an *experiment* when it imposes treatment conditions on some subjects; measurements on that subject are called the *response variable*. On the other hand, *observational studies* examine relationships between response variables and potential explanatory variables, which could become treatments, but there is no active administration of a treatment. A *confounding factor* (or *confounding variable*) is one that affects both the response variable and an explanatory variable; its effects on the response can not be distinguished from the effects of the explanatory variable of interest on the response.

The particular subjects being experimented upon may have special characteristics that make them different than those about which one may wish to draw conclusions. In many situations, carefully designed experiments can avoid these difficulties. *Randomization*, meaning the random allocation of the treatment to the subject provides a way of avoiding confounding variables; *double-blind* experiments can avoid hidden biases in the response measurements. It is also important to keep in mind that response variables and explanatory variables may not accurately capture what they are purported to be measuring. Strict adherence to the experimental *protocol* can also help avoid mismeasured variables. More generally, errors that can result from failure to adhere to protocol have been emphasized by Simmons et al. (2011).

Well-designed, randomized experiments can support causal explanations for associations between response and explanatory variables. More specifically, based on a well-designed experiment, it may be possible to say that, up to some degree of statistical uncertainty (represented by a standard error or confidence interval), a response will on average increase or decrease by a particular amount when an explanatory variable changes its value by some number of units (including being present rather than absent, as is the case for typical treatments).

In fact, it is possible to define a *causal effect*, and the corresponding association effect that would be observed in data. There is then a theorem saying that in a randomized experiment the causal effect is equal to the association effect (e.g., Wasserman (2004, Chapter 16)). In other words, for a randomized experiment, association *is* causation (see Section 12.4.2).

13.4.2 Observational studies can produce substantial evidence.

Although it is preferable to have data from a well-designed randomized experiment, there are situations in which it is impossible to randomly assign subjects to treatments. For example, one can not tell people whether they will be in "smoking" or "nonsmoking" groups. Still, very convincing evidence can accumulate from observational studies—as in fact has happened in the case of smoking. Several observed patterns may increase the plausibility of an explanatory variable as a cause of a response variable:[9]

- The explanatory variable or treatment precedes observation of the response, and in terms of timing can thus act as a cause.
- Large effects are observed; this makes it less likely that the association is due to a confounding variable. One often-cited example is that mortality due to scrotum cancer among chimney sweeps was about 200 times above the population levels early in the 20th century.

[9] A widely-cited source for many of these ideas is Hill (1971).

- A quantitative "dose-response" relationship is observed, in which an increase in the explanatory variable increases (or decreases) the observed response, as opposed to simply an observation of an effect when a treatment is applied versus not applied.
- There is physiological evidence to support a theory that could explain the putative causal relationship.
- There are no anomalous results that seem difficult to explain; anomalous results may signal the presence of confounding variables.
- Similar results are obtained under differing experimental studies; confounding variables are often less likely to be present in each of the different studies.

Example 13.6 (continued) Now, let us reexamine the IQ and breast milk results with these principles in mind. First, the study is prospective, in the sense that children received some percentage of breast milk and then were followed over time to see what IQ score they got many years later. Second, the estimated effect is reasonably large—8 IQ points is about half of a standard deviation in the population as a whole. Third, there is physiological relevance: pediatricians recommend that mothers breast-feed their babies for nutritional reasons. We have not done a careful review of the literature, however, and do not have the expertise to comment critically on this basic scientific issue.

Concerning the dose-response relationship, in the regression reported above the breast milk variable merely indicates whether or not the infant received breast milk; but the authors reported a similar regression using instead *percentage* breast milk where the regression coefficient was .09, which says that holding the same variables fixed, for every 10% increase in breast milk the subsequent IQ would go up on average by nearly a full point. This last result is important: by removing the decision of whether or not to use breast milk as an explanatory variable, the confounding variables associated with that decision are no longer a concern.[10] Now we must shift to the question of whether some confounding variables may affect both the amount of milk a mother can express and the subsequent IQ of the child. If not, we would be regarding the percentage breast milk actually delivered as if it were a randomly-determined percentage. One possible confounding variable would be the health of the mother during pregnancy: mothers who are unable to express much milk might conceivably have been providing worse nutrition to the fetus.

As far as anomalous results are concerned, here are two possibilities: first, given the other variables, subsequent IQ decreases as social class increases, which is surprising; second, given the other variables, female babies have higher subsequent IQs. There should be explanations for these outcomes. Otherwise, they raise doubts.[11]

Overall, from the report of this study we have given here, there is clearly a substantial association between increased administration of breast milk and increased

[10] We are here assuming that the reported regression is not being driven primarily by inclusion of lots of babies with zero percent breast milk, but rather holds among the non-zero percentage babies.

[11] We do not have the full results when percentage breast milk is used, so we don't know whether these associations diminish or change sign in that case.

IQ, when social class (measured in the way the authors did), mother's education, and days of ventilation are held fixed. However, it remains possible that some confounding variables affect breast-milk expression and IQ. As we write this, 20 years has passed since the publication of the 1992 paper. While the topic remains controversial, subsequent research has been informative. For further information see Brion et al. (2011) and the references therein. □

Example 13.5 (continued) Returning to the alcohol metabolism example, let us now consider the possibility of confounding due to the use of volunteers in the study. The chief concern is whether volunteers are different than the rest of the population with respect to alcohol metabolism. This is at least plausible, though in order to affect the study, the volunteer men and women would have to be different. For example, if the women who volunteered tended to have trouble with alcohol metabolism (perhaps they thought the study sounded interesting because they knew they had a high susceptibility to the effects of alcohol) but men just wanted the money, then the differential effect would tend to be larger in this sample than in the population. Is this kind of hypothetical scenario reasonable, or really a stretch of the imagination? Your answer to this question determines how much faith you will put in the results. □

Chapter 14
Generalized Linear and Nonlinear Regression

Multiple linear regression is a powerful method of exploring relationships between a response Y and a set of potential explanatory variables $x_1, \ldots, x_p$, but it has an obvious limitation: it assumes the predictive relationship is, on average, linear. In addition, in its standard form it assumes that the noise contributions are homogeneous and follow, roughly, a normal distribution. During the latter part of the twentieth century a great deal of attention was directed toward the development of generalized regression methods that could be applied to nonlinear relationships, with non-constant and non-normal noise variation. In this chapter and in Chapter 15 we discuss several of the most common techniques that come under the heading *modern regression.*

We alluded to modern regression in Chapter 12 by displaying diagram (12.4),

$$Y \leftarrow \begin{cases} \text{noise} \\ f(x_1, \ldots, x_p). \end{cases}$$

To be more specific about the models involved in modern regression let us write the multiple linear regression model (12.44) in the form

$$Y_i = \mu_i + \epsilon_i \tag{14.1}$$
$$\mu_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} \tag{14.2}$$

where $\epsilon_i \sim N(0, \sigma^2)$. In (14.1) and (14.2) we are separating two parts of the model. The deviations from the mean appear in (14.1) as additive noise while, according to Eq. (14.2), the mean itself is a linear function of the x variables. Modern regression models have the more general form

$$Y_i \sim f_{Y_i}(y_i|\theta_i) \tag{14.3}$$
$$\theta_i = f(x_{1i}, \ldots, x_{pi}) \tag{14.4}$$

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_14,

where $f_{Y_i}(y|\theta)$ is some family of pdfs that depend on a parameter θ, which[1] is related to $x_1, \ldots, x_p$ according to a function $f(x_1, \ldots, x_p)$. Here, not only is $f(x_1, \ldots, x_p)$ in (14.4) allowed to be nonlinear, but also the probabilistic representation of noise in (14.3) is more general than in (14.1). The family of pdfs $f_{Y_i}(y|\theta)$ must be specified. In Sections 14.1.1–14.1.3 and 14.1.4–14.1.5 we take the response distributions in (14.3) to be binomial and Poisson, respectively, but in applying (14.4) we retain the linear dependence on $x_1, \ldots, x_p$ for suitable parameters θ_i. In Section 14.1.6 we discuss the formal framework known as *generalized linear models* that encompasses methods based on normal, binomial, and Poisson distributions, along with several others. In Section 14.2 we describe the use of nonlinear functions $f(x_1, \ldots, x_p) = f(x_1, \ldots, x_p; \theta)$ that remain determined by a specified vector of parameters θ (such as $f(x; \theta) = \theta_1 \exp(-\theta_2 x)$).

Modern regression is also used to analyze spike trains, where it becomes *point process regression*. We discuss this in Chapter 19. We lay the foundation for point process regression with our description of Poisson regression, especially in Examples 14.4 and 14.5 in Section 14.2.2.

We hope that our presentation will make the generalization of the regression framework to (14.3) and (14.4) seem straightforward. From our current perspective, it is. Historically, however, the step from least squares to generalized linear models was huge: it required not only the advent of ML estimation, but also the recognition that some widely-used probability distributions had well-behaved likelihood functions (see Section 14.1.6) together with sufficient computational power to perform the fitting in a reasonable amount of time. All of this came together in the publication Nelder and Wedderburn (1972).

14.1 Logistic Regression, Poisson Regression, and Generalized Linear Models

14.1.1 Logistic regression may be used to analyze binary responses.

There are many situations where some y should be a noisy representation of some function of $x_1, \ldots, x_p$, but the response outcomes y are binary. For instance, behavioral responses are sometimes either correct or incorrect and we may wish to consider the probability of correct response as a function of some explanatory variable or variables, or across experimental conditions. Sometimes groups of binary responses are collected into proportions.

Example 5.5 (continued from p. 226) In Fig. 8.9 we displayed a sigmoidal curve fitted to the classic psychophysical data of Hecht et al. (1942) on perception of dim light. There, each response was binary and the 50 binary responses at a given light

[1] We apologize for the double use of f to mean both a pdf in $f_{Y_i}(y|\theta)$ and a general function in $f(x_1, \ldots, x_p)$. These two distinct uses of f are very common. We hope by pointing them out explicitly we will avoid confusion.

intensity could be collected into a proportion out of 50 that resulted in perception. We fit the data by applying maximum likelihood estimation to the logistic regression model in (8.43) and (8.44). This[2] is known as *logistic regression*. □

Example 2.1 (continued from p. 378) In Section 13.2.2 we discussed ANOVA interactions in the context of the study by Behrmann et al. (2002) on hemispatial neglect, where the response was saccadic reaction time and one of the explanatory variables was angle of the starting fixation point of the eyes away from "straight ahead." A second response variable of interest in that study was saccadic error, i.e., whether the patient failed to execute the saccade within a given time window. Errors may be coded as 0 and successful execution as 1. Behrmann et al. (2002) used logistic regression to analyze the error rate as a function of the same explanatory variables. They found, for example, that the probability of error increased as eyes fixated further to the right. □

From (14.1) and (14.2) together with normality, for a single explanatory variable x, in linear regression we assume

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2).$$

There are three problems in applying ordinary linear regression with binary responses to obtain fitted probabilities: (i) a line won't be constrained to (0, 1), (ii) the variances are not equal, and (iii) the responses are not normal (unless we have proportions among large samples, in which case the proportions would be binomial for large n and thus would be approximately normal, as in Section 5.2.2). The first problem, illustrated in Fig. 8.9, is that the linear regression may not make sense beyond a limited range of x values: if $y = a + bx$ and $b > 0$ then y must become infinitely large, or small, as x does. In many data sets with dichotomous or proportional responses there is a clear sigmoidal shape to the relationship with x. The second problem was discussed in the simpler context of estimating a mean, in Section 8.1.3. There we derived the best set of weights to be used for that problem, and showed that an estimator that omits weights can be very much more variable, effectively throwing away a substantial portion of the data. Much more generally it is also possible to solve

[2] The analysis of Hecht et al. (1942) was different, but related. They wished to obtain the minimum number of quanta, n, that would produce perception. Because quanta are considered to follow a Poisson distribution, in the notation we used above, they took $W \sim P(\lambda)$ and $c = n$, with λ, the mean number of quanta falling on the retina, being proportional to the intensity. This latter statement may be rewritten in the form $\log \lambda = \beta_0 + x$, with x again being the log intensity. Then $Y = 1$ (light is perceived) if $W \geq n$ which occurs with probability $p = 1 - P(W \leq n-1) = 1 - F(n-1|\lambda)$, where F is the Poisson cdf. This is a latent-variable model for the proportional data (similar to but different than the one on p. 399). It could be fitted by finding the MLE of β_0, though Hecht et al. apparently did the fitting by eye. Hecht et al. then determined the value of n that provided the best fit. They concluded that a very small number of quanta sufficed to produce perception, but see also Teich et al. (1982).

problem (ii) by using weighted least squares, as discussed surrounding Eq. (12.64), and such solutions apply to the logistic regression setting. The third problem can make distributional results (standard errors and p-values) suspect. The method of logistic regression, which applies maximum likelihood to the logistic regression model, fixes all three problems.

The logistic regression model begins with the log-odds transformation. Recall that when p is a probability the associated *odds* are $p/(1-p)$. The number p lies in the range $(0, 1)$ while the associated odds is in the range $(0, \infty)$. If we then take logs, the number $\log(p/(1-p))$ will lie in the range $(-\infty, \infty)$, which corresponds to what we need for infinite straight lines. Therefore, instead of taking the expected value of Y to be linear in x ($E(Y_i) = \beta_0 + \beta_1 x_i$) we note that when $Y_i \sim B(n_i, p_i)$ we have $E(Y_i/n_i) = p_i$ and we apply $\log(p_i/(1-p_i)) = \beta_0 + \beta_1 x_i$. First, from the algebraic manipulations given in our discussion of Example 5.5 on p. 226, substituting z for u and w for p in (9.8) and (9.9), we have

$$z = \log\left(\frac{w}{1-w}\right) \Longleftrightarrow w = \frac{\exp(z)}{1+\exp(z)}. \tag{14.5}$$

In (14.5) we replace w with p_i and z with $\beta_0 + \beta_1 x_i$. The logistic regression model (8.43) and (8.44) may then be written in the form

$$\begin{aligned} Y_i &\sim B(n_i, p_i) \\ \log\frac{p_i}{1-p_i} &= \beta_0 + \beta_1 x_i. \end{aligned}$$

The log-odds (or *logit*) transformation is helpful in interpreting results. The log odds (of a response) are linear in x. Thus, β_1 is the change in the log odds for a unit change in x.

The log odds scale itself is a bit awkward to think about, though if the base of the logarithm is changed from e to 2 or 10 it becomes easier. It is often useful to transform back to the odds scale, where an increase of 1 unit in x is associated with an increase in the odds (that $Y = 1$) by a factor of $\exp(\beta_1)$. If we wish to interpret the change in probabilities, we must pick a particular probability p and conclude that a unit increase in x is associated with an increase from p to $expit(logit(p) + \beta_1)$, where $logit(z) = \log(z/(1-z))$ and $expit(w) = \exp(w)/(1+\exp(w))$. To illustrate, we provide some interpretation in the context of Example 5.5.

Example 5.5 (continued) On p. 213 we found $\hat{\beta}_1 = 10.7$ with standard error $SE = 1.2$. We interpret the fitted model as saying that, on average, for every increase of intensity by a factor of 10 (1 unit on the scale of the explanatory variable) there is a 10.7 ± 1.2 increase in the log odds of a response. To get an approximate 95 % CI for the factor by which the odds increase we exponentiate, $\exp(10.7 \pm 2(1.2))$, i.e., (4023, 489000). A more interpretable intensity change, perhaps, would be doubling. An increase in intensity by a factor of 2 corresponds to .30 units on the scale of the explanatory variable (because $\log_{10}(2) = .301$). For an increase of intensity by a

factor of 2 the log odds thus increase by $3.22 \pm .72$ (where $3.22 = (.301)(10.7)$ and $.72 = (.301)(2.4)$). This gives an approximate 95 % CI for the factor by which the odds increase, when the intensity doubles, of $\exp(3.22 \pm .72) = (12.2, 51.4)$.

We can go somewhat further by converting odds to the probability scale by inverting

$$\text{odds} = \frac{p}{1-p}$$

to get

$$p = \frac{\text{odds}}{1+\text{odds}}.$$

Let us pick $p = .5$, so that the odds are 1. If we increase the odds by a factor ranging from 12.2 to 51.4 then the probability would go from .5 to somewhere between .92 and .98 (where $.92 = 12.2/(1+12.2)$ and $.98 = 51.4/(1+51.4)$). Thus, if we begin at the x_{50} intensity (where $p = .5$) and then double the intensity, we would obtain a probability of perception between .92 and .98, with 95 % confidence. This kind of calculation may help indicate what the fitted model implies. □

Logistic regression extends immediately to multiple explanatory variables: for m variables $x_1, \ldots, x_m$ we write

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_m x_{mi}.$$

The multiple logistic regression model may be written in the form

$$\begin{aligned} Y_i &\sim B(n_i, p_i) \\ \log \frac{p_i}{1-p_i} &= x_i \beta \end{aligned} \tag{14.6}$$

where β is the coefficient vector and x_i is the $1 \times (m+1)$ vector of values of the several explanatory variables corresponding the ith unit under study.

14.1.2 In logistic regression, ML is used to estimate the regression coefficients and the likelihood ratio test is used to assess evidence of a logistic-linear trend with x.

It is not hard to write down the likelihood function for logistic regression. The responses Y_i are independent observations from $B(n_i, p_i)$ distributions, so each pdf has the form $\binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i}$ and the likelihood function is

Table 14.1 Linear regression results for data from subject S.S. in Example 5.5.

Variable	Coefficients	SE	t_{obs}	p-value
Intercept	−1.78	.30	−5.9	.0042
Intensity	1.20	.16	7.5	.0017

Table 14.2 Logistic regression results for data from subject S.S. in Example 5.5.

Variable	Coefficients	SE	t_{obs}	p-value
Intercept	−20.5	2.4	−8.6	$p < 10^{-6}$
Intensity	10.7	1.2	8.9	$p < 10^{-6}$

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{n_i - y_i}$$
$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

where the second equation is substituted into the first. Standard statistical software may be used to maximize this likelihood. The standard errors are obtained from the observed information matrix, as described in Section 8.3.2.

For a single explanatory variable, the likelihood ratio test of Section 11.1.3 may be used to test $H_0 : \beta_1 = 0$. More generally, if there are variables $x_1, \ldots, x_p$ in model 1 and additional variable $x_{p+1}, \ldots, x_{p+m}$ in model 2, then the likelihood ratio test may again be applied to test $H_0 : \beta_{p+1} = \cdots = \beta_{p+m} = 0$. The log likelihood ratio has the form

$$-2 \log LR = -2[\log(\hat{L}_1) - \log(\hat{L}_2)]$$

where $\hat{L}_i$ is the maximum value of the likelihood under model i. For large samples, under H_0, $-2 \log LR$ follows the χ^2 distribution with m degrees of freedom.

In some software, the results are given in terms of "deviance." The *deviance* for a given model is $-2 \log(\hat{L})$. The *null deviance* is the deviance for the "intercept-only" model, and we denote it by $-2 \log \hat{L}(0)$. Often, the deviance from the full fitted model is called the *residual deviance*. In this terminology, the usual test of $H_0 : \beta_1 = 0$ is based on the difference between the null deviance and the residual deviance.

Example 5.5 (continued) The output from least-squares regression software is given in Table 14.1. The F statistic in this case is the square of t_{obs} and gives the $p = .0017$, as in Table 14.1. The results for logistic regression are given in Table 14.2. The null deviance was 257.3 on 5 degrees of freedom and the residual deviance was 2.9 on 4 degrees of freedom. The difference in deviance is

$$\text{null deviance} - \text{residual deviance} = 257.3 - 2.9 = 256.4$$

Table 14.3 Quadratic logistic regression results for data from subject S.S. in Example 5.5.

Variable	Coefficients	SE	t_{obs}	p-value
Intercept	−4.3	15.8	−.27	.78
Intensity	−6.6	17.0	−.39	.70
Intsq	4.6	4.6	1.0	.31

Table 14.4 Quadratic logistic regression results for data from subject S.S. in Example 5.5, after first centering the intensity variable.

Variable	Coefficient	SE	t_{obs}	p-value
Intercept	−20.3	2.3	−8.7	$p < 10^{-6}$
Intensity	10.5	1.2	8.6	$p < 10^{-6}$
Int2	4.6	4.6	1.0	.31

which should be compared to the chi-squared distribution on 1 degree of freedom. It is very highly significant, consistently with the result in Table 14.2. □

Polynomial terms in x may be handled in logistic regression just as they are in linear regression (Section 12.5.4).

Example 5.5 (continued) To consider whether an additional, nonlinear component might contribute usefully to the linear logistic regression model, we may square the intensity and try including it in a two-variable logistic regression model. In this case it is interesting to note that intensity and its square are highly correlated. To reduce the correlation it helps to subtract the mean before squaring. Thus, we define intsq $=$ (intensity)^2 and int2 $=$ (intensity $-$ mean(intensity))^2. The results using the alternative variables intsq and int2 are shown in Tables 14.3 and 14.4, respectively. Using either of these two logistic regression summaries we would conclude the quadratic term does not improve the fit. The results in Table 14.3 might, at first, be confusing because of the nonsignificant p-values. As we noted in Section 12.5.5, this is a fairly common occurrence with highly correlated explanatory variables, as x and x^2 often are. Recall that each nonsignificant p-value leads to the conclusion that its corresponding variable contributes little *in addition to* the other variable. Since we already found a very highly significant logistic linear relationship, we would conclude that the quadratic doesn't improve the fit. Again, though, the interpretation appears cleaner in the second formulation. □

In non-normal regression models there is no fully satisfactory generalization of the measure of fit R^2. One useful measure, proposed by Nagelkerke (1991) and usually called the *Nagelkerke* R^2, is defined by

$$R_N^2 = 1 - \left(\frac{\hat{L}(0)}{\hat{L}} \right)^{\frac{2}{n}}$$

where, again, $\hat{L}(0)$ is the maximized likelihood for the intercept-only model and $\hat{L}$ is the maximized likelihood for the model being considered. Because the maximum value of R_N^2 may be less than 1, a scaled version is often used:

$$R^2_{\text{scaled } N} = \frac{R_N^2}{R^2_{max}}$$

where

$$R^2_{max} = 1 - \left(\hat{L}(0)\right)^{\frac{2}{n}}.$$

14.1.3 The logit transformation is one among many that may be used for binomial responses, but it is the most commonly applied.

The *expit* function $\exp(x)/(1 + \exp(x))$, defined in Section 14.1.1, is one of many possible sigmoidal curves and thus logistic regression is only one of many possible models for binary or proportion data. In fact, expit(x) has an asymptote at 0 as $x \to -\infty$ and at 1 as $x \to \infty$, and is increasing, so it is a cumulative distribution function. The distribution having expit(x) as its cdf is called the *logistic distribution*, but the cdf of any continuous distribution could be used instead. One important alternative to logistic regression is the Probit regression model, which substitutes the normal cdf in place of the *expit*: specifically, the probit model is

$$\begin{aligned} Y_i &\sim B(n_i, p_i) \\ \Phi^{-1}(p_i) &= \beta_0 + \beta_1 x_i \end{aligned}$$

where $\Phi(z) = P(Z \le z)$, with $Z \sim N(0, 1)$. The fitted curve is then obtained from $y = \Phi(\hat{\beta}_0 + \hat{\beta}_1 x)$.

Example 5.5 (continued) Figure 14.1 displays the fitted curves from probit and logistic regression for the data shown previously in Fig. 8.9. The two models produce nearly identical fitted curves. □

As with the threshold data, the fitted curves from probit and logistic regression are generally very close to each other. This is because the graph of the logistic cdf (the *expit* function) is close to the graph of the normal cdf. Two things are special about the logistic regression model. First, it gives a nice interpretation of the coefficients in terms of log odds. Second, in the logistic regression model (but not the Probit or other versions) the loglikelihood function is necessarily concave (as long as there are at least two distinct values of x). This means that there is a unique MLE, which can be obtained from an arbitrary starting value in the iterative algorithm. Logistic

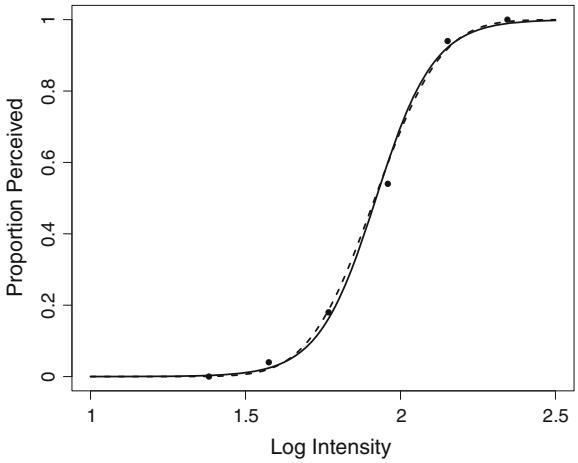

Fig. 14.1 Two curves fitted to the data in Fig. 8.9. The fitted curve from probit regression (*dashed line*) is shown together with the fitted curve from logistic regression. The fits are very close to each other.

regression is the standard method for analyzing dichotomous or proportional data, though in some contexts probit regression remains popular.[3]

An interesting interpretation of binary phenomena involves the introduction of *latent variables*, meaning random variables that become part of the statistical model but are never observed (see the illustration on p. 216 and Section 16.2). Let us discuss this in terms of perception, and let us imagine that the binary experience of perception, as "perceived" or "not perceived" is controlled by an underlying continuous random variable, which we label W. We may think of W as summarizing the transduction process (from light striking the retina to firing rate among multiple ganglion cells), so that perception occurs whenever $W > c$ for some constant c. Neither the precise meaning of W, nor the units of c need concern us. Let us take W to be normally distributed and, because the units are arbitrary, we take its standard deviation to be 1. Finally, we take this latent transduction variable, on average, to be a linear function of the log intensity of light x and we write this in the form $\mu_W = c + \beta_0 + \beta_1 x$. We now have the probit regression model: $Y = 1$ when $W > c$ but, defining $-Z = W - \mu_W$ (so that $-Z \sim N(0, 1)$ and $Z \sim N(0, 1)$),

$$W > c \iff W - \mu_W > c - \mu_W \iff -Z > c - \mu_W \iff Z < \mu_W - c.$$

In other words, $Y = 1$ when $Z < \beta_0 + \beta_1 x$, which occurs with probability $p = \Phi(\beta_0 + \beta_1 x)$.

This latent-variable interpretation helps transfer the intuition of linear regression models over to the binary case, and provides an appealing way to think about many

[3] We have not discussed residual analysis here. It may be performed using *deviance residuals,* or other forms of residuals. See Agresti (1990) or McCullagh and Nelder (1989).

Table 14.5 Spike counts from an SEF neuron during directional saccades.

left	9	6	9	9	6	6	8	5	7	9	4	8	8	3	6
Up	2	0	6	4	4	0	0	0	5	2	1	0	3	0	
Right	4	8	2	2	4	0	3	4	1	1	0	3	4	0	2
Down	1	5	1	2	0	4	4	4	4	3	6	1	1	1	

phenomena. Note that logistic regression is obtained by taking W to have a *logistic distribution*,[4] having cdf

$$F(w) = \frac{1}{1+e^{-w}}.$$

14.1.4 The usual Poisson regression model transforms the mean λ to $\log \lambda$.

The simplest distribution for counts is Poisson, $Y \sim P(\lambda)$. Here, the Poisson mean must be positive and it is therefore natural to introduce dependence on explanatory variables through $\log \lambda$. In Section 14.1.6 we will note that models defined in terms $\log \lambda$ have special properties. The usual multiple Poisson regression model is

$$\begin{aligned} Y_i &\sim P(\lambda_i) \\ \log \lambda_i &= x_i \beta \end{aligned}$$

where β is the coefficient vector and x_i is the $1 \times (m+1)$ vector of values of the explanatory variables corresponding to the ith unit under study. Poisson regression is useful when we have counts depending on one or more explanatory variables.

Example 14.1 Directional sensitivity of an SEF neuron Olson et al. (2000) reported data collected from many individually-recorded neurons in the supplementary eye field (SEF). In this experiment, a monkey was trained to translate one of four possible icons displayed at the fixation point into an instruction of a location to which he was to move his eyes: either left, up, right, or down. SEF neurons tend to be directionally sensitive. To establish direction sensitivity, Olson et al. examined the number of spikes occurring 600–750 ms after presentation of the cue. The spike count data for one neuron across the various trials are given in Table 14.5. Is this neuron directionally sensitive?

By eye it appears that the firing rate is higher for the "left" condition than for the other conditions. There are various versions of ANOVA that may be used to check this. Analysis of spiking activity from these SEF neurons revealed that while the

[4] Probit regression was introduced by Chester Bliss in 1934, but the latent variable idea and normal cdf-transformation was part of Fechner's thinking about psychophysics in 1860; logistic regression was apparently discussed first by Ronald Fisher and Frank Yates in 1938. See Agresti (1990) for much more extensive discussion of the methods described briefly here.

spike counts deviated from that predicted by a Poisson distribution, the deviation was small (Ventura et al. 2002). Here we will use the data to illustrate a version of ANOVA based on Poisson regression. Note that in Table 14.5 there are a total of 58 spike counts, from 58 trials. □

The problem of fitting counts is analogous to, though less extreme than, that of fitting proportions. For proportions, the (0,1) range could make linear regression clearly inappropriate. Counts have a range of $(0, \infty)$. Because the ordinary regression line is not constrained, it will eventually go negative. The simple solution is to use a log transformation of the underlying mean. The usual Poisson regression model is

$$Y_i \sim P(\lambda_i) \tag{14.7}$$

$$\lambda_i = \exp(\beta_0 + \beta_1 x_i). \tag{14.8}$$

To interpret the model we use the log transformation:

$$\log \lambda_i = \beta_0 + \beta_1 x_i.$$

For example, in the SEF data of Example 14.1 Y_i is the spike count and x_i is the experimental condition (up, down, left, right) for the ith trial. The advantage of viewing ANOVA as a special case of regression is apparent: we immediately generalize Poisson ANOVA by applying our generalization of linear regression to the Poisson regression model above.

14.1.5 In Poisson regression, ML is used to estimate coefficients and the likelihood ratio test is used to examine trends.

As in logistic regression we use ML estimation and the likelihood ratio test ("analysis of deviance").

Example 14.1 (continued) We perform Poisson regression using indicator variables as described in Section 13.2.1 to achieve an ANOVA-like model. Specifically, we concatenate the data in Table 14.5 so that the counts form a 58×1 vector and define a variable *left* to be 1 for all data corresponding to the left saccade direction and 0 otherwise, and similarly define vectors *up* and *right*. The results from ordinary least-squares regression are shown in Table 14.6. The F-statistic was 18.76 on 3 and 54 degrees of freedom, giving $p < 10^{-6}$. The Poisson regression output, shown in Table 14.7 is similar in structure. Here the null Deviance was 149.8 on 57 degrees of freedom and the residual Deviance was 92.5 on 54 degrees of freedom. The difference in deviances is

$$\text{null deviance - residual deviance} = 149.8 - 92.5 = 57.3$$

Table 14.6 ANOVA Results for the SEF data in Table 14.5 shown in the form of regression output.

Variable	Coefficient	SE	t_{obs}	p-value
Intercept	3.49	.26	13.2	$p < 10^{-6}$
Left	2.11	.37	5.6	$p < 10^{-6}$
Up	−.74	.21	−3.5	.0011
Right	−.52	0.15	−3.4	.0014

Table 14.7 Poisson regression results for the SEF data in Table 14.5. The form of the results is similar to that given in Table 14.6.

Variable	Coefficients	SE	t_{obs}	p-value
Intercept	1.12	.079	14.2	$p < 10^{-6}$
Left	.475	.096	4.9	3×10^{-6}
Up	−.173	.063	−2.76	.0039
Right	−.155	.052	−2.96	.0023

which should be compared to the chi-squared distribution on 3 degrees of freedom. It is very highly significant. □

In Example 14.1 the results from Poisson regression were the same as with ordinary linear regression (standard ANOVA), but the details are different. In some situations the conclusions drawn from the two methods could be different.

14.1.6 Generalized linear models extend regression methods to response distributions from exponential families.

We began this chapter by saying that modern regression models have the form given by (14.3) and (14.4), which for convenience we repeat:

$$\begin{aligned} Y_i &\sim p(y_i|\theta_i) \\ \theta_i &= f(x_i). \end{aligned}$$

The simple logistic regression model may be put into this form by writing

$$\begin{aligned} Y_i &\sim B(n_i, p_i) \\ \theta_i &= \beta_0 + x_i\beta_1 \end{aligned}$$

where

$$\theta_i = \log \frac{p_i}{1 - p_i}$$

or, more succinctly,

$$Y_i \sim B(n_i, p_i)$$
$$\log \frac{p_i}{1-p_i} = \beta_0 + x_i\beta_1.$$

Similarly, the simple Poisson regression model may be written

$$Y_i \sim P(\lambda_i)$$
$$\log \lambda_i = \beta_0 + x_i\beta_1.$$

Logistic and Poisson regression are special cases of *generalized linear models*. These generalize linear regression by allowing the response variable to follow a distribution from a certain class known as *exponential families*. They also use a *link* function that links the expected value (the mean) μ_i of the data with the linear model $\beta_0 + \beta_1 x_i$. For example, the usual link functions for binomial and Poisson data are the log odds and the log, respectively, as shown above.

Exponential families have pdfs of the form

$$f_Y(y|\eta(\theta)) = h(y)\exp(\eta(\theta)T(y) - B(\theta)). \tag{14.9}$$

For instance, in the Poisson case $Y \sim P(\lambda)$, the pdf (from Chapter 5, p. 112) is

$$P(Y = y) = \frac{1}{y!}\lambda^y e^{-\lambda}.$$

We can rewrite this in the form

$$\frac{1}{y!}\lambda^y e^{-\lambda} = \frac{1}{y!}\exp(y\log\lambda - \lambda).$$

If we let $\theta = \lambda$, $\eta(\theta) = \log\lambda$, $B(\lambda) = \lambda$, $T(y) = y$ and $h(y) = 1/y!$ we obtain (14.9). Now, with $\mu = \lambda$, if we define the link function to be

$$g(\mu) = \log\mu \tag{14.10}$$

the simple Poisson regression model becomes

$$g(\mu) = \beta_0 + \beta_1 x_i.$$

Here, the log provides the link in the sense that it is the function by which the mean is transformed before being equated to the linear model.

We may rewrite (14.9) in the form

$$f_Y(y|\eta) = h(y)\exp(\eta T(y) - A(\eta))$$

in which case $\eta = \eta(\theta)$ is called the *natural parameter* (or *canonical parameter*). In the Poisson case the natural parameter is $\log \lambda$. The logarithmic link function is thus often called *the canonical link*. In the binomial case the log odds function becomes the canonical link. The statistic $T(y)$ is *sufficient* in the sense described on p. 200. The extension to the multiparameter case, in which η and $T(y)$ are vectors, is immediate:

$$f_Y(y|\eta) = h(y) \exp(\eta^T T(y) - A(\eta)). \tag{14.11}$$

Assuming that Y_i comes from an exponential family, we obtain a generalized linear model by writing

$$g(\mu_i) = \beta_0 + \beta_1 x_i, \tag{14.12}$$

where $\mu_i = E(Y_i)$. Equation (14.10) provided an example in the Poisson case, but in (14.12) $g(\mu)$ may be any link function.

Common distributions forming exponential families include binomial, multinomial, Poisson, normal, inverse Gaussian, gamma, and beta. The introduction of generalized linear models allowed regression methods to be extended immediately to all of these families, and a multiple-variable generalized linear model may be written

$$\begin{aligned} Y_i &\sim f_{Y_i}(y_i|\eta_i) \\ g(\mu_i) &= x_i\beta \end{aligned} \tag{14.13}$$

where $f_{Y_i}(y_i|\eta_i)$ is an exponential family pdf as in (14.11), $\mu_i = E(Y_i)$, and $g(\mu)$ is the link function. The unification of mathematical form meant that implementation of maximum likelihood, and likelihood ratio tests, could use the same algorithms with only minor changes in each particular case. Furthermore, for the canonical link it turns out (under relatively mild conditions on the x and y variables[5]) that the loglikelihood function is concave so that the MLE is unique. This guarantees that the maximum of the loglikelihood function will be found by the function maximizer (using *Newton's method*, i.e., iterative quadratic approximation) beginning with any starting value, and convergence will tend to be fast. Generalized linear models are part of most statistical software.

In addition to the canonical link, several other link functions are usually available in software. For example, it is usually possible to perform binomial regression using the probit link instead of the log odds, or logit link. Similarly, a Poisson regression could be performed using the identity link so that

$$\log \lambda_i = \beta_0 + \beta_1 x_i$$

is replaced by

[5] The regularity conditions insure non-degeneracy. For example, if there is only one x variable, it must take on at least two distinct values so that a line may be fitted. The y observations also must correspond to values that are possible according to the model; in dealing with proportions, for instance, the observed proportions can not all be zero.

$$\lambda_i = \beta_0 + \beta_1 x_i.$$

Occasionally, the identity link provides a better description of the data than the canonical link, as in Example 14.3 on p. 406.

The terminology "generalized linear model" should not to be confused with "the general linear model," which is the matrix form of regression and includes ANOVA. Both have the acronym GLM. Also, the "linear" part of the terminology is misleading because the framework really includes *nonlinear* and *nonparametric* models, as well. Specifically, while linear models with the canonical link have especially nice properties, more generally in Equation (14.4) $f(x_i)$ does not need to be linear. See Examples 14.3, 14.4, and 14.5 in Section 14.2.1 and 14.2.2.

14.2 Nonlinear Regression

14.2.1 Nonlinear regression models may be fitted by least squares.

In Section 12.5.4 we pointed out that when $f(x)$ is a polynomial in x, linear regression could be used to fit a function of the form $y = f(x)$ to (x, y) data. This involved the "trick" of starting with an initial definition of x, relabeling it as x_1 and then defining the new variable $x_2 = x_1^2$, and so on for higher-order polynomials. The resulting expectation of Y,

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

followed the form required in the linear regression model. In particular, although the relationship of Y and x, on average, was nonlinear, the *coefficients* entered linearly into the model and therefore—as in any linear regression model—the likelihood equations could be solved easily by linear algebra. A similar trick was used to fit directional tuning data with a cosine function.

There are, however, many nonlinear relationships where this sort of manipulation does not apply. For example, if

$$E(Y) = \theta_1 e^{-\theta_2 x}$$

it is not possible to redefine the x variable so that the form becomes linear in the parameters. Instead, we have the *nonlinear regression model*,

$$Y_i = f(x_i; \theta) + \epsilon_i \tag{14.14}$$

$$f(x_i; \theta) = \theta_1 e^{\theta_2 x_i}. \tag{14.15}$$

Here, the usual assumption is $\epsilon_i \sim N(0, \sigma^2)$, independently (though, again, normality is not crucial).

Models of the form (14.14)–(14.15) may still be fit by least-squares and, in fact, least squares remains a special case of ML estimation. What is different is that the equations defining the least-squares solution (the likelihood equations) are no longer solved by a single linear algebraic step. Instead, they must be solved iteratively. The problem is thus usually called *nonlinear least squares*. Example 1.6 on p. 14 provided an illustration, with the nonlinear function given by (1.6) and the fit based on nonlinear squares given in Fig. 1.5.

Example 14.2 Magnesium block of NMDA receptors NMDA receptors, which are ubiquitous in the vertebrate central nervous system, may be blocked by Magnesium ions (Mg^{2+}). To investigate the quantitative dependence of NMDA-receptor currents on the concentration of Mg^{2+}, Qian et al. (2005) measured currents at various concentrations, then summarized the data using the equation

$$\frac{I}{I_0} = \frac{1}{1 + (\frac{[Mg^{2+}]}{IC_{50}})^{n_H}}$$

where the measurements are the current I and the Magnesium concentration $[Mg^{2+}]$, I_0 being the current in the absence of Mg^{2+}. The free parameters are the "Hill constant" n_H and the 50 % inhibition concentration IC_{50} (when $[Mg^{2+}] = IC_{50}$ we get $I/I_0 = .5$). The authors estimated these constants using nonlinear least squares, and they examined IC_{50} across voltages, and across receptor subunit types. □

The term "nonlinear regression" usually refers to models of the form (14.14). However, similar models may be used with binomial or Poisson responses, and may be fit using ML. The next example illustrates nonlinear regression models using both normal and Poisson distributions.

Example 14.3 Non-cosine directional tuning of motor cortical neurons Amirikian and Georgopoulos (2000) investigated cosine and non-cosine directional tuning for 2-dimensional hand movement among motor cortical neurons. In Section 12.5.4 we considered the cosine tuning model given by (12.67) and (12.68) where, according to (12.67), a neuron's firing rate $\mu(v)$ when the movement is in direction v was linear in the components v_1 and v_2 and the model could be fit using linear regression. To investigate departures from cosine tuning, Amirikian and Georgopoulos used a class of functions involving exponentials that are not amenable to reconfiguration in a linear model and, as a result, reported that the tuning curves in motor cortical neurons, for 2-dimensional hand movement, tend to be substantially narrower than cosine tuning curves.

Examples of nonlinear fits to data from two neurons are shown in Fig. 14.2. The functions fitted were

$$\mu(v) = \mu + \beta \exp(\kappa \cos(\theta - \tau + \eta \cos(\theta - \tau))) \tag{14.16}$$

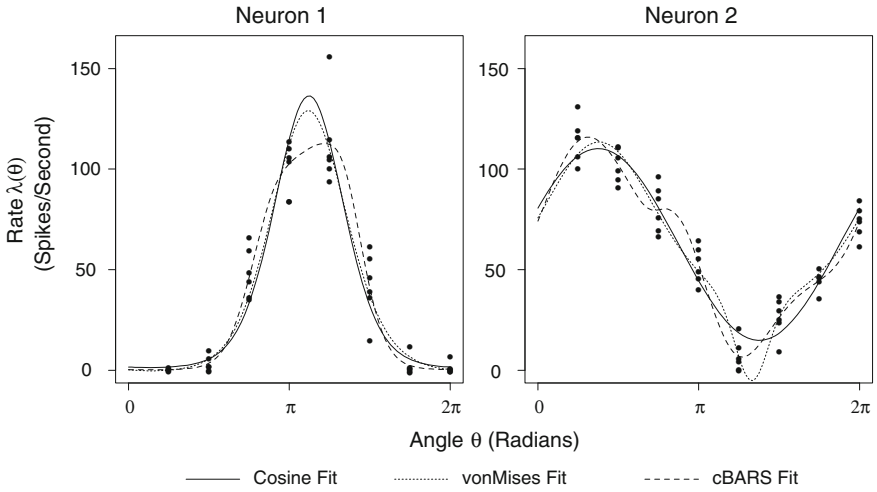

Fig. 14.2 Fits to activity of two neurons in primate motor cortex (reprinted with permission from Kaufman et al. 2005). Each datapoint represents the observed firing rate of a neuron in the motor cortex of a monkey during one repetition of a wrist movement to a particular target. The cosine fits use the cosine function in Eq. (12.67) and the von Mises fits use more complicated parametric forms given by Eq. (14.16), for Neuron 1, and Eq. (14.17) for Neuron 2. The cosine and von Mises parametric fits use Poisson maximum likelihood for Neuron 1 and least squares for Neuron 2. Also shown is the fit from a nonparametric regression method called cBARS, described by Kaufman et al. (2005).

for the first neuron, where $\theta = \arctan(v_2/v_1)$, and

$$\mu(v) = \mu + \beta_1 \exp(\kappa_1 \cos(\theta - \tau_1)) + \beta_2 \exp(\kappa_2 \cos(\theta - \tau_2)) \tag{14.17}$$

for the second neuron. These results come from Kaufman et al. (2005), who also considered nonparametric methods, discussed in Chapter 15. The function in (14.16) includes parameters corresponding roughly to the baseline firing rate, the amplitude, width, and location of the mode, and the skewness about the mode. The function in (14.17) includes parameters corresponding to two modes, one of which is constrained to be in the positive direction and the other in the negative direction. This is of use in fitting the data for the Neuron 2 in Fig. 14.2. For both neurons the data indicate mild but noticeable departures from cosine tuning.

In fact, the data in Fig. 14.2 coming from Neuron 1 exhibited roughly Poisson variation. The fits shown there were based on $Y_i \sim P(\mu_i)$ with $\mu_i = \mu(v)$ given by Eq. (14.16). This is a Poisson nonlinear regression model (with the identity link, as defined in Section 14.1.6). □

Another example of nonlinear least squares has been discussed in earlier chapters. We provide some more details here.

Example 8.2 (continued from p. 241) In presenting this example on p. 193 we said the model took Y to be the spike width and x the preceding ISI length, and assumed there was an ISI length τ such that, on average, Y is quadratic for $x < \tau$ and constant for all $x \geq \tau$. As we noted, τ is called a change point. Specifically, the statistical model was

$$Y_i \sim N(\mu(x_i), \sigma^2) \tag{14.18}$$

independently for $i = 1, \ldots, n$ where

$$\mu(x; \beta_0, \beta_1, \tau) = \begin{cases} \beta_0 + \beta_1 (x - \tau)^2 & \text{if } x < \tau \\ \beta_0 & \text{if } x \geq \tau \end{cases} \tag{14.19}$$

and the least-squares estimate $(\hat{\beta}_1, \hat{\beta}_1, \hat{\tau})$ becomes defined by

$$\sum_{i=1}^{n} \left(y_i - \mu(x_i; \hat{\beta}_0, \hat{\beta}_1, \hat{\tau}) \right)^2 = \min_{\beta_0, \beta_1, \tau} \sum_{i=1}^{n} (y_i - \mu(x_i; \beta_0, \beta_1, \tau))^2 . \tag{14.20}$$

The parameter τ enters nonlinearly into the statistical model, and this makes (14.20) a nonlinear least squares problem. However, for every value of τ we may formulate a simple linear regression problem as follows. Let us define new values $u_1(\tau), \ldots, u_n(\tau)$ by

$$u_i(\tau) = \begin{cases} (x_i - \tau)^2 & \text{if } x_i < \tau \\ 0 & \text{if } x \geq \tau \end{cases}$$

so that $\mu(x_i)$ in (14.19) may be rewritten as

$$\mu(x_i; \beta_0, \beta_1, \tau) = \beta_0(\tau) + \beta_1(\tau) u_i(\tau).$$

We then define $(\hat{\beta}_0(\tau), \hat{\beta}_1(\tau))$ by

$$\sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0(\tau) + \hat{\beta}_1(\tau) u_i) \right)^2 = \min_{\beta_0(\tau), \beta_1(\tau)} \sum_{i=1}^{n} (y_i - (\beta_0(\tau) + \beta_1(\tau) u_i))^2$$

which has the form of the simple least-squares regression problem on p. 12 and thus is easily solved. Finally, defining

$$g(\tau) = \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0(\tau) + \hat{\beta}_1(\tau) u_i) \right)^2 ,$$

the nonlinear least squares problem in (14.20) is found by minimizing $g(\tau)$. This can be achieved in software (e.g., in Matlab) with one-dimensional nonlinear minimization. Therefore, it was easy to implement nonlinear least squares for this change-point problem. □

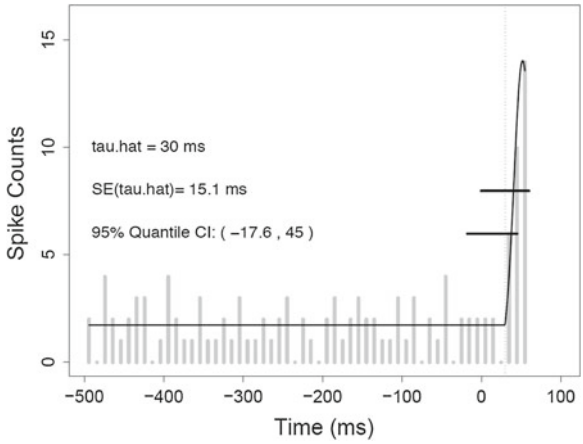

Fig. 14.3 Initiation of firing in a neuron from the basal ganglia: change-point and bootstrap confidence intervals when a quadratic model is used for the post-change-point firing rate. Two forms of approximate 95 % confidence intervals are shown. The first is the usual estimate $\pm 2SE$ interval. The second is the interval formed by the .025 and .975 quantiles among the bootstrap samples. The latter typically performs somewhat better, in the sense of having coverage probability closer to .95. See Sect. 9.2.2.

14.2.2 Generalized nonlinear models may be fitted using maximum likelihood.

Nonlinear relationships also arise in the presence of non-normal noise. We use the term *generalized nonlinear model* to refer to a model in which the linear function $g(\mu_i)$ in (14.13) is replaced by a nonlinear function. We give two examples of nonlinear Poisson regression. The first involves determination of a change-point, and is similar to Example 8.2 in Section 14.2.1.

Example 14.4 Onset latency in a basal ganglia neuron An unfortunate symptom of Parkinson's disease (PD) is muscular rigidity. This has been associated with increased gain and inappropriate timing of the long latency component of the stretch reflex, which is a muscular response to sudden perturbations of limb position. One of the important components of the stretch reflex is mediated by a trans-cortical reflex, probably via corticospinal neurons in primary motor cortex that are sensitive to kinesthetic input. To investigate the neural correlates of degradation in stretch reflex, Dr. Robert Turner and colleagues at the University of Pittsburgh have recorded neurons in primary motor cortex of monkeys before and after experimental production of PD-like symptoms. One part of this line of work aims at characterizing neuronal response latency following a limb perturbation (see Turner and DeLong 2000). Figure 14.3 displays a PSTH from one neuron prior to drug-induced PD symptoms. The statistical problem is to identify the time at which the neuron begins to increase

its firing rate, with the goal being to compare these latencies in the population of neurons before and after induction of PD.

To solve this problem we used a change-point model similar to that used in Example 8.2 on p. 408. In this case, we assume the counts within the PSTH time bins—after pooling the data across trials—follow Poisson distributions. Let Y_t be the pooled spike count in the bin centered at time t and let $\mu(t)$ be its mean. The change-point model assumes the mean counts are constant up until time $t = \tau$, at which time they increase. For simplicity, we assume the count increases as a quadratic. This gives us the Poisson change-point model

$$Y_t \sim P(\mu(t))$$

with

$$\mu(t) = \begin{cases} \beta_0 & \text{if } t \leq \tau \\ \beta_0 + \beta_1(t-\tau)^2 & \text{if } t > \tau. \end{cases}$$

The value τ is the change point. For any fixed τ the change-point model becomes simply a Poisson regression model. Specifically, for a given τ we define

$$x = \begin{cases} 0 & \text{if } t \leq \tau \\ (t-\tau)^2 & \text{if } t > \tau. \end{cases}$$

We then apply Poisson regression with the regression variable x.

However, the parameter τ is unknown and is, in fact, the object of interest. We may maximize the likelihood function iteratively over τ. That is, in software such as *R* or Matlab we set up a loop within which, for a fixed τ, we perform Poisson regression and obtain the value of the loglikelihood. We then iterate until we maximize the loglikelihood across values of τ. This gives us the MLE of τ. We may then obtain a SE for τ by applying a parametric bootstrap. Results are given in Fig. 14.3. □

Here is another example of a nonlinear model for spike counts.

Example 14.5 A Poisson regression model for a hippocampal place cell Neurons in rodent hippocampus have spatially specific firing properties, whereby the spiking intensity is highest when the animal is at a specific location in an environment, and falls off as the animal moves further away from that point (e.g., Brown et al., 1998). Such receptive fields are called *place fields*, and neurons that have such firing properties are called *place cells*. The left panel of Fig. 14.4 shows an example of the spiking activity of one such place cell, as a rat executes a free-foraging task in a circular environment. The rat's path through this environment is shown, and the location of the animal at spike times is overlain as dark dots. It is clear that the firing intensity is highest slightly to the southwest of the center of the environment, and decreases when the rat moves away from this point.

One very simple way to describe this hippocampal neural activity is to use a Poisson generalized linear model for spike counts in successive time bins while the rat forages, and to assume that the spike count depends on location in the environment

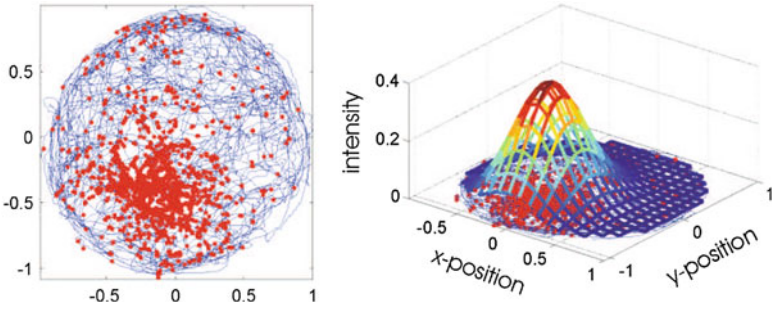

Fig. 14.4 Spiking activity of a rat Hippocampal place cell during a free-foraging task in a circular environment. *Left* Visualization of animal's path and locations of spikes. *Right* Place field model for this neuron, with parameters fit by the method of maximum likelihood.

based on a 2-dimensional bell-shaped curve. For this purpose of specifying the dependence of spiking activity on location a normal pdf may be used. Let us take $Y_t \sim P(\lambda_t)$, with t signifying time, and then define

$$\lambda_t = \exp\left\{\alpha - \frac{1}{2}\left(x(t) - \mu_x \; y(t) - \mu_y\right)\begin{pmatrix}\sigma_x^2 & \sigma_{xy}\\ \sigma_{xy} & \sigma_y^2\end{pmatrix}^{-1}\begin{pmatrix}x(t)-\mu_x\\ y(t)-\mu_y\end{pmatrix}\right\}. \quad (14.21)$$

The explanatory variables in this model are $x(t)$ and $y(t)$, the animal's x and y-position. The model parameters are $(\alpha, \mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \sigma_{xy})$, where (μ_x, μ_y) is the center of the place field, $\exp\alpha$ is the maximum firing intensity at that point, and σ_x^2, σ_y^2, and σ_{xy} express how the intensity drops off away from the center. Note that it is the shape of the place field that is assumed normal, not the distribution of the spiking activity. The right panel of Fig. 14.4 displays a fit of the place field to the data in the left panel. We will discuss models of this sort when we discuss point processes in Chapter 19. □

14.2.3 In solving nonlinear optimization problems, good starting values are important, and it can be helpful to reparameterize.

As in maximization of any likelihood, use of the numerical procedures requires care. Two important issues are the choice of initial values, and of parameterization. Both of these may be illustrated with the exponential model (14.15).

Illustration: Exponential regression To fit the exponential model (14.15) a first step is to reparameterized from θ to ω using $\omega_1 = \log(\theta_1)$ and $\omega_2 = \theta_2$ so that the expected values have the form

$$E(Y) = \exp(\omega_1 + \omega_2 x).$$

The loglikelihood is typically closer to being quadratic as a function of ω than as a function of θ. Taking logs of both sides of this expectation equation gives

$$\log E(Y) = \omega_1 + \omega_2 x.$$

This suggests we may define $U_i = \log(Y_i)$ and apply the linear model,

$$U_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \tag{14.22}$$

The resulting fitted values $\hat{\beta}_0$ $\hat{\beta}_1$ make good starting values for the iterative procedure used to obtain ω_1 and ω_2. □

It is important to recognize the distinction between the exponential model in (14.14) and (14.15) and the linearized version (14.22). Either could be used to fit data, but they make different assumptions about the way the noise contributes. In many examples, the fits based on (14.14) and (14.22) would be very close, but sometimes the resulting inferences would be different. It is an empirical question which model does a better job of describing the data. The point here, however, is that if the exponential form is preferred, the log-linear form may still be used to obtain starting values for the parameters. The linearization method of obtaining starting values is frequently used in fitting nonlinear models. (See Bates and Watts (1988) for further discussion.) These issues also arise in generalized nonlinear models.

Chapter 15
Nonparametric Regression

At the beginning of Chapter 14 we said that modern regression applies models displayed in Eqs. (14.3) and (14.4):

$$Y_i \sim f_{Y_i}(y_i|\theta_i)$$
$$\theta_i = f(x_{1i}, \ldots, x_{pi})$$

where $f_{Y_i}(y|\theta)$ is some family of pdfs that depend on a parameter θ, which is related to $x_1, \ldots, x_p$ according to a function $f(x_1, \ldots, x_p)$. In Section 14.1 we discussed the replacement of the normal assumption in (14.3) with binomial, Poisson, or other exponential-family assumptions. In Section 14.2 we showed how the linear assumption for $f(x_1, \ldots, x_p)$ in (14.4) may be replaced with a specified nonlinear modeling assumption. What if we are unable or unwilling to specify the form of the function $f(x_1, \ldots, x_p)$? In this chapter we consider fitting general functions, which are chosen to provide flexibility for fitting purposes. This is the subject of *nonparametric regression*. The terminology "nonparametric" refers to the absence of a specified parametric form, such as in (14.6) or (14.15). We focus almost exclusively on the simplest case of a single explanatory variable x, and thus consider functions $f(x)$. Here is an example.

Example 15.1 Peak minus trough differences in response of an IT neuron Some neurons in the inferotemporal cortex (IT) of the macaque monkey respond to visual stimuli by firing action potentials in a series of sharply defined bursts. Rollenhagen and Olson (2005) found that displaying an object image in the presence of a different, already-visible "flanker" image could enhance the strength of the oscillatory bursts. Figure 15.1 displays data (in the form of PSTHs) from an IT neuron under two conditions: in the first, a black patterned object was displayed as the stimulus for 600 ms; in the second condition, prior to the display of the stimulus a pair of blue rectangles appeared (as a flanker image) and these remained illuminated while the patterned-object stimulus was displayed. Overlaid on the PSTHs are fits obtained by the nonparametric regression method BARS, which will be explained briefly in Section 15.2.6. In part b of Fig. 15.1 the BARS fits are displayed together, to highlight

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_15,

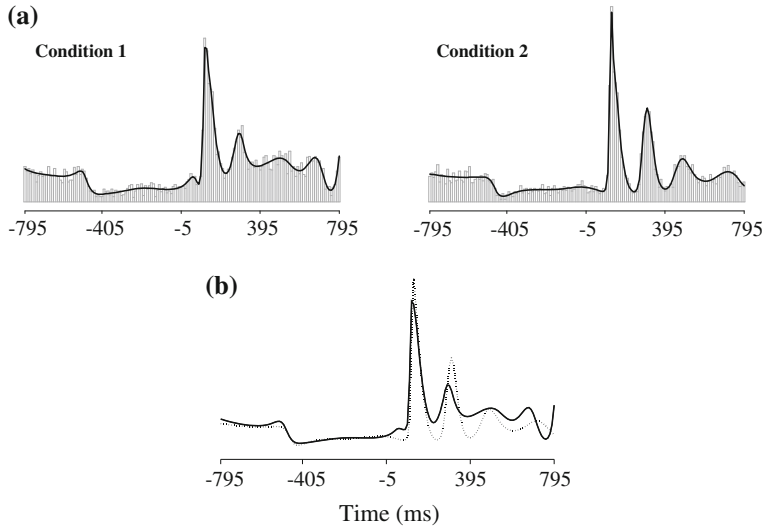

Fig. 15.1 **a** PSTHs and BARS fits for an IT neuron recorded by Rollenhagen and Olson (2005) under two conditions. **b** The two BARS fits are overlaid for ease of comparison. See text for explanation. Adapted from DiMatteo et al. (2001).

the differential response. One way to quantify the comparison is to estimate the drop in firing rate from its peak (the maximal firing rate) to the trough immediately following the peak in each condition. Let us call these peak minus trough differences, under the two conditions, ϕ^1 and ϕ^2. BARS was used to propagate the error (see DiMatteo et al. 2001). The results, for this neuron, were $\hat{\phi}^1 = 131.8(\pm 4.4)$, $\hat{\phi}^2 = 181.8(\pm 20.4)$ spikes per second, and $\hat{\phi}^1 - \hat{\phi}^2 = 50.0(\pm 20.8)$ spikes per second (where parenthetical values are *SE*s). □

There are two general approaches to nonparametric regression. The first attempts to represent a function $f(x)$ in terms of a set of more primitive functions, such as polynomials, which are often called *basis functions*. The methods following the second approach estimate $f(x)$ by weighting the data (x_i, y_i) according to the proximity of x_i to x, a process called *local fitting*. We take up these two topics in Sections 15.2 and 15.3. The fitted values $\hat{y}_i = \hat{f}(x_i)$ produce fitted points $(x_i, \hat{y}_i)$ which collectively become a *smoothed* version of the original data points. Thus, the nonparametric regression algorithm that is applied to the data is often called a *smoother*. The problem of smoothing (x_i, y_i) data to obtain a curve $y = \hat{f}(x)$ is also called *curve-fitting*.

15.1 Smoothers

As always, we are concerned with the use of statistical models both to generate estimates of scientifically interesting quantities and to provide measures of uncertainty. For both purposes we need to begin by defining the quantities we want to know

about. In linear regression and generalized linear models, and in their nonlinear counterparts, these are usually coefficients or simple functions of them such as x_{50} in Example 5.5 of Chapter 9, where we discussed propagation of uncertainty. With nonparametric regression the trick is to phrase inferential problems in terms of the function values themselves, which avoids any reference to a specific functional form. In fact, x_{50} in Example 5.5 could be considered an example of this, because even if some other function (some nonlinear, nonparametric function) were used to link log odds of perception with light intensity, that function would necessarily define a value x_{50} of the intensity at which the probability of perception would be 50 %.

A variety of nonparametric regression methods have been proposed. Some are linear and some nonlinear in a sense spelled out in Section 15.1.1.

15.1.1 Linear smoothers are fast.

We say that a nonparametric regression method results from a *linear smoother* if the fitted function values $\hat{f}(x_i)$ are obtained by linear operations on the data vector $y = (y_1, \ldots, y_n)^T$, that is, if we can write

$$(\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_n))^T = Hy \tag{15.1}$$

for a suitable matrix H. In other words, according to (15.1), for these linear smoothers, each fitted value is a linear combination of the data values y_i. The only nonlinear smoothing method we mention is that used in Example 15.1, BARS.

Because the multiplication in (15.1) involves relatively few arithmetic operations, linear smoothers are fast. They are therefore advantageous especially for large data sets, where computational speed becomes important.

15.1.2 For linear smoothers, the fitted function values are obtained via a "hat matrix," and it is easy to apply propagation of uncertainty.

The matrix H in (15.1) is called the *hat matrix*, because it produces estimates denoted with "hats." For example, in linear regression we have

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

(see Chapter 12) so that

$$\begin{aligned}(\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_n))^T &= X\hat{\beta} \\ &= X(X^T X)^{-1} X^T y\end{aligned}$$

and the hat matrix is $H = X(X^TX)^{-1}X^T$. In the case of linear regression we are able to propagate uncertainty using the distribution of $\hat{\beta}$ (as we did, similarly, for logistic regression in Chapter 9), but we could instead propagate the uncertainty from the distributions of the fitted values $X\hat{\beta}$: we simply need the variance

$$\begin{aligned} V((\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_n))^T) &= HV(Y)H^T \\ &= \sigma^2 HH^T. \end{aligned} \tag{15.2}$$

In the case of linear regression this simplifies because (as is easily checked) $H^T = H$ and $HH^T = H$ so that

$$V((\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_n))^T) = \sigma^2 H.$$

For linear smoothers more generally, $H \neq HH^T$ but, in the case of data for which $V(Y_i) = \sigma^2$ with the Y_is being independent of each other, the variance formula (15.2) continues to hold, and it remains easy to apply propagation of uncertainty. In other words, even though we do not have an estimated parameter vector, such as $\hat{\beta}$, from which to compute quantities of interest and their SEs, we can often compute quantities of interest directly from the fitted values, as in the peak minus trough example above, and can then obtain SEs from the variance formula (15.2) together with the large-sample result that the fitted values are approximately normally distributed. Similarly, when linear smoothing methods extend to logistic or Poisson regression it again remains easy to propagate uncertainty.

15.2 Basis Functions

Suppose $f(x)$ is a continuous function on an interval $[a, b]$. A famous theorem in mathematical analysis, the Weierstrass Approximation Theorem, says that $f(x)$ may be approximated arbitrarily well by a polynomial of sufficiently high order. One might therefore think that polynomials could be effective for curve fitting. That is, we could try to fit an unknown function $y = f(x)$ by instead fitting a pth order polynomial

$$y = b_0 + b_1 x + b_2 x^2 + \cdots + b_p x^p,$$

which we can do using least squares, as described in Section 12.5.4. It turns out that polynomials do not perform as well as the theoretical result might suggest. As illustrated in Fig. 15.2, even a twentieth-order polynomial can fail to represent adequately a relatively well-behaved function in the presence of minimal noise. The idea of replacing $f(x)$ with a set of simple functions, however, is very powerful. In the case of polynomials, for data $(x_1, y_1), \ldots, (x_n, y_n)$ we could fit a quadratic using (12.65) and (12.66) and regressing $y = (y_1, \ldots, y_n)$ on w_1 and w_2, and we could similarly define higher-order terms up to

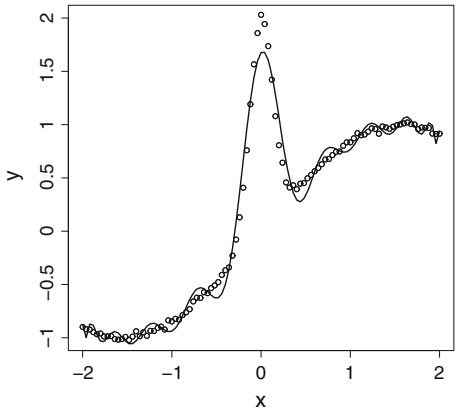

Fig. 15.2 Data simulated from function $f(x) = \sin(x) + 2\exp(-30x^2)$ together with twentieth-order polynomial fit (shown as line). Note that the polynomial is over-fitting (under-smoothing) in the relatively smooth regions of $f(x)$, and under-fitting (over-smoothing) in the peak. (In the data shown here, the noise standard deviation is 1/50 times the standard deviation of the function values.)

$$w_p = \begin{pmatrix} x_1^p \\ x_2^p \\ \vdots \\ x_n^p \end{pmatrix} \tag{15.3}$$

and could regress $y = (y_1, \ldots, y_n)$ on $w_1, w_2, \ldots, w_p$. This is an example of regression using basis functions.

The "basis function" terminology comes from the conception that the theoretical functions $f(x)$ that are, in principle, to be fitted make up an infinite-dimensional vector space for which the chosen simple functions (such as polynomials), form[1] a *basis* (see Section A.9 of the Appendix). In practice we use data $(x_1, y_1), \ldots, (x_n, y_n)$ to fit only the values $(f(x_1), f(x_2), \ldots, f(x_n))$ and thus we have an n-dimensional

[1] In Section A.9 of the Appendix we give the definition of a basis for R^n, which is an n-dimensional vector space. The basis function terminology refers to an extension of this idea to infinitely many dimensions: the functions $f(x)$ on an interval $[a, b]$ that satisfy

$$\int_a^b f(x)dx < \infty$$

(here the Lebesgue integral is used) form an infinite-dimensional vector space and if the functions $B_j(x)$ form a basis then every $f(x)$ may be written as

$$f(x) = \sum_{j=1}^{\infty} c_j B_j(x).$$

vector space for which n vectors, defined by the simple functions (such as those in (12.65), (12.66), up through (15.3) with $p = n$), form a basis.

A pth order polynomial regression will work well for functions $y = f(x)$ that look a lot like pth order polynomials. The inability of a 20th-order polynomial to fit the function in Fig. 15.2 is an indication that the function is different than a 20th-order polynomial. The challenge of nonparametric regression using basis functions is to find simple alternatives to polynomials that are flexible enough to fit a variety of functions with relatively few terms.

15.2.1 Splines may be used to represent complicated functions.

The problem in Fig. 15.2 is that the function $f(x)$ is not very close to being a low-order polynomial. In particular, it has a different form near $x = 0$ than it does as the magnitude of x increases. A possible solution here, and in other problems, is to glue together several pieces of polynomials. If the pieces are joined in such a way that the resulting function remains smooth, then it is called a *spline*. We will discuss cubic splines. Let $[a, b]$ be an interval and suppose we have values $\xi_1, \xi_2, \ldots, \xi_p$, where $a < \xi_1 < \xi_2 < \cdots < \xi_p < b$. There are then $p + 2$ subintervals $[a, \xi_1], [\xi_1, \xi_2], \ldots, [\xi_{p-1}, \xi_p], [\xi_p, b]$. A function $f(x)$ on $[a, b]$ is a *cubic spline* with *knots* $\xi_1, \xi_2, \ldots, \xi_p$ if $f(x)$ is a cubic polynomial on each of the $p+2$ subintervals defined by the knots such that $f(x)$ is continuous and its first two derivatives $f'(x)$, and $f''(x)$ are also continuous. This restriction of continuity, and continuity of derivative, applies at the knots; in between the knots, each cubic polynomial is already continuous with continuous derivatives. A cubic spline is shown in Fig. 15.3, and the result of fitting a cubic spline to the data of Fig. 15.2 is shown in Fig. 15.4. In contrast to the 20th order polynomial in Fig. 15.2, the cubic spline in Fig. 15.4 fits the data remarkably well.

15.2.2 Splines may be fit to data using linear models.

It is easy to define a cubic spline having knots at $\xi_1, \xi_2, \ldots, \xi_p$. Let $(x - \xi_j)_+$ be equal to $x - \xi_j$ for $x \geq \xi_j$ and 0 otherwise. Then the function

$$\begin{aligned} f(x) = {} & \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \\ & + \beta_4 (x - \xi_1)_+^3 + \beta_5 (x - \xi_2)_+^3 + \cdots + \beta_{p+3} (x - \xi_p)_+^3 \end{aligned} \tag{15.4}$$

is twice continuously differentiable, and is a cubic polynomial on each segment $[\xi_j, \xi_{j+1}]$. Furthermore, with $f(x)$ defined by (15.4),

$$Y_i = f(x_i) + \varepsilon_i$$

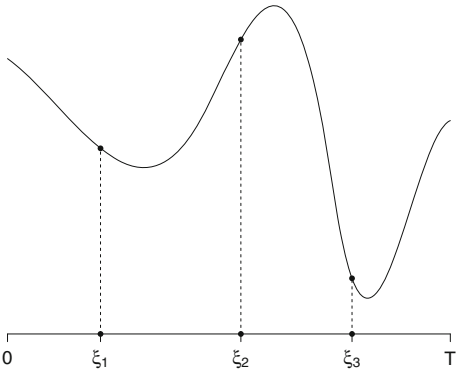

Fig. 15.3 A cubic spline with three knots, on an interval $[0, T]$. The function $f(x)$ depicted here is made up of distinct cubic polynomials (cubic polynomials with different coefficients) on each sub-interval $[0, \xi_1]$, $[\xi_1, \xi_2]$, $[\xi_2, \xi_3]$, $[\xi_3, T]$.

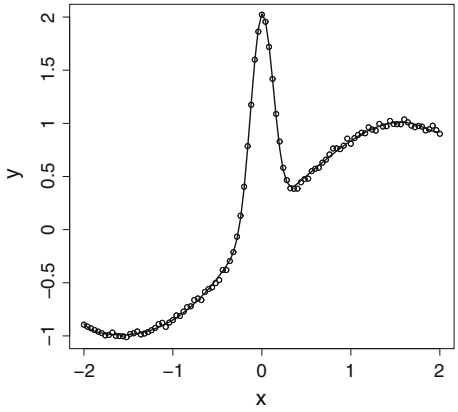

Fig. 15.4 A cubic spline fit to the data from Fig. 15.2. The spline has knots $(\xi_1, \xi_2, \ldots, \xi_7) = (-1.8, -.4, -.2, 0, .2, .4, 1.8)$.

becomes an instance of the usual linear regression model (assuming $\epsilon_i \sim N(0, \sigma^2)$, independently), so that regression software may be used to obtain spline-based curve fitting. Specifically, we define $x_1 = x$, $x_2 = x^2$, $x_3 = x^3$, $x_4 = (x - \xi_1)_+^3$, ..., $x_{p+3} = (x - \xi_p)_+^3$ and then regress Y on $x_1, x_2, \ldots, x_{p+3}$. To be concrete, let us take a simple special case. Suppose we have 7 data values $y_1, \ldots, y_7$ observed at 7 x values $(-3, -2, -1, 0, 1, 2, 3)$ and we want to fit a spline with knots at $\xi_1 = -1$ and $\xi_2 = 1$. Then we define $y = (y_1, \ldots, y_7)^T$, $x_1 = (-3, -2, -1, 0, 1, 2, 3)^T$, $x_2 = (9, 4, 1, 0, 1, 4, 9)^T$, $x_3 = (-27, -8, -1, 0, 1, 8, 27)^T$. The variables x_1, x_2, x_3 represent x, x^2, x^3. We continue by defining $x_4 = (0, 0, 0, 1, 8, 27, 64)^T$ and $x_5 = (0, 0, 0, 0, 0, 1, 8)^T$, which represent $(x - \xi_1)_+^3$ (which takes the value 0 for $x \leq -1$) and $(x - \xi_2)_+^3$ (which takes the value 0 for $x \leq 1$). Having defined these variables we regress y on x_1, x_2, x_3, x_4, x_5. Putting this regression in the form of (12.53) we have

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \end{pmatrix} = \begin{pmatrix} 1 & -3 & 9 & -27 & 0 & 0 \\ 1 & -2 & 4 & -8 & 0 & 0 \\ 1 & -1 & 1 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 8 & 0 \\ 1 & 2 & 4 & 8 & 27 & 1 \\ 1 & 3 & 9 & 27 & 64 & 8 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \end{pmatrix}. \tag{15.5}$$

When (15.4) is used the variables $x, x^2, x^3, (x-\xi_1)_+^3, \ldots, (x-\xi_p)_+^3$ are said to form the *power basis* for the set of cubic splines with knot set $\xi_1, \ldots, \xi_p$. This terminology indicates that any cubic spline with knots $\xi_1, \ldots, \xi_p$ may be represented in the form (15.4), which is a linear combination of $x, x^2, x^3, (x-\xi_1)_+^3, \ldots, (x-\xi_p)_+^3$ (together with the constant intercept).

An important caveat in applying (15.4), however, is that the variables $x_1, x_2, \ldots, x_{p+3}$ will be highly correlated. The possibility of polynomial x variables being correlated was considered in Section 12.5.4 and again in Section 14.1.2. Here there are two good solutions to this problem. The first is to *orthogonalize* the x variables. The trick of subtracting the mean, used in the earlier sections, is a special case of orthogonalization. The general method is to first replace x with $x_1^* = x - \bar{x}$; then regress $(x_1^*)^2$ on x_1^* and replace x^2 with x_2^* defined to be the residual from that regression; then regress $(x_1^*)^3$ on x_1^* and x_2^* and replace x^3 with x_3^* defined to be the residual from that regression; etc., continuing through the remainder of the regression variables to get a new set of variables $x_1^*, x_2^*, \ldots, x_{p+3}^*$ which are used instead of $x_1, x_2, \ldots, x_{p+3}$. The second, more commonly-applied alternative is to use a different version of splines, known as *B-splines*. *B*-splines may be used to form an alternative basis with which to represent cubic splines having knots $\xi_1, \ldots, \xi_p$, replacing the power basis in (15.4). The power basis and the *B*-spline basis represent the same set of cubic splines, but the *B*-spline basis offers better numerical stability. Thus, statistical software using *B*-splines for nonparametric regression will typically take the knot locations as input, and then will compute the X matrix as in (15.5), except that the columns will change because *B*-splines are used.[2] A variant of *B*-splines, known as *natural splines*, assumes the function is linear for x outside a specified range—which is often taken to be the range of the data (i.e., the function is linear for $x < x_{min}$ and $x > x_{max}$ where x_{min} and x_{max} are the smallest and largest values of x in the data). Because there is very little data near x_{min} and x_{max}, and none outside the range of the data, the fits based on the power basis and *B*-spline basis are often highly variable near the extremes of x. By introducing a strong assumption, natural splines are much less variable at the extreme values of x and typically provide nicer-looking fits. Natural splines are often recommended, and are an option in most statistical curve-fitting software. The power basis and *B*-spline basis each have $p + 4$ free parameters. Due to the addi-

[2] Because the span of the columns of the X matrix using *B*-splines will be the same as the span of X matrix using the orthogonalized power basis, the resulting least-squares estimated fits $X\hat{\beta}$ will be the same in both cases.

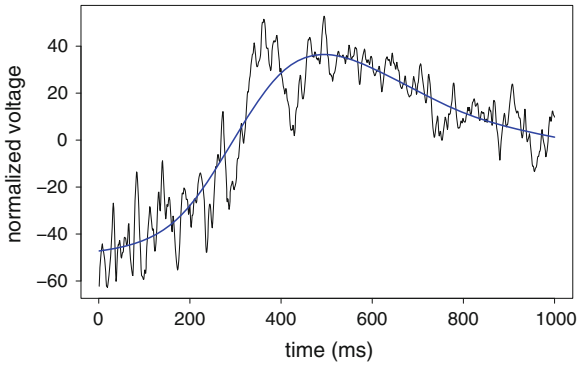

Fig. 15.5 LFP and smoothed version representing slowly-varying trend. A 1 s (seconds) sample of data is shown together with a smooth fit using natural splines.

tional constraints at each end of the range of x, the natural spline basis has $p + 2$ free parameters.

Example 15.2 Local field potential in primary visual cortex Kelly et al. (2010) examined the activity of multiple, simultaneously-recorded neurons in primary visual cortex in response to visual stimuli under anesthesia. As we noted in Example 2.2, under anesthesia the EEG displays strong delta range (1–4 Hz) wave-like activity. It is also common to see even lower frequency activity (less than 1 Hz), often called "slow waves," the effects of which are visible in Fig. 2.2. This activity appears in local field potential (LFP) recordings as well. In the data analyzed by Kelly et al., waves of firing activity were observed across the population of recorded neurons, and these were correlated with the waves of activity in the LFP. A short snippet of LFP is displayed in Fig. 15.5. In Chapter 18 we will examine the oscillatory content of this sample of the LFP. A preliminary step, discussed on p. 517, is to remove any slow trends in the data. Spline-based regression is useful for this purpose. A fit based on the natural-spline basis using knots at time points 200, 400, 600, 800 is shown in Fig. 15.5. □

15.2.3 Splines are also easy to use in generalized linear models.

Splines may also be used with logistic regression or Poisson regression, or other generalized regression models. When splines are used in regression models, they are often called *regression splines*. Standard statistical software usually includes options for using regression splines in generalized linear models.

Example 1.1 (continued from p. 187) In Chapter 1, p. 3, we discussed the problem of describing a neural response to a stimulus under two different experimental conditions in the context of recordings made from the SEF. In Chapter 8 we returned to the example to describe the value of smoothing the PSTH, using Fig. 8.3, on p. 187 to illustrate. We did not, however, say specifically how the smoothing was done. We obtained the smooth curve in Part b of Fig. 8.3 by fitting a Poisson regression spline. Specifically, spike counts Y were pooled across trials in 10 ms bins centered at times $x = -295, -285, -275, \ldots, 635, 645$ relative to appearance of the cue at time $x = 0$. Then the statistical model was

$$\begin{aligned} Y_i &\sim P(\lambda_i) \\ \log \lambda_i &= f(x_i) \end{aligned}$$

with $f(x)$ being a regression spline having knots at $-200, 200$. The fitted values $\hat{f}(x_i)$ were obtained using generalized linear model software and x_{max} was the value of x_i at which maximum among the $\hat{f}(x_i)$ values occurred. (Interpolation could be have been used to get a more refined maximum, but this was not considered necessary.) In Fig. 8.3, the arrow indicating the maximum of the fitted curve was plotted at $x = x_{max}$. A standard error for $x = x_{max}$ may be obtained by propagation of uncertainty (see Chapter 9).

In this example we would get similar results using a normal kernel density estimator (a Gaussian filter), which is discussed on pp. 431 and 578. □

15.2.4 With regression splines, the number and location of knots controls the smoothness of the fit.

Splines are very easy to use because the problem of spline fitting may be formulated in terms of a linear model. This, however, assumes that the knot set $\xi_1, \xi_2, \ldots, \xi_p$ has been determined. The choice of knots can be consequential: with more knots, the spline has greater flexibility, but also provides less smoothness. In addition, the placement of knots can be important. Figure 15.6 displays three alternative spline fits. The first two use splines with five and 15 knots having locations that are equally-spaced according to the quantiles of x so, for example, 5 knots would be placed at the $\frac{1}{6}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{5}{6}$ quantiles. Spacing the knots according to the quantiles of x allows more knots to be placed where there are more data values. The third spline uses seven knots chosen by eye. The spline with seven knots fits well because five knots are placed in the middle of the range, where the function variation is large, while only two are placed on the flanks where the variation is small.

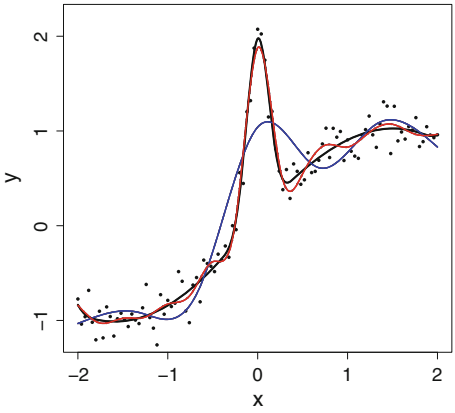

Fig. 15.6 Three cubic spline fits to data generated from the same test function as Fig. 15.2, but with more noise. Splines with 5 and 15 knots are shown (*blue* and *red lines*), with knot locations selected by default in R. The spline with five knots provides more smoothing than the spline with 15 knots and, as a result, does a poorer job of capturing the peak in the function. The spline shown in the *black line* has seven knots chosen to be $\xi = (-1.8, -.4, -.2, 0, .2, .4, 1.8)$.

15.2.5 Smoothing splines are splines with knots at each x_i, but with reduced coefficients obtained by penalized ML.

The problem of choosing knots may be solved in various ways, and in many situations it is adequate to select knots based on preliminary examination of the data and/or some knowledge of the way the function $f(x)$ is likely to behave. This is admittedly somewhat arbitrary, and two kinds of alternatives have been proposed that are more automated.

The first approach is to use a large number of knots, but to reduce, or "shrink," the values of the coefficients. One intuition here is that using a large number of knots in a regression spline would allow it to follow the function well, but would make it very wiggly; reducing the size of the coefficients will tend to smooth out the wiggles. A second intuition is obtained by replacing the least-squares problem of minimizing the sum of squares

$$SS = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

with the *penalized least squares* problem of minimizing the penalized sum of squares

$$PSS = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int (f''(x))^2 dx$$

where λ is a constant. The problem of minimizing *PSS* is similar to that of minimizing the penalized regression sum of squares in (12.71). Here, the squared second derivative is a *roughness penalty*: wherever $(f''(x))^2$ is large, the function is fluctuating substantially, and the integral of this quantity is a measure of the total fluctuation, or roughness. Thus, the value of the coefficient vector β^* that minimizes *PSS* will achieve some compromise between fitting the y_i values and keeping the function smooth. As λ increases, the resulting fit becomes increasingly smooth, and in the limit $\lambda \to \infty$ it becomes a line. It turns out that the solution to the penalized least squares problem is a cubic spline with knots at every value of x_i, but with coefficients that are smaller in magnitude than those of the regression spline with knots at every x_i (which would correspond to taking $\lambda = 0$). This solution is called *a smoothing spline*.

Smoothing spline technology has a strong theoretical foundation, and is among the most widely-used methods for nonparametric regression. There is also much well-developed software for smoothing splines. In the case of binomial or Poisson regression, the smoothing spline will maximize a penalized likelihood.

There remains the problem of choosing λ. Various alternative choices of λ may be tried. Statistical software typically provides options for choosing λ automatically by a variant of cross-validation (see p. 356) known as *generalized cross-validation* or by variants of ML called *generalized maximum likelihood* or *restricted maximum likelihood*. A smoothing spline fit to the data of Fig. 15.2 is visually indistinguishable from the spline fit in Fig. 15.4.

15.2.6 A method called BARS chooses knot sets automatically, according to a Bayesian criterion.

One defect of smoothing spline technology, and many other nonparametric methods, is that it assumes the degree of smoothness of $f(x)$ remains about the same across its domain, i.e., throughout the range of x values. An alternative is to devise a method that selects good knot sets based on the data. One of the most successful such procedures is called BARS (DiMatteo et al., 2001). In Fig. 1.6 of Example 1.7 BARS was applied to data from an electrooculogram, which produces voltage traces that are similar to many others, including EEG, ECoG, and LFP. There, BARS was able to retain the high-frequency signal (the sudden drop and sudden increase in voltage associated with an eye blink) while filtering high-frequency noise. In Figure 15.1 of Section 15.1 we displayed BARS fits to two peristimulus time histograms. BARS uses a Bayesian framework, and produces a posterior probability distribution on knot sets (see Section 16.1). Knot sets are then generated by simulation from the posterior distribution (Section 16.1.6). Based on each simulated knot set a fitted curve is obtained (the mean of these fitted curves is used for displays, as in Figs. 1.6 and 15.1). Finally, propagation of uncertainty is used to provide standard errors

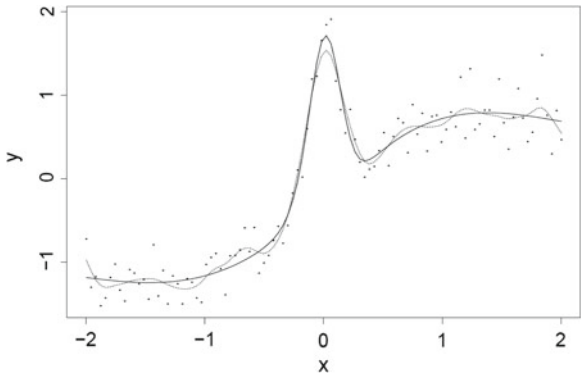

Fig. 15.7 Data from the test function of Fig. 15.2, but with more noise, as in Fig. 15.6, together with smoothing spline fit (*dotted line*) and BARS fit (*solid line*).

or intervals for quantities of interest. Figure 15.7 compares BARS and smoothing spline fits to the data from Fig. 15.6.

15.2.7 Spline smoothing may be used with multiple explanatory variables.

At the beginning of this chapter we recalled Eqs. (14.3) and (14.4), which we had used to define modern regression. In Section 15.2.2 we showed how splines are used to define a function $f(x)$ in ordinary linear regression and in Section 15.2.3 we gave the extension to binomial and Poisson regression. Those sections involved a single explanatory variable x. With p variables $x_1, \ldots, x_p$ it is too difficult to fit a function $f(x_1, \ldots, x_p)$ in full generality: there are too many possible ways that the variables may interact in defining $f(x_1, \ldots, x_p)$. However, a useful way to proceed is to make the strong assumption of an additive form:

$$f(x_1, \ldots, x_p) = \sum_{j=1}^{p} f_j(x_j). \tag{15.6}$$

With this restriction, spline smoothing (or alternative smoothing methods) may be applied to each variable successively in order to fit the model

$$Y_i = \sum_{j=1}^{p} f_j(x_j) + \epsilon_i \tag{15.7}$$

under the usual assumptions for linear regression. More specifically, an iterative algorithm may be used[3] to find the least-squares fit when a spline basis represents each function $f_j(x_j)$.

Example 15.3 Decoding natural images from V1 fMRI Kay et al. (2008) showed that natural images could be identified with above-chance accuracy from V1 activity picked up in fMRI responses. Vu et al. (2011) re-analyzed the data and showed how decoding accuracy could be improved by 30 % when additive models of the general form (15.7) were used. Kay et al. had applied a model of fMRI activity in a V1 voxel based on *Gabor wavelet filters*. Briefly, as shown in Fig. 15.8, a Gabor wavelet is a product of a sinusoidal factor and a factor based on a Gaussian (normal) pdf (see Section 15.2.8). The Gaussian factor is similar to that used in the hippocampal place cell model in (14.21). It has the effect of producing a response, for a particular voxel, based only on a small region in the visual image. The sinusoidal factor produces a central peak together with neighboring troughs that represent lateral inhibition, as is characteristic of the response of V1 neurons. The response due to each filter also has a particular orientation. The activity of each voxel in response to a particular image was regressed on filtered representations of the image. A set of 48 Gabor filters at 8 orientations and 6 spatial scales, as shown in Fig. 15.8, was used. Each image in the stimulus set produced a set of magnitudes $x_j(v)$, with $j = 1, \ldots, 48$, corresponding to the 48 filters, for each voxel v. These were the explanatory variables in the regression model, while the fMRI voxel activity was the response. Due to visible nonlinearities, Kay et al. performed a version of least squares based on $\sqrt{x_j(v)}$. Vu et al. found substantial nonlinearity in the residuals from the model of Kay et al., see Fig. 15.9. They then applied a model of the form (15.7) based on splines having 9 knots placed at the 10th, 20th, …, 90th percentiles of each explanatory variable. Because they had relatively large numbers of regression variables for each voxel, they applied a version of L1 penalized regression (see p. 358). The resulting additive model greatly improved the residual plots, see[4] Fig. 15.10. Vu et al. also showed that the additive model is more sensitive to weak stimuli, and this has the effect of broadening voxel tuning in space, frequency, and contrast. This, presumably, was the main source of improved performance. □

Equation (14.13) may be generalized to

$$\begin{aligned} Y_i &\sim f_{Y_i}(y_i|\eta_i) \\ g(\mu_i) &= \sum_{j=1}^{p} f_j(x_j) \end{aligned} \tag{15.8}$$

[3] One method, known as *backfitting*, cycles through the variables x_j, using smoothing (here, spline smoothing) to fit the residuals from a regression on all other variables.

[4] There remain upward trends in the residual plots. This is due to the penalized fitting, which induces correlation of residuals and fitted values.

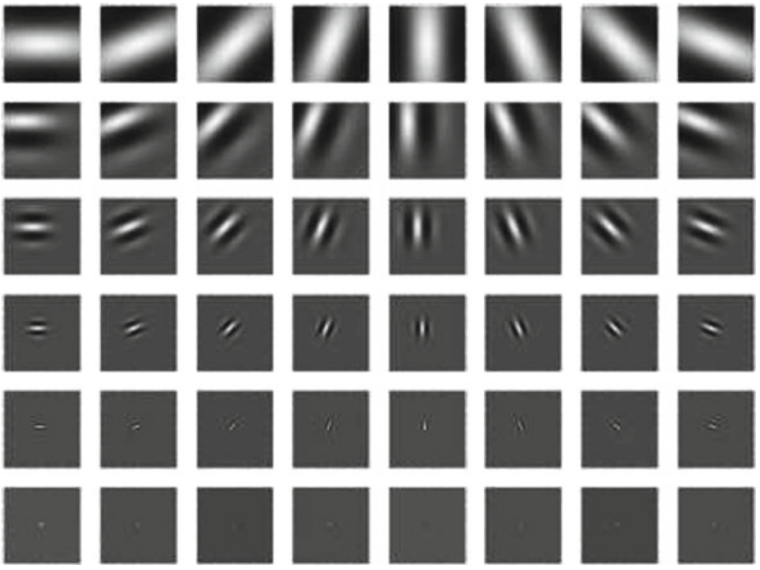

Fig. 15.8 Examples of Gabor wavelets at eight orientations (*columns*) and 6 spatial scales (*rows*). Adapted from Vu et al. (2011).

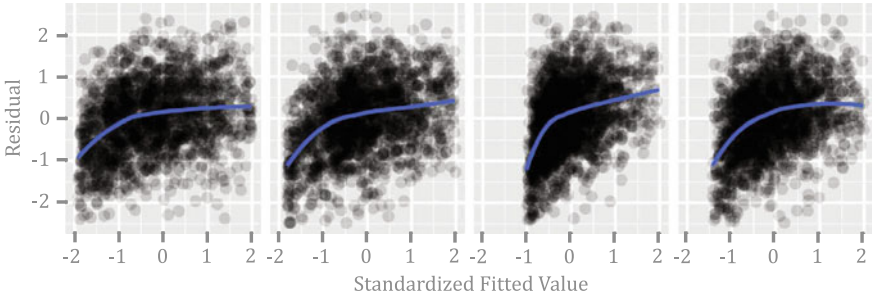

Fig. 15.9 Plots of residuals versus fitted values at four selected voxels for the model based on $\sqrt{x_j(v)}$. *Solid curve* is a local linear fit, as outlined in Section 15.3.2. Adapted from Vu et al. (2011).

where $f_{Y_i}(y_i|\eta_i)$ is an exponential family pdf as in (14.11), $\mu_i = E(Y_i)$, and $g(\mu)$ is the link function. The model (15.8) is known as a *generalized additive model*.

15.2.8 Alternatives to splines are often used in nonparametric regression.

We have discussed splines at some length because they are effective, easy to understand, and easy to use with available software. Other basis functions are often used.

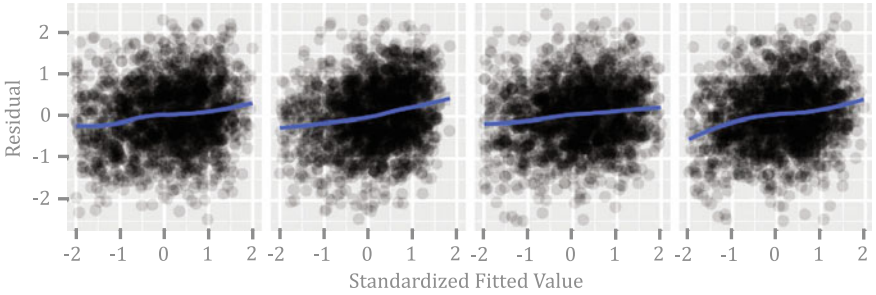

Fig. 15.10 Plots of residuals versus fitted values at the same four voxels as in Fig. 15.9, but using the additive model. *Solid curve* is a local linear fit, as outlined in Section 15.3.2. Adapted from Vu et al. (2011).

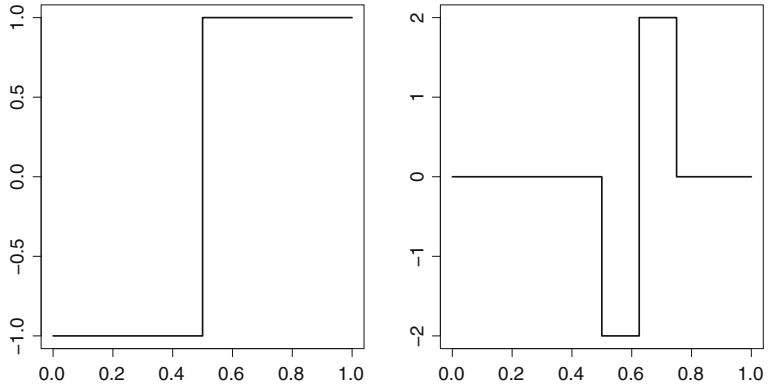

Fig. 15.11 *Left* The Haar mother wavelet $\psi(x)$. *Right* The Haar wavelet $\psi_{2,2}(x)$.

One popular choice is *wavelets*, which are often applied in time-frequency analysis of neural signals (see Section 18.3.7).

Wavelets use a *wavelet function*, or *mother wavelet* $\psi(x)$, and a *scaling function*, or *father wavelet* $\phi(x)$. A set of wavelet functions used for fitting is then defined by

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k) \tag{15.9}$$

together with the scaling function. For example, the Haar wavelets begin with

$$\psi(x) = \begin{cases} -1 \text{ if } 0 \le x \le \frac{1}{2} \\ 1 \quad \text{if } \frac{1}{2} < x \le 1 \end{cases}$$

and

$$\phi(x) = \begin{cases} 1 \text{ if } 0 \le x \le 1 \\ 0 \text{ otherwise.} \end{cases}$$

Figure 15.11 shows the Haar mother wavelet together with the Haar wavelet $\psi_{2,2}(x)$, which is concentrated in a narrow range. From the definition (15.9), the range of $\psi_{j,k}(x)$ becomes narrower as j increases. Haar wavelets are very simple and display the locality and scaling structure of wavelets in general. In practice other forms are used for the mother and father wavelets. For example, for *Gabor wavelets* or *Morlet wavelets*, the mother wavelet is a product[5] of a normal pdf and a sinusoidal term (see Fig. 15.8).

Wavelet-based nonparametric regression proceeds by defining a relatively large set of wavelets and then *shrinking* the coefficients (see p. 357), typically in such a way that most coefficients become zero, leaving a sparse representation involving few non-zero terms. The computations can be performed fast, using a method called the *discrete wavelet transform*.

Wavelets tend to be very good for automated, sparse representation of low-noise signals. When noise becomes more substantial there is unlikely to be a great advantage in using wavelets for nonparametric regression. In general, the choice of basis functions is largely a matter of preference and convenience.

15.3 Local Fitting

The second general approach to nonparametric regression is to use local fitting. Recall that in ordinary linear regression, the regression line is the expectation of Y as a function of x: we have $E(Y_i) = \beta_0 + \beta_1 x_i$ and could extend this to some newly-observed value of x by writing

$$E(Y|x) = \beta_0 + \beta_1 x. \tag{15.10}$$

In (15.10) we mean to include the case in which the data collection process makes it more reasonable to think of x as non-random. However, we have written $E(Y|x)$ to be reminiscent of our discussion, in Section 4.2.4, where we said that the regression of Y on a random variable X is the conditional expectation of Y given $X = x$. See the prediction theorem on p. 89.

Now, just as the expectation of a random variable is generally estimated by a sample mean, so the conditional expectation in (15.10) may be estimated as the mean of y_i values for which $X = x_i$, at least approximately. This is indicated in Fig. 4.3. When we generalize (15.10) to

$$E(Y|x) = f(x) \tag{15.11}$$

[5] The names Gabor and Morlet both get attached to what is perhaps more properly known as the Morlet wavelet, which has the form of a product of a normal pdf and a complex exponential, the real and imaginary parts of which are sinusoidal.

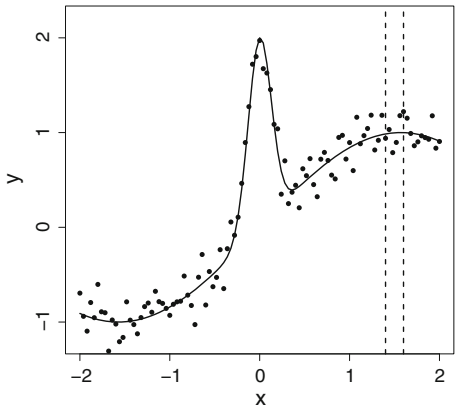

Fig. 15.12 Data simulated from function $f(x) = \sin(x)2\exp(-30x^2)$ (shown as *dark line*). The idea of local fitting begins with the notion that, just as in linear regression, for large data sets, the regression curve $f(x)$ at $x = 1.5$ should average the y-values among the points within the dashed lines. However, for smaller data sets, like that shown here, the region within the dashed lines contains relatively few points.

we may, in principle, also estimate $f(x)$ by averaging y_i values for $X = x_i$, approximately, as illustrated in Fig. 15.12. For large data sets the average gives an answer very close to the expectation. An immediate issue, however, is how to choose the size of the *window* (between the dashed lines in Fig. 15.12). Furthermore, in estimating $f(x)$ even with moderate-size data sets, it is possible to improve on the arithmetic mean among y_i values corresponding to x_i near x. For instance, in Fig. 15.12, there are not many values of x_i that are very close to any particular x. The idea of local fitting is to consider x_i values that are somewhat more distant from x, but to *weight* the various x_i values according to their proximity to x.

Two different ways to accomplish local fitting are distinguished by the names *kernel regression* and *local polynomial regression*.

15.3.1 Kernel regression estimates $f(x)$ with a weighted mean defined by a pdf.

In Section 8.1.3 we defined the the weighted mean of $y_1, \ldots, y_n$ to be

$$\bar{y}_w = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}$$

where $w_1, \ldots, w_n$ are positive numbers and w_i becomes the weight attached to the ith value. In kernel regression, each value $f(x)$ is estimated as a weighted mean of the observations y_i, with the weights increasing as x_i gets closer to x. The weights

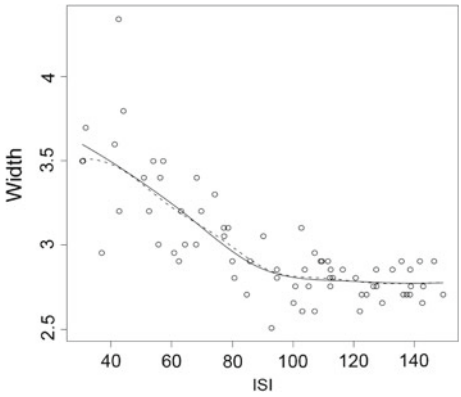

Fig. 15.13 Data showing the relationship of spike width to preceding ISI length for a neuron recorded in slice preparation. A kernel regression estimator is superimposed on the plot (*dashed line*) together with a local linear fit (*solid line*).

are defined by

$$w_i = K(\frac{x - x_i}{h}) \tag{15.12}$$

for a suitable function $K(u)$, which is called a *kernel*. The constant h is usually called[6] the *bandwidth*. The most commonly-used kernel is the $N(0, 1)$ pdf, in which case h effectively plays the role of a standard deviation, i.e., we have $w_i \propto K_h(x - x_i)$ where $K_h(u)$ is the $N(0, h^2)$ pdf. That is, $K(\frac{x-x_i}{h})$ is proportional to a normal pdf centered at zero having standard deviation h. This puts very nearly zero weight on y_i values for which $|x - x_i| > 3h$. Because many applications of smoothing arise in signal processing, some of the terminology is taken from that domain. In particular, when a normal kernel is used, it is often called a *normal filter* or *Gaussian filter*. Filtering is explained in Section 18.3.4.

More generally, any pdf could be used as a kernel. The formula for the kernel-regression fit is

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} K(\frac{x-x_i}{h}) y_i}{\sum_{i=1}^{n} K(\frac{x-x_i}{h})}. \tag{15.13}$$

Example 8.2 (continued, see p. 193) Previously we provided some results from a study of action potential width as a function of the preceding ISI, and Fig. 8.6 displayed a plot of some data from one neuron recorded from rat barrel cortex in a slice preparation. A portion of the data are shown again here, in Fig. 15.13. Only the data points for which ISI was less than 200 ms are displayed, and the analysis here only considered this truncated data set. Kernel regression, with a normal kernel,

[6] The terminology comes from spectral analysis (see Section 18.3.3) where the width corresponds to a band of frequencies.

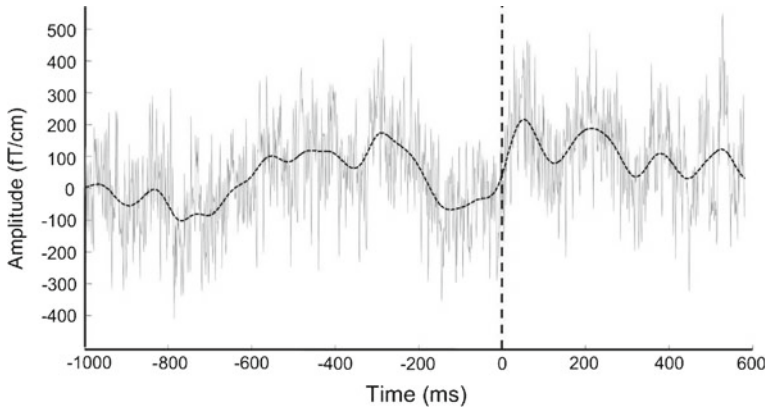

Fig. 15.14 MEG signal from a single sensor on a single trial. (Adapted from Wang et al. 2010.) This trial involved wrist movement, and time $t = 0$ corresponded to onset of movement. The dashed line through the sensor tracing is the smoothed version obtained from the normal kernel regression (a Gaussian filter).

produced the fitted relationship shown by the dashed line in the figure. The bandwidth used was 30 ms. □

The choice of bandwidth h in kernel regression is important, and affects smoothness: when h is small, the estimate tends to follow the data closely, but is very rough, while when h is large the estimate becomes smooth but may ignore places where the function seems to vary. Bandwidth selection involves a "bias versus variance" trade-off: small h reduces bias (and increases variance) while large h reduces variance (but increases bias). See Section 15.3.3.

Example 4.7 (continued from p. 358) The MEG decoding study of Wang et al. (2010), described on p. 100, involved predicting actual or imagined wrist movement from sensor signals. A preliminary step was to smooth each sensor signal, recorded on each trial. One such signal is shown in Fig. 15.14 together with a smoothed version based on a normal kernel. The bandwidth was 25 ms. This value of the bandwidth was chosen because it is a round number and provided what seemed to be a reasonable amount of smoothing when many plots were examined by eye, taking into consideration the temporal accuracy required in the subsequent analyses. □

15.3.2 Local polynomial regression solves a weighted least squares problem with weights defined by a kernel.

A second idea in local fitting of $f(x)$ is to solve a weighted least-squares problem defined at x by suitable weights $w_i = w_i(x)$. In particular, *local linear regression* at x minimizes

$$WSS(x) = \sum_{i=1}^{n} w_i(y_i - \beta_0 - \beta_1(x - x_i))^2 \tag{15.14}$$

where the weights w_i are defined in terms of a kernel, as in (15.12). A normal pdf may be used as the kernel, but an alternative is

$$K(u) = (1 - |u|^3)^3$$

for $|u| < 1$ and $K(u) = 0$ otherwise. The latter form of the kernel is used in some statistical software. Extensive study of this methodology has shown that local linear regression is effective in many situations. As with kernel regression, in local polynomial regression[7] there remains a choice of bandwidth. See Loader (1999) for further discussion, references, and extensions.

Example 8.2 (continued) In Fig. 15.13 we displayed a plot of some action potential width data together with a nonparametric regression fit based on a normal kernel (or Gaussian filter). A local linear fit is also shown in Fig. 15.13. In this example the local linear fit is nearly identical with the kernel regression fit. □

An important feature of local linear regression is that it may be extended to non-normal families such as binomial and Poisson. The idea is very simple. In place of the locally weighted sum of squares in (15.14) we can, for any value of the explanatory variable x_i, maximize a locally weighted loglikelihood having the form

$$WLL(x) = \sum_{i=1}^{n} w_i \ell(\beta_0 - \beta_1 x_i).$$

More specifically, in the case of binomial local linear fitting, with $Y_i \sim B(n_i, p_i)$, we have

$$WLL(x) = \sum_{i=1}^{n} w_i(y_i \log p_i + (n - y_i) \log(1 - p_i))$$
$$\log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_i.$$

Maximizing this loglikelihood for each successive x_i produces the fit at x_i.

[7] A popular variation on this theme, called *loess*, modifies the weights so that large residuals (outliers) exert less influence on the fit. The terminology comes from the English meaning of loess, which is a silt-like sediment, and is derived from German word *löss*, which means "loose."

15.3.3 Theoretical considerations lead to bandwidth recommendations for linear smoothers.

Recall, from Section 8.1.1, that $MSE = \text{Bias}^2 + \text{Variance}$. A minimal requirement of an estimator, in large samples, is that its bias and variance vanish (as $n \to \infty$). Consider estimation of $f(x)$ at the single point x. A linear smoother is, at x, a linear combination of the data response values y_i, so that the estimator may be written in the form

$$\hat{f}(x) = \sum_{i=1}^{n} w_i(x) y_i$$

where $w_i(x)$ emphasizes that the weights are determined for each x. We want

$$E(\hat{f}(x)) \to f(x) \tag{15.15}$$

and

$$V(\hat{f}(x)) \to 0. \tag{15.16}$$

Because $E(Y_i) = f(x_i)$ we also have

$$E\hat{f}(x) = \sum_{i=1}^{n} w_i(x) f(x_i)$$

so that the bias vanishes, as stated in (15.15), if the weights $w_i(x)$ become concentrated near x and the function $f(x)$ is smooth. For the weights to become concentrated it is sufficient that

$$\sum_{i=1}^{n} (x_i - x)^2 w_i(x) \to 0.$$

Assuming $V(Y_i) = \sigma^2$ (or, at least, that the variances do not vary rapidly), the variance vanishes if

$$\sum_{i=1}^{n} w_i(x)^2 \to 0.$$

Conditions like these on the weights, to guarantee (15.15) and (15.16), need to be assumed by any large-sample theoretical justification of a linear smoothing method. An explicit expression for the MSE of kernel estimators was given by Gasser and Muller (1984). This allows a theoretical bias versus variance trade-off, i.e., a formula for bandwidth selection as a function of n.

15.4 Density Estimation

Suppose we have a sample $U_1, \ldots, U_n$ from a distribution having pdf $f_U(u)$. If $f_U(u)$ is specified by a parameter vector θ (so that $f_U(u) = f_U(u|\theta)$) we may apply ML to estimate θ and thereby determine $f_U(u)$. Sometimes, however, we do not wish to assume a particular parametric form, yet we still want to obtain an estimate of the pdf. This presents the problem of nonparametric *density estimation*.

15.4.1 Kernels may be used to estimate a pdf.

One of the most popular ways to estimate a density is to apply a kernel, in the form we give below. It is possible to view the problem of density estimation as a special case of the problem of nonparametric regression, and in particular to derive a kernel density estimate from (15.13). We provide some discussion of this in the next subsection. Here we consider a somewhat simpler motivation for the procedure.

Recall that, for small h,

$$f_U(u) \approx \frac{P(u-h < U < u+h)}{2h}.$$

Then a direct estimate of $f_U(u)$ is

$$\hat{f}_U(u) \approx \frac{\text{no. obsn's falling in (u} - \text{h, u} + \text{h)}}{2nh}. \tag{15.17}$$

This estimate can be written in terms of the kernel $K(z) = \frac{1}{2}$ for $|z| < 1$ and 0 otherwise: we have

$$\hat{f}_U(u) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{u-u_i}{h}). \tag{15.18}$$

This direct (or "näive") estimate is also essentially a histogram with bins centered at the observations: if we normalize an ordinary histogram to give it the form of a pdf we get

$$\hat{f}_{U,hist}(u) = \frac{\text{no. obsn's in same bin as u}}{2nh}.$$

Both the histogram and the estimate in (15.17) suffer from being rectangular, and thus unable to produce a smooth curve as an estimate of the pdf. If we instead replace the kernel $K(z) = \frac{1}{2}$ for $|z| < 1$ with a smooth kernel, such as the normal pdf, we will get a smooth density estimate. In this general form the result of applying (15.18) produces what is known as a *kernel density estimate*. Kernel density estimation may be considered a way of getting a smooth density to replace the histogram. The normal (Gaussian) kernel is often used, though other choices are generally available

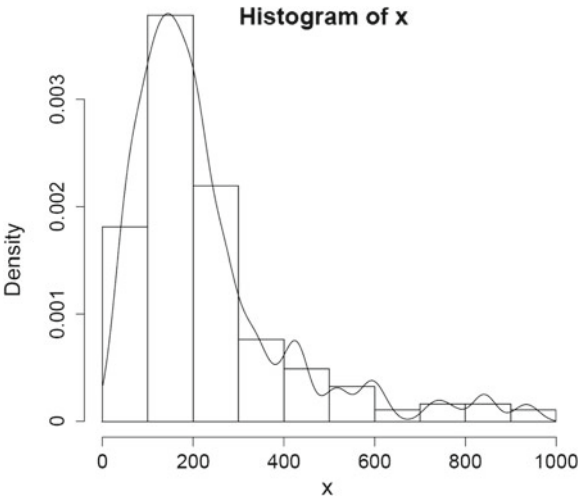

Fig. 15.15 A Gaussian kernel density estimator superimposed on an ISI histogram for the ISI data of Fig. 15.13. Here the histogram bin width was chosen using the "oversmoothed" rule from Scott (1992, p. 46), which produced 10 bins of width 100 ms; the bandwidth of the Gaussian kernel was set at 100 ms.

in density estimation software. As stated in Section 15.3.1, when a normal kernel is used it is often called a *normal filter* or *Gaussian filter*. Filtering is explained in Section 18.3.4.

As in kernel regression, the bandwidth parameter h is important. As we discussed in Chapter 2, choice of bin width is similarly important when using a histogram. For small h the estimate will tend to follow the data, but will be wiggly, while for large h the estimate will be smooth, but may not respond quickly to bunching of points that should indicate an increase in probability density. A variety of methods have been proposed for automatic selection of h, but many analysts choose h based on examination of the data, and experience with similar data (often picking a round number for h, which indicates the arbitrariness in the choice).

Example 8.2 (continued): We now examine only the ISI component of the data considered earlier, including all ISIs under 1,000 ms. A Gaussian kernel density estimate is shown in Fig. 15.15 superimposed on an ISI histogram. □

15.4.2 Other nonparametric regression methods may be used to estimate a pdf.

Many alternatives to kernel density estimation have been studied, and some of these can provide better estimates in certain situations. The virtue of kernel density estimation is that it is fast, easy, and often effective. When some imprecision in the estimate is tolerable, kernel density estimation is often a method of choice.

It is possible to view density estimation as a problem in binary nonparametric regression: we consider a very fine grid of values of u and define a variable that is one whenever a grid interval contains an observation, and 0 otherwise; estimating the expectation of these binary random variables amounts to estimating the pdf of U. Thus, with any method of nonparametric regression for binary data, after the regression estimate is normalized so that it integrates to one it may be considered a density estimate.

Details: Let us suppose we wish to obtain $\hat{f}_U(u)$ at some grid of u values, as we would in order to plot $\hat{f}_U(u)$, and let us write the grid as $x_1, x_2, \ldots, x_m$, so that the pairs we would plot would be $(x_j, \hat{f}_U(x_j))$, for $j = 1, \ldots, m$. We are using the notation x_j to distinguish the grid points from the random variable observations u_i. For the purpose of plotting this pdf we would, typically—as in plotting any function—choose m to be a fairly large value (such as 200), so that the plotted graph would not appear jagged. For convenience, let us take $\Delta x = x_j - x_{j-1}$, assuming the grid points to be equally spaced. Then taking a large m is equivalent to making Δx small. Let us assume that the grid is chosen to be sufficiently fine that there is at most one observation u_i in any given interval (x_{j-1}, x_j). (We may take $x_0 = x_1 - \Delta x$.) Viewing this procedure probabilistically, we can set up our grid prior to observing $U_1, \ldots, U_n$ and take it to be sufficiently fine that the probability of obtaining more than one observation in any given interval is negligible. The probability that an observation U_i will fall in interval (x_{j-1}, x_j) is approximately $f_U(x_j)\Delta x$. (We could improve the approximation somewhat by instead taking it to be $f_U(\frac{x_j + x_{j-1}}{2})\Delta x$, but will ignore this distinction here, as we are assuming Δx is small, so that $f_U(x_j) \approx f_U(\frac{x_j + x_{j-1}}{2})$.) Now let $Y_j = 1$ if the interval (x_{j-1}, x_j) contains an observation U_i (for some i) and 0 otherwise. Then Y_j, for $j = 1, \ldots, m$, forms a sequence of binomial random variables with

$$E(Y_j) \approx n f_U(x_j)\Delta x. \tag{15.19}$$

Because Y_j varies with j, it varies also with x_j and we may think of this expectation as a conditional expectation $E(Y_j|x_j)$; and because nonparametric regression methods estimate such conditional expectations, we may apply a kernel method to the estimation of the left-hand side of (15.19) in order to obtain an estimate of $f_U(u)$, which appears on the right-hand side. Specifically, writing $x = x_j$ and applying (15.13), we have

$$n\hat{f}_U(x)\Delta x = \frac{\sum_{j=1}^{m} K(\frac{x-x_j}{h})y_j}{\sum_{j=1}^{m} K(\frac{x-x_j}{h})} = \frac{\Delta x \sum_{j=1}^{n} K(\frac{x-x_j}{h})y_j}{\sum_{j=1}^{m} K(\frac{x-x_j}{h})\Delta x}. \tag{15.20}$$

For large m we have

$$\sum_{j=1}^{m} K(\frac{u-x_j}{h})\Delta x \approx \int K(\frac{u-x}{h})dx$$

and, setting $z = (u-x)/h$, because $K(z)$ is itself a pdf its integral is one, so from $x = u - hz$ we have

$$\int K(\frac{u-x}{h})dx = h\int K(z)dz = h.$$

Thus, for large m, the sum appearing in the denominator of (15.20) is approximately h. In the numerator, we note that $y_j = 0$ except when there is an observation u_i in (x_{j-1}, x_j), in which case $x_j \approx u_i$. Plugging these into (15.20), canceling Δx from the left-hand side and the numerator of (15.20), and replacing x with u then gives (15.18). □

Chapter 16
Bayesian Methods

Few results are as consequential for data analysis as Bayes' Theorem. The theorem itself, which we introduced in Sections 3.1.4 and 4.3.3, is a simple re-formulation of conditional probability and is easy to derive. Its conceptual power has been illustrated already in three previous chapters. In Example 3.2, p. 44, we calculated the probability of having vascular dementia based on a positive result from a screening test and found it to be surprisingly small. In Section 4.3.4 we showed that Bayes classifiers are optimal, in the sense of having the smallest possible mis-classification rate, and mentioned that such optimality has been used in theoretical studies of vision, where object-recognition is often considered a problem of classification. Then, in Section 7.3.9, we described the Bayesian approach to providing inferential intervals, called *credible* intervals, that display knowledge and uncertainty about the value of an unknown parameter; and in the context of Example 1.4 (on p. 175) we showed that credible intervals can have good frequentist coverage probability, getting closer to the nominal .95 probability than the standard approximate confidence interval we had applied earlier. These vignettes indicate some of the ways Bayes' theorem can generate important data analytic procedures. There are many texts on Bayesian statistical methods. In this chapter we limit our discussion to several fundamental topics. In Section 16.1 we present key concepts that help in understanding, and computing, posterior distributions. In Section 16.2 we describe a setting in which Bayesian formalism is especially valuable: the use of latent variables, including so-called "hidden states." Finally, in Section 16.3 we return to the Bayesian approach to hypothesis testing, which we mentioned in Section 10.4.5, and discuss it at greater length.

Bayes' theorem has been the source of great debates across more than 200 years (see McGrayne 2011). On the one hand, Bayesian inference has seemed compelling to many people. In the first place, as we pointed out on p. 174, Bayes' theorem provides a straightforward interpretation of what we know based on the data at hand. Secondly, because Bayes' theorem is a law of probability, Bayesian inference is self-consistent (it does not yield[1] logical paradoxes). Thirdly, as we showed in Section 4.3.4, it yields

[1] Exceptions occur for improper priors; see Section 16.1.4.

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_16,

optimal decisions. Furthermore, as we say on p. 450, it produces one of the guiding principles of science, namely that with sufficiently good data any two investigators will come to agreement. Taken together, these properties inspire awe among many who are able to appreciate them. Yet when people become captivated by the spell of Bayes' theorem they tend to proselytize, and become blinded to its fundamental vulnerability: its powers depend on the accuracy of its probabilistic inputs. Just as other kinds of zealots have always found combative foes, so Bayesians for many years joined battle with non-believers. The arguments turned out to be partly productive, but mostly futile. The contemporary view is much more civilized and Bayes' theorem is now[2] widely recognized as a crucial tool for data analysis.

Within neuroscience enthusiasm for Bayesian inference has been voiced, with many theoreticians claiming that it can yield important insights into human behavior and the functioning of the nervous system. We hinted at this when we mentioned Bayesian decision-making, beginning on p. 102. Further discussion may be found in Griffiths et al. (2012), Jacobs and Kruschke (2010), Knill and Pouget (2004), Körding (2007), and Wolpert et al. (2011), and the references therein. Bayesian inference is not a panacea, but it has supplied a fruitful conceptual framework. While the aim of this chapter is to present essential Bayesian concepts for analysis of neural data, the constructions we review here are also crucial components of Bayesian thinking about neural systems.

16.1 Posterior Distributions

A central formula in deriving Bayesian methods is (7.28), which gives the posterior pdf of a parameter θ given data x in terms of the likelihood function $L(\theta)$ and the prior pdf $\pi(\theta)$. We repeat that formula here:

$$f_{\theta|x}(\theta|x) = \frac{L(\theta)\pi(\theta)}{\int L(\theta)\pi(\theta)d\theta}. \tag{16.1}$$

In this chapter we drop the subscript on the posterior pdf and write it as $f(\theta|x)$.

As we have already seen in Section 7.3.9, estimation and uncertainty about an unknown parameter θ may be summarized with the posterior mean and standard deviation, which become an alternative to the MLE and SE. For skewed posterior distributions, where the mean and mode may differ, sometimes the mode is used instead of the mean; and sometimes the mode is easier to compute than the mean. Because the mode maximizes the posterior pdf it is often called the *maximum a posteriori* (*MAP*) estimate. As an estimator of θ, the posterior mean is sometimes motivated by the theorem on p. 90 which, in this context, says that the posterior

[2] See Kass (2011) for an elaboration of the current philosophical pragmatism among most practicing statisticians.

mean will be optimal in terms of minimizing the posterior mean squared error, i.e., the value of d that minimizes $E((\theta - d)^2|x)$ is $d = E(\theta|x)$.

16.1.1 *Bayesian inference equates descriptive and epistemic probability.*

Bayesian methods may be considered to supply good solutions to a limited range of statistical problems—particularly those involving latent variables, as we discuss in Section 16.2. In this sense, they are specialized. However, the application of Bayes' theorem to problems of statistical inference as in (16.1) involves a broadly important conceptualization. This is worth appreciating.

As we stated on p. 14, probability is used in two distinct ways. It is used *descriptively* in modeling variation, as when we said on p. 37 that the probability of rolling an even number with a fair six-sided die is $\frac{1}{2}$ or, in fact, as in any of the statistical models we have used throughout the book. Probability is also used to quantify a state of uncertain knowledge, as we illustrated on p. 172 with the statement "I am 90 % sure the capital of Louisiana is Baton Rouge." This second use of probability, to indicate a state of knowledge is often called *epistemic*. In Eq. (16.1), the likelihood function comes from the pdf $f(x|\theta)$, which is a descriptive use of probability. The prior and posterior distributions, however, use probability to quantify uncertain knowledge. As we highlighted in Section 7.3.9, the indirect interpretation of confidence intervals, which is based solely on descriptive probability, contrasts sharply with the stronger and more intuitive epistemic interpretation of posterior credible intervals. Bayesian analysis is based on an *inferential principle of equivalence*, which asserts that there is only one kind of probability for *both* descriptive and epistemic purposes, and that epistemic statements can be made using descriptive quantifications merely by applying the laws of probability, i.e., by Bayes' theorem. Let us return to the version of formula (16.1) in which we see explicitly how Bayes' theorem is used, namely Eq. (7.27), which we repeat here:

$$f_{\theta|x}(\theta|x) = \frac{f_{X|\theta}(x|\theta)f_\theta(\theta)}{\int f_{X|\theta}(x|\theta)f_\theta(\theta)d\theta}. \tag{16.2}$$

The pdf $f_{X|\theta}(x|\theta)$ in (16.2) is where probability enters descriptively (describing the variation in the data), but the posterior pdf $f_{\theta|x}(\theta|x)$ produces epistemic statements, as on p. 174. The inferential principle of equivalence says that we may convert descriptive uses of probability to epistemic ones, as in (16.2).

This simple principle is attractive partly for the reasons we enumerated in the introduction, on p. 439, and also because it is remarkably powerful in its ability to solve complicated statistical problems: once a statistical model is able to do a reasonably good job in describing variation, and prior information is formulated, at

least approximately, then Bayesian inference can produce useful quantifications of the way the data produce new knowledge.

16.1.2 Conjugate priors are convenient.

Let us return to the binomial setting discussed in Sections 7.3.9 and 8.3.3. There we used a uniform prior $\pi(\theta) = 1$ and obtained the posterior pdf

$$f(\theta|x) = \frac{\theta^x(1-\theta)^{n-x}}{\int \theta^x(1-\theta)^{n-x}d\theta}$$

which matched the beta pdf form given on p. 125,

$$f(w) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}w^{\alpha-1}(1-w)^{\beta-1}, \tag{16.3}$$

producing the posterior distribution $\theta|X = x \sim Beta(x+1, n-x+1)$. Interestingly, as may be seen from Eq. (16.3), the uniform distribution on (0, 1), which has pdf $f(w) = 1$, is the same as the $Beta(1, 1)$ distribution. Therefore, with the binomial likelihood a $Beta(1, 1)$ prior produces a $Beta(x + 1, n + 1)$ posterior. We may generalize by instead using a $Beta(\alpha_\pi, \beta_\pi)$ prior: in this case the posterior pdf becomes

$$f(\theta|x) = \frac{\theta^x(1-\theta)^{n-x}\theta^{\alpha_\pi}(1-\theta)^{\beta_\pi}}{\int \theta^x(1-\theta)^{n-x}\theta^{\alpha_\pi}(1-\theta)^{\beta_\pi}d\theta} \tag{16.4}$$

and, from Eq. (5.15), above, this may be recognized as a $Beta(\alpha_{post}, \beta_{post})$ pdf where

$$\begin{aligned}\alpha_{post} &= x + 1 + \alpha_\pi \\ \beta_{post} &= n - x + 1 + \beta_\pi.\end{aligned}$$

Thus, in conjunction with the binomial likelihood, a beta prior distribution produces a beta posterior poster distribution. In such situations, where a prior distribution leads to a posterior within the same parametric family of distributions, the prior is called *conjugate*.

Here is another example. Let $X_1, X_2, \ldots, X_n$ be a sample of $N(\theta, \sigma^2)$ random variables, write $X = (X_1, \ldots, X_n)$, take $\overline{X}$ to be the usual sample mean of the X_i variables so that $\overline{X} \sim N(\theta, \sigma^2/n)$, and assume σ is known. If we let the prior distribution be normal with $\theta \sim N(\mu_\pi, \sigma_\pi^2)$ then the posterior distribution is also normal. This is because the likelihood function based on the data $(x_1, \ldots, x_n)$ is the same as the likelihood function based on $\overline{x}$ (the sample mean is sufficient, and Bayes sufficient, as mentioned on p. 200), and is given by

$$L(\theta) = \exp\left(-\frac{n}{2\sigma^2}(\bar{x} - \theta)^2\right) \tag{16.5}$$

which may be written as $L(\theta) = \exp(Q_1(\theta))$ for a quadratic function $Q_1(\theta)$, while the prior similarly has the form $\pi(\theta) = \exp(Q_2(\theta))$ for another quadratic function $Q_2(\theta)$. The posterior therefore has the form

$$f(\theta|x) \propto \exp(Q_3(\theta)) \tag{16.6}$$

where $Q_3(\theta) = Q_1(\theta) + Q_2(\theta)$ is quadratic and the symbol $\propto$ means "proportional to," i.e.,

$$f(\theta|x) = c\exp(Q_3(\theta)) \tag{16.7}$$

for some nonzero constant c. Equation (16.6) implies that the posterior pdf is normal. (The normal pdf has the form (16.7) where the constant c is chosen so that $f(\theta|x)$ integrates to 1.) Thus, for the problem of estimating a normal mean, the normal prior is conjugate. In Section 16.1.3 we give the formulas for the posterior mean and variance in this case.

More generally, exponential family models (see Section 14.1.6) have conjugate priors. For instance, for Poisson likelihood functions gamma distributions become conjugate priors. Conjugacy is advantageous because formulas for pdfs, and also means and variances, may be derived and software for conjugate families is available.

16.1.3 *For exponential families with conjugate priors the posterior mean is a weighted combination of the MLE and the prior mean.*

In Section 16.1.2 we noted that when $X_1, \ldots, X_n$ form a sample from a $N(\theta, \sigma^2)$ distribution and the prior is a $N(\mu_\pi, \sigma_\pi^2)$ distribution, the posterior is also normal, due to Eq. (16.6). To be explicit, with $X = (X_1, \ldots, X_n)$, we have

$$\theta|X = x \sim N(\mu_{post}, \sigma_{post}^2)$$

where

$$\mu_{post} = \frac{\sigma_{\bar{x}}^2}{\sigma_{\bar{x}}^2 + \sigma_\pi^2}\mu_\pi + \frac{\sigma_\pi^2}{\sigma_{\bar{x}}^2 + \sigma_\pi^2}\bar{x} \tag{16.8}$$

$$\sigma_{post}^2 = \left(\frac{1}{\sigma_{\bar{x}}^2} + \frac{1}{\sigma_\pi^2}\right)^{-1} \tag{16.9}$$

in which

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}. \tag{16.10}$$

Details: Taking logs in Eq. (16.6) and inserting $\sigma_{\bar{x}}$ according to (16.10), we have

$$\log f(\theta|x) = -\frac{1}{2}\left(\frac{\bar{x}-\theta}{\sigma_{\bar{x}}}\right)^2 - \frac{1}{2}\left(\frac{\theta-\mu_\pi}{\sigma_\pi}\right)^2 + \text{constant}$$

where "constant" refers to all terms that are constant in θ. We then note that

$$\frac{1}{\sigma_{\bar{x}}^2}(\theta-\bar{x})^2 + \frac{1}{\sigma_\pi^2}(\theta-\mu_\pi)^2 = (\frac{1}{\sigma_{\bar{x}}^2} + \frac{1}{\sigma_\pi^2})(\theta-\mu_{post})^2 + \text{constant}$$

where μ_{post} is given by (16.8). Therefore, we have

$$\log f(\theta|x) = (\frac{1}{\sigma_{\bar{x}}^2} + \frac{1}{\sigma_\pi^2})(\theta-\mu_{post})^2 + \text{constant}$$

and, exponentiating,

$$f(\theta|x) \propto \exp\left((\frac{1}{\sigma_{\bar{x}}^2} + \frac{1}{\sigma_\pi^2})(\theta-\mu_{post})^2\right)$$

which shows the posterior is normal with mean μ_{post} and variance σ_{post}^2 given by (16.9). □

Equation (16.8) has a deeply useful interpretation. Let us rewrite it in the form

$$\mu_{post} = w\bar{x} + (1-w)\mu \tag{16.11}$$

where

$$w = \frac{\sigma_\pi^2}{\sigma_{\bar{x}}^2 + \sigma_\pi^2}.$$

In the special case $n = 1$ we write $x = x_1$ and get

$$\mu_{post} = wx + (1-w)\mu. \tag{16.12}$$

Equations (16.11) and (16.12) say that the posterior mean is a weighted combination of the MLE and the prior mean, with the weights determined by the relative precision (the inverse of the variance) of data and prior. In (16.11), as the precision in the data increases relative to the prior (i.e., as $\sigma_\pi^2/\sigma_{\bar{x}}^2$ increases), w increases, more weight is placed on $\bar{x}$, and the posterior mean becomes nearly the same as $\bar{x}$. Intuitively, when

the weight w is large, the data contribute more knowledge than the prior, and so the posterior is centered near the data value $\bar{x}$. When the data are imprecise relative to the prior (i.e., when $\sigma_\pi^2/\sigma_{\bar{x}}^2$ is small), more weight is placed on the prior mean, so that the posterior mean is pulled away from $\bar{x}$ and toward the prior mean. The posterior mean is often said to *shrink* the value $\bar{x}$ toward μ, particularly when $\mu = 0$ (so that the magnitude of μ_{post} is smaller than that of $\bar{x}$). In this terminology, the amount of *shrinkage* is determined by $1 - w$. In Section 16.2.3 we discuss the connection between the use of "shrinkage" in this context and in regression (see p. 357).

Example 16.1 Sensorimotor learning Körding and Wolpert (2004) designed an experiment in which visual input could be combined with a learned prior distribution in order to produce a finger movement. Subjects moved their index finger from a starting location toward a target, which was represented on a computer monitor. Halfway through the finger movement they were given visual feedback as to where their finger was at that moment (a cursor was shown briefly on the monitor) but, relative to a straight path between starting location and target, it was (a) corrupted by noise and (b) displaced to the right. The noisy location was indicated by a cloud of points drawn from a spherical bivariate normal distribution with one of 4 possible values of standard deviation (the standard deviation here refers to the standard deviation of each marginal distribution determined by the bivariate normal). This standard deviation would correspond to $\sigma_{\bar{x}}$ in Eqs. (16.8) and (16.9), and the center of the displayed cloud of points would correspond to $\bar{x}$. The size of the displacement varied with each trial, and was drawn from another normal distribution. The mean and standard deviation of this displacement distribution would correspond to μ_π and σ_π in Eqs. (16.8) and (16.9). In other words, the displacement distribution formed a prior and the center of the cloud of points (together with the standard deviation) became the subject's input data for each trial.

Subjects were given 1,000 training trials during which they could learn the prior displacement distribution. When queried afterwards they had no awareness of the displacement. The authors used an additional 1,000 trials to collect experimental data about the final location of each subject's finger. The authors showed that the displacement of the final location from the target was predicted well by Eqs. (16.8) and (16.9). In other words, in attempting to reach the target, subjects combined the visual input information with their prior knowledge of the displacement, at least approximately, as if their nervous system were computing a posterior mean according to Eqs. (16.8) and (16.9). □

Formulas analogous to Eq. (16.8) also hold for other exponential families with conjugate priors. For example, in the binomial setting let us reparameterize the $Beta(\alpha, \beta)$ distribution by defining

$$\mu = \frac{\alpha}{\alpha + \beta} \tag{16.13}$$

$$\nu = \alpha + \beta. \tag{16.14}$$

Here, μ is the mean of the beta distribution and ν is its variance -1 (see Section 5.4.5). The beta distribution may then be written $Beta(\mu\nu, (1-\mu)\nu)$ and the beta prior $\theta \sim Beta(\alpha_\pi, \beta_\pi)$ instead becomes $\theta \sim Beta(\mu_\pi\nu_\pi, (1-\mu_\pi)\nu_\pi)$. Similarly, the posterior based on a binomial $B(n, \theta)$ observation $X = x$ becomes

$$\theta|X = x \sim Beta(\mu_{post}\nu_{post}, (1-\mu_{post})\nu_{post})$$

and we have

$$\mu_{post} = \frac{\nu_\pi}{n+\nu_\pi}\,\mu_\pi + \frac{n}{n+\nu_\pi}\bar{x} \tag{16.15}$$

where we have written the observed proportion in the form $\bar{x} = x/n$ (thinking of the binomial as resulting from n Bernoulli trials). In (16.15) the data precision is not exactly the reciprocal of the variance but is instead represented by n and the prior precision is represented by ν_π. With these definitions of precision it is again true that as the precision in the data increases relative to the prior more weight is placed on the observed proportion $\bar{x}$ and the posterior mean becomes nearly the same as $\bar{x}$, while when the data precision gets relatively smaller the posterior mean is pulled away from $\bar{x}$ toward the prior mean.

The binomial posterior mean may be interpreted as equivalent to the MLE that would be obtained from the original data x together with some *pseudo-data* represented by the prior. For example, with a uniform prior (so that $\alpha = \beta = 1$), the posterior is a $Beta(x+1, n-x+1)$ distribution which, from (16.13), has mean

$$\mu_{post} = \frac{x+1}{n+2},$$

and this is equal to the MLE based on $x+1$ successes (1s) and $n-x+1$ failures (0s). That is, we imagine first supplementing the actual data with 1 success and 1 failure, and then finding the observed proportion of successes; this is the posterior mean. A similar statement remains true whenever α and β are integers. The non-integer case is sometimes interpreted by analogy. For example, if we use the conjugate prior with $\alpha = \beta = \frac{1}{2}$ the posterior mean is equal to the MLE we would get by "adding half a success and half a failure" to the data before finding the proportion of successes.[3]

The normal case may be interpreted similarly. Let us suppose, first, that $\sigma_\pi = \sigma$. In this case, in (16.11), we have $w = n/(n+1)$ and the posterior mean is the same as the sample mean from the original n observations supplemented by 1 observation having the value μ_π. Similarly, if $\sigma_\pi^2 = \sigma^2/k$, in (16.11), we have $w = n/(n+k)$ and the posterior mean is the same as the sample mean from the original n observations supplemented by k observations having mean μ_π. When the ratio σ^2/σ_π^2 is not an integer the interpretation works by analogy: the prior again injects some additional

[3] Adding 2 successes and two failures has been advocated as a way of achieving good frequentist coverage probability of approximate 95% CIs, i.e., intervals based on (7.22) with $\hat{p}$ replaced by $(x+2)/(n+4)$. See Agresti and Caffo (2000).

information, beyond the data, represented as if based on other data having sample mean μ_π and variance of that mean equal to σ_π^2.

16.1.4 There is no compelling choice of prior distribution.

In the case of a binomial $B(n, \theta)$ distribution, it is intuitive to take the prior for θ to be uniform on $(0, 1)$, so that the prior pdf is $\pi(\theta) = 1$. This uniformity seems to capture the notion that the prior is not favoring particular values of the parameter above others. Working by analogy, in the case of estimating a normal mean based on a sample from a $N(\theta, \sigma^2)$ distribution, with σ known, the prior on θ could taken to be uniform on $(-\infty, \infty)$ with $\pi(\theta) = 1$. This is not a pdf because its integral is

$$\int_{-\infty}^{\infty} 1 d\theta = \infty$$

while the integral of a pdf must equal 1. However, in this particular case the posterior turns out to be a well-defined probability distribution: it is normal with mean and variance given by Eqs. (16.8) and (16.9) where $1/\sigma_\pi = 0$. That is, in this case the posterior is normal with mean $\mu_{post} = \bar{x}$ and standard deviation is $\sigma_{post} = \sigma/\sqrt{n}$. Formal priors that have infinite integrals are called *improper*. In estimating a normal mean there is an identification of ML estimation with the improper Bayesian solution based on the prior $\pi(\theta) = 1$.

It is tempting to call a uniform prior "non-informative," and to apply it in other problems. An immediate difficulty, however, is that such a choice is not invariant to reparameterization: if θ is given a uniform distribution, the pdf of $\phi = g(\theta)$ for any one-to-one nonlinear function g (such as $\phi = e^\theta$) will be non-uniform, according to the theorem on p. 62, and one needs some justification for being "non-informative" for θ rather than ϕ. This forces the data analyst to decree a particular parameterization to be special. In some problems, such as the binomial or normal, a particular parameter, such as the binomial mean or the normal mean, may indeed play a special role in the problem. In particular, in estimating a normal mean, different values of $\bar{x}$ produce likelihood functions $L(\theta)$ in (16.5) that are exactly the same aside from the location of their peak (which is at $\bar{x}$). That is, as the value of $\bar{x}$ is translated along the real line, so too is the likelihood function; and this would not be the case for the likelihood $L(\phi)$ for any nonlinearly transformed parameter ϕ. It has been argued that if we wish the prior to represent maximal uncertainty, and inject minimal information, this special translational property should be respected by the posterior as well and, therefore, the prior should be uniform on θ. In other words, it has been argued that the special form of the likelihood function in (16.5) makes θ an exceptional parameterization for this

model, so that it becomes reasonable to consider the uniform prior $\pi(\theta) = 1$ to be[4] non-informative.

The binomial parameter θ, and the normal mean μ have a special status. In most statistical models, however, it is difficult to argue in favor of some particular choice of parameterization about which one might wish to be non-informative. In Example 7.1, for instance, there was a memory capacity parameter, and at first glance one might wish to be non-informative about it. But it might seem equally plausible to use its reciprocal or logarithm. There is no strong analogy with the argument about a normal μ in this case. Similarly, for the gamma model discussed in Section 7.2 it is not clear what parameterization should gain some special consideration. Furthermore, especially in multidimensional problems, there can be unanticipated consequences of uniformity, including especially the possibility that the posterior itself is improper and therefore no longer able to provide inferential statements such as credible intervals. As reviewed by Kass and Wasserman (1996), numerous methods have been proposed in an attempt to resolve these issues and provide rules for selecting prior distributions. (The bibliography in Kass and Wasserman (1996), has over 200 entries.) While particular choices often seem reasonable, there is a degree of arbitrariness in all and no consensus has emerged. Although this situation may seem problematic, it is balanced by considerations reviewed in Section 16.1.5.

16.1.5 For large samples, posteriors are approximately normal and centered at the MLE.

In Section 8.3.3 we cited the result that for large samples the posterior distribution of θ is approximately normal, with the normal distribution centered at the MLE $\hat{\theta}$ and having variance equal to the inverse of the observed information. We illustrated this for the binomial setting in Fig. 8.8, using a plot of a normal approximation to a beta posterior. We now elaborate.

The general argument uses the quadratic approximation to the loglikelihood function. For simplicity let us assume θ is a scalar. Because $\ell'(\hat{\theta}) = 0$ we may simplify the quadratic approximation (second-order Taylor series approximation),

$$\ell(\theta) \approx \ell'(\hat{\theta})(\theta - \hat{\theta}) + \frac{1}{2}\ell''(\hat{\theta})(\theta - \hat{\theta})^2$$

to get

$$\ell(\theta) \approx \frac{1}{2}\ell''(\hat{\theta})(\theta - \hat{\theta})^2. \tag{16.16}$$

[4] A similar argument may be applied to the case of estimating a normal standard deviation σ when the mean μ is assumed known, and this produces a uniform prior $\pi_\xi(\xi) = 1$ on $\xi = \log\sigma$, with the change of variables formula (see the theorem on p. 62) giving $\pi(\sigma) = 1/\sigma$.

This approximation holds for θ close to $\hat{\theta}$, in the sense that for sufficiently small values of $|\theta - \hat{\theta}|$ the difference between the left and right-hand sides of (16.16) may be made arbitrarily small. The argument then proceeds by showing that for sufficiently large n, the only values of θ that have appreciable posterior probability are those for which $|\theta - \hat{\theta}|$ is small, and the prior contributes negligibly to the posterior, so that the posterior pdf satisfies

$$\log f(\theta|x) \approx \frac{1}{2}\ell''(\hat{\theta})(\theta - \hat{\theta})^2 + \text{constant} \tag{16.17}$$

where "constant" means constant in θ (and determined by the condition $\int f(\theta|x)d\theta = 1$). Thus, the posterior pdf may be written, approximately, as

$$f(\theta|x) \approx k \exp\left(\frac{1}{2}\ell''(\hat{\theta})(\theta - \hat{\theta})^2\right) \tag{16.18}$$

where k is a proportionality constant.[5] In other words, the posterior pdf is approximately the same as a normal pdf with mean $\hat{\theta}$ and variance given by the inverse of the observed information $I_{OBS}(\hat{\theta}) = -\ell''(\hat{\theta})$. We now say a few more words[6] about the steps involved in obtaining (16.17).

First, as we said in Section 8.3.1, MLEs are consistent,

$$\hat{\theta} \xrightarrow{P} \theta. \tag{16.19}$$

For example, in the binomial case $X \sim B(n, \theta)$, we have $\hat{\theta} = \frac{x}{n}$ and we get $\hat{\theta} \xrightarrow{P} \theta$ by the law of large numbers (because X may be considered a sum of n Bernoulli trials so that X/n is a sample mean).

Second, the observed information increases at the rate of n, and in fact (for i.i.d. samples)

$$\frac{1}{n} I_{OBS}(\hat{\theta}) \xrightarrow{P} I_F(\theta). \tag{16.20}$$

(Compare the details following Eq. (8.35).) For example, in the binomial case we have

$$I_{OBS}(\hat{\theta}) = \frac{n}{\hat{\theta}(1 - \hat{\theta})}$$

and

$$I_F(\theta) = \frac{1}{\theta(1 - \theta)}.$$

---

[5] In order for the right-hand side to integrate to 1 we must have $k = \sqrt{-\ell''(\hat{\theta})/2\pi}$.

[6] Additional details may be found in many sources including Kass and Vos (1997, Theorem 2.2.13) and Chen (1985).

so that (16.20) holds due to (16.19) (by an application of Slutsky's theorem). Because the observed information increases as $n \to \infty$, the loglikelihood function becomes more highly peaked about its maximum as $n \to \infty$. Furthermore, the likelihood function may be approximated by the normal pdf found by exponentiating the right-hand side of (16.16), which has standard deviation $SE = I_{OBS}(\hat{\theta})^{-1/2}$ (the standard error associated with the MLE). From Eq. (16.20), this standard deviation is decreasing as $1/\sqrt{n}$. Therefore the width of the peak in the likelihood function is decreasing at the rate of $1/\sqrt{n}$. We can write the values of θ for which the likelihood is substantial in the form

$$(a_n, b_n) = (\hat{\theta} - c \cdot SE, \hat{\theta} + c \cdot SE) \tag{16.21}$$

where we take c to be a positive constant, such as $c = 4$.

Finally, we consider the contribution of the prior to the posterior as $n \to \infty$. As in (16.19), the posterior pdf is a normalized product of the likelihood function and prior pdf. While the likelihood function becomes increasingly close to the form of a normal pdf, with standard deviation decreasing as $1/\sqrt{n}$, the prior pdf $\pi(\theta)$ is a fixed function that does not change with n. The intervals in (16.21) will have lengths $b_n - a_n$ that decrease as $1/\sqrt{n}$ and, for θ in these intervals, when n is large the value of $\theta - \hat{\theta}$ will be small so that we get

$$\pi(\theta) \approx \pi(\hat{\theta}). \tag{16.22}$$

In other words, for values "within the peak" of the loglikelihood function, the prior is approximately constant. Therefore, (16.16) gives us (16.17) and the posterior becomes concentrated near θ with an approximately normal form, as in (16.18). We have sketched the argument in the scalar case, but the steps are the same when θ is a vector.

The approximate $N(\hat{\theta}, I_{OBS}(\hat{\theta})^{-1})$ distribution of the posterior not only gives an easy way to compute posterior probabilities, for large samples, but it also provides a very nice mathematical expression of one of the guiding principles of science: any two investigators who start with differing beliefs (in the form of two different priors $\pi_1(\theta)$ and $\pi_2(\theta)$), will, with sufficiently much data, come to agreement (their posterior distributions will be essentially the same).

16.1.6 Powerful methods exist for computing posterior distributions.

In our introduction to this chapter we reviewed briefly the conceptual appeal of Bayesian inference. Bayesian methods have become indispensable in the analysis of neural data mainly because (i) inferences agree reasonably well with those based on ML estimation (due to the result outlined in Section 16.1.5), (ii) sometimes

there is available structure that can be formalized as part of a prior specification (see Section 16.2), and (iii) in many complicated statistical models there are general computational tools for computing posteriors. In this section we sketch the essential ideas behind the main such computational tool, *posterior simulation*.

We have already, in Chapter 9, described the great utility of simulation methods in statistical inference. In posterior simulation a sequence $\theta^{(1)}, \ldots, \theta^{(G)}$ of observations from the posterior distribution is generated, and inference is based on the methods outlined in our discussion of simulation-based propagation of uncertainty (p. 225). For example, to compute the probability in

$$P(a < \theta < b|x) = \int_a^b f(\theta|x)d\theta \tag{16.23}$$

we could use

$$P(a < \theta < b|x) \approx \frac{N_1}{G}$$

where N_1 is the number of $\theta^{(g)}$ such that $a < \theta^{(g)} < b$. Similarly, if $\phi = f(\theta)$ for some function $f(x)$ we could compute probabilities involving ϕ (again, as on p. 225), as

$$P(c < \phi < d|x) \approx \frac{N_2}{G}$$

where we let $W^{(g)} = f(\theta^{(g)})$ and N_2 is the number of $W^{(g)}$ such that $c < W^{(g)} < d$. This kind of computation is used in the following example.

Example 16.2 Methylphenidate-Induced Emergence from General Anesthesia When general anesthesia is administered for surgery, or for an invasive diagnostic procedure, patients recover by resting until the anesthesia's effects wear off. As an alternative, Solt et al. (2011) considered the possibility that methylphenidate might induce emergence from general anesthesia. Methylphenidate (Ritalin) is widely used to treat Attention Deficit Hyperactivity Disorder (ADHD), and acts primarily by inhibiting dopamine and norepinephrine reuptake. But dopamine and norepinephrine can also promote arousal. The authors applied isoflurane anesthesia to rats at a dose sufficient to maintain them in a supine position (lying down) for 40 min. Five minutes after establishing their anesthetized state (from an equilibration procedure) the animals were given one of three doses of methylphenidate intravenously ranging from a maximum of 5 mg/kg to a minimum of .05 mg/kg. At the maximum dose, 12 out of 12 rats regained their upright position and made purposeful movements within 30 s (seconds) of drug administration. At the minimum dose, 0 out of 6 regained their upright position within 30 min. Apparently, 5 mg/kg of methylphenidate is sufficient to remove the immobilizing effects of isoflurane-induced anesthesia in rats. (At the intermediate dose of .5 mg/kg 11 out of 12 regained their upright position.)

To evaluate the strength of this evidence, 12/12 versus 0/6, the authors considered the binomial model $X_1 \sim B(12, p_1)$ and $X_2 \sim B(6, p_2)$, introduced independent

uniform priors on p_1 and p_2 as in Example 1.4 on p. 174, and computed the posterior probability $P(p_1 > p_2|X_1 = 12, X_2 = 0)$. This may be done very easily by posterior simulation: the two posterior distributions on p_1 and p_2 are $Beta(13, 13)$ and $Beta(1, 7)$, and they are independent. (It is easy to check that if X_1 and X_2 are independent, and the prior distributions on p_1 and p_2 are independent, then the posterior distributions on p_1 and p_2 are independent.) We therefore do the following:

1. Draw $G = 10,000$ observations from a $Beta(13, 13)$ distribution and put them in a vector A.
2. Draw $G = 10,000$ observations from a $Beta(1, 7)$ distribution and put them in a vector B.
3. Compute the number of components i for which $A[i] > B[i]$, and divide by G. This is, approximately, the desired posterior probability.

Performing the calculation gives $P(p_1 > p_2|X_1 = 12, X_2 = 0) = .986$. The authors concluded that methylphenidate actively induces emergence from isoflurane anesthesia. We re-evaluate this evidence using Bayes factors on p. 478. □

In Section 16.1.2 we noted that posterior probabilities may be computed easily when conjugate priors are used, and Example 16.2 made use of posterior simulation with conjugate beta posterior distributions. As soon as we leave conjugacy, numerical difficulties become apparent. Even in the simple case of estimating a normal mean θ from a sample $X_1, \ldots, X_n$, with $X_i \sim N(\theta, \sigma^2)$ and σ known, if we take the prior to be a non-normal probability distribution on $(-\infty, \infty)$, the posterior pdf becomes intractable, in the sense that $L(\theta)\pi(\theta)$ in Eq. (16.1) has a non-normal form, and we can not evaluate analytically the integrals needed to compute posterior probabilities such as that in (16.23). The usual approach to solving this problem is to apply posterior simulation based on *Markov chain Monte Carlo* (*MCMC*).

The nomenclature is descriptive of the idea behind MCMC: "Monte Carlo" refers to[7] simulation methods, and "Markov chain Monte Carlo" indicates that the approach is based on Markov chains. To explain, we begin by returning to an example.

Example 3.5 (Continued, see p. 58) In our discussion of Colquhoun and Sakman's results on ion channel openings we noted from Fig. 3.8, panel B, the good fit of an exponential distribution to the histogram of open durations, when there was only one opening in an activation burst. The major purpose of the paper was to demonstrate the existence of activation bursts. Let us, however, ignore bursts and imagine an ideal ion channel that opens and closes with open and closed durations governed by exponential distributions. The defining property of exponential distributions is that they are memoryless (see the theorem on p. 120). Now consider an ion channel that is observed to be either open or closed for a sequence of discrete time values, e.g., every ms for 10 min, and let $X_t = 1$ if it is open at time t and $X_t = 0$ if it is closed at time t. We refer to the channel's *state* at time t as the value of X_t, with 1 or 0

[7] When computer-based simulation methods were first being used, Monte Carlo was the site of a famous gambling establishment, which was frequented by the uncle of one of the developers of these methods. See Metropolis (1987).

signifying either open or closed. If we assume[8] the ion channel is memoryless, then its state at time $t+1$ will depend on its state at time t, but not on any of the preceding states prior to time t. There are then four possibilities: the channel can be closed at time t and stay closed at $t+1$, it can be closed at t and be open at $t+1$, it can be open at t and close at $t+1$, or it can be open at t and stay open at $t+1$. The four possibilities have conditional probabilities given by $P(X_{t+1} = j|X_t = i)$ where i and j can take values 0 or 1. □

Abstracting from this example, suppose we have a sequence of random variables $X_1, X_2, \ldots, X_t, \ldots$, which take values 0 and 1, and suppose further that

$$P(X_{t+1}|X_1, X_2, \ldots, X_t) = P(X_{t+1}|X_t) \tag{16.24}$$

and that these conditional probabilities are time-invariant in the sense that

$$P(X_{t+1} = j|X_t = i) = P(X_{s+1} = j|X_s = i)$$

for all $s, t = 0, 1, 2, \ldots$. Then the sequence $X_1, X_2, \ldots, X_t, \ldots$ is said to form a two-state *Markov chain* having *transition probabilities* $P(X_{t+1} = j|X_t = i)$, which we write as

$$P_{ij} = P(X_{t+1} = j|X_t = i). \tag{16.25}$$

Let us note that (16.25) implies

$$P(X_{t+1} = 0) = P(X_t = 0)P_{00} + P(X_t = 1)P_{10} \tag{16.26}$$
$$P(X_{t+1} = 1) = P(X_t = 0)P_{01} + P(X_t = 1)P_{11}. \tag{16.27}$$

The definition extends immediately to the case of m states, for m an integer with $m \geq 2$.

The key property of a Markov chain is its lack of memory: the probability of being in state j at time $t+1$ depends only on the state of the chain at time t. Under some mild conditions[9] it is possible to say something about the long-run behavior of the chain. In the case of the ideal ion channel considered above, we may ask for the probability that the channel is open at time $t = 600{,}000$, corresponding to 10 min after the commencement of observation. In principle this probability depends on the initial condition, whether the channel was open at time $t = 1$. However, because the state at time $t = 600{,}000$ is the result of 599,999 random draws from the distributions given by the transition probabilities (16.25) the influence of the initial

[8] Because we are assuming discrete time the memoryless distribution of durations becomes geometric rather than exponential, as we noted on p. 120.

[9] The chain must be *irreducible* (if the chain is in state i at time t it is possible for it to get to state j in the future), *aperiodic* (the chain does not cycle deterministically through the states), and *recurrent* (if the chain is in state i at time t it will eventually return to state i in the future), see for example, Ross (1996, Theorem 4.3.3).

state is miniscule.[10] Thus, we have a limiting distribution which we write as

$$\lim_{t\to\infty} P(X_t = 1) = P_\infty(1)$$

and $P_\infty(0) = 1 - P_\infty(1)$. This limiting distribution is called the *stationary distribution* because it satisfies

$$P_\infty(0) = P_\infty(0)P_{00} + P_\infty(1)P_{10} \tag{16.28}$$
$$P_\infty(1) = P_\infty(0)P_{01} + P_\infty(1)P_{11}. \tag{16.29}$$

Comparing (16.28) and (16.29) to (16.26) and (16.27) we see that if at any time t the chain satisfies $P(X_t = 1) = P_\infty(1)$ (and thus also $P(X_t = 0) = 1 - P(X + t = 1) = 1 - P_\infty(1) = P_\infty(0)$) it continues to satisfy $P(X_{t+h} = 1) = P_\infty(1)$ for all positive h. In words, once the chain reaches its stationary distribution, it stays there. Again, all of this extends immediately to the case of m states.

The existence of a stationary distribution for a Markov chain is profoundly important for posterior simulation.[11] If a simulation procedure is set up as a sequence of random draws that form a Markov chain, and if the chain has the posterior as its stationary distribution, then once the chain runs for a sufficiently long time that it reaches its stationary distribution, it will thereafter be generating observations from the posterior distribution. This is the idea behind MCMC.

Let us backtrack just a little to acknowledge that we are glossing over the distinction between discrete and continuous distributions: in our discussion here we are considering the discrete case, with discrete states (in fact, two states), while for posterior simulation we would usually be concerned with a continuous posterior distribution.[12]

There is a remarkably simple algorithm that creates a Markov chain having the posterior as its stationary distribution. It is the *Metropolis-Hastings algorithm* (Hastings 1970; Metropolis et al. 1953).

Let $q(u|v)$ be a pdf of an m-dimensional random vector U that depends on an m-dimensional vector v. For example, $q(u|v)$ could be an m-dimensional multivariate normal pdf with mean v. Suppose we have simulated an m-dimensional parameter vector $\theta^{(g)}$. The Metropolis algorithm proceeds by generating a *candidate* vector $\theta_c^{(g+1)}$ from the *proposal* pdf $q(\theta_c^{(g+1)}|\theta^{(g)})$ and then either accepts $\theta_c^{(g+1)}$ as the next

[10] It is not too difficult to derive the formula, but we omit the arithmetic. For large t the probability of the channel being open is

$$P(X_t = 1) \approx \frac{P_{01}}{P_{01} + P_{10}}$$

which depends on the probability P_{01} of switching from closed to open relative to the probability P_{10} of switching from open to closed.

[11] It is also very important in many other situations, where Markov chains are used as statistical models.

[12] Details concerning the continuous case may be found in many sources (for example, Robert and Casella 2004) for a parameter θ.

simulated vector, $\theta^{(g+1)} = \theta_c^{(g+1)}$, or rejects it and sets $\theta^{(g+1)} = \theta^{(g)}$. Acceptance is determined probabilistically from the quantity

$$\alpha^{(g+1)} = \frac{L(\theta_c^{(g+1)})\pi(\theta_c^{(g+1)})q(\theta^{(g)}|\theta_c^{(g+1)})}{L(\theta^{(g)})\pi(\theta^{(g)})q(\theta_c^{(g+1)}|\theta^{(g)})}. \tag{16.30}$$

Metropolis-Hastings Algorithm:
Initialize by setting a value of $\theta^{(0)}$.
For $g = 0, \ldots, G-1$, draw a candidate $\theta_c^{(g+1)}$ from the proposal pdf $q(\theta_c^{(g+1)}|\theta^{(g)})$ and compute $\alpha^{(g+1)}$ from (16.30), then

$$\texttt{if } \alpha^{(g+1)} > 1 \texttt{ set } \theta^{(g+1)} = \theta_c^{(g+1)}$$
$$\texttt{otherwise}$$
$$\texttt{set } \theta^{(g+1)} = \begin{cases} \theta_c^{(g+1)} & \text{with probability } \alpha^{(g+1)} \\ \theta^{(g)} & \text{with probability } 1 - \alpha^{(g+1)} \end{cases}$$

In theory the Metropolis-Hastings algorithm "works" in the sense that it eventually converges to its stationary distribution, and then starts generating observations from the posterior. In practice, however, the number of iterations leading to convergence, often called the *burn-in* period, is crucially important and depends on the choice of the proposal pdf. Notice that if the candidate proposal pdf is equal to the posterior pdf, then $\alpha^{(g+1)} = 1$ so that every candidate is accepted and convergence is immediate (for $g \geq 1$, every $\theta^{(g)}$ is simulated from the posterior distribution). In practice, the proposal pdf is chosen for convenience, but with the hope that it will provide at least a rough approximation to the posterior pdf and a reasonable fraction of the candidates will be accepted. If the dimension of θ is small, it is not hard to find a proposal distribution for which the number of iterations needed to reach convergence is manageable (perhaps several hundred or a few thousand) in the sense that the necessary computing time is tolerable. Furthermore, especially when the statistical model has an advantageous structure, general-purpose MCMC algorithms, which are variations on Metropolis-Hastings, can be effective. We return to this comment at the end of Section 16.2.2.

A second practical concern with MCMC is that it is necessary to remove the simulated values $\theta^{(g)}$ that occur during burn-in, and retain only those that are generated after convergence. The data analyst must, therefore, apply some method aimed at determining when convergence is reached. These and other issues are addressed in the literature (see Robert and Casella (2004), and references therein).

Derivation of the Metropolis-Hastings algorithm:
We consider the discrete case and begin with an arbitrary target distribution, which we label P_∞. We wish to construct a Markov chain such that P_∞ is the stationary distribution for the chain. We start with

another Markov chain having transition probabilities $\{Q_{ij}\}$ and show how to modify it so that we obtain a chain with transition probabilities $\{P_{ij}\}$ having P_∞ as its stationary distribution. The idea is that $\{Q_{ij}\}$ would represent transition probabilities for the chain based on the candidate proposal distributions, and $\{P_{ij}\}$ would be the transition probabilities for the resulting Metropolis-Hastings chain.

First, let us note that a distribution P_∞ is said to satisfy *detailed balance* for a Markov chain having transition probabilities $\{P_{ij}\}$ if for all i, j, $P_\infty(i)P_{ij} = P_\infty(j)P_{ji}$. If P_∞ satisfies detailed balance then we also have

$$\begin{aligned} P_\infty(j) &= P_\infty(j) \sum_i P_{ji} \\ &= \sum_i P_\infty(j) P_{ji} \\ &= \sum_i P_\infty(i) P_{ij} \end{aligned}$$

which shows that P_∞ is a stationary distribution for the chain. Thus, detailed balance implies stationarity.

Now suppose P_∞ is a given target distribution and we have available a Markov chain with transition probabilities $\{Q_{ij}\}$. This available chain will be used to generate candidates. We would like to define a Markov chain for which P_∞ is its stationary distribution. If detailed balance were satisfied by P_∞ for the chain with transition probabilities $\{Q_{ij}\}$ then we would be done. If not, then $P_\infty(i)Q_{ij} \neq P_\infty(j)Q_{ji}$. For definiteness, suppose

$$P_\infty(i)Q_{ij} > P_\infty(j)Q_{ji}. \tag{16.31}$$

We wish to construct transition probabilities $\{P_{ij}\}$ such that equality holds in Eq. (16.31) when $\{P_{ij}\}$ is substituted for $\{Q_{ij}\}$. Examining (16.31), we need to make the values P_{ij} that we will substitute on the left-hand side of (16.31) smaller than Q_{ij} while setting $P_{ji} = Q_{ji}$ for the values we will substitute on the right-hand side of (16.31). Toward this end, we introduce a set of numbers α_{ij} with $0 < \alpha_{ij} \leq 1$, and define $P_{ij} = Q_{ij}\alpha_{ij}$. To make sure we are setting $P_{ji} = Q_{ji}$ for the values we will substitute on the right-hand side of (16.31), we make the restriction

$$\begin{aligned} P_{ij} &= Q_{ij}\alpha_{ij} \quad \text{when} \quad (16.31) \quad \text{holds} \\ P_{ij} &= Q_{ij} \quad \text{otherwise} \end{aligned}$$

(because the Q_{ji} that appear on the right-hand side of (16.31) are simply those that appear on the left-hand side when the inequality is reversed). We then solve for the values of α_{ij} that produce equality in (16.31). That is, we write

$$P_\infty(i)Q_{ij}\alpha_{ij} = P_\infty(j)Q_{ji}$$

and solve for α_{ij}:

$$\alpha_{ij} = \frac{P_\infty(j)}{P_\infty(i)}\frac{Q_{ji}}{Q_{ij}}. \tag{16.32}$$

Note that $\alpha_{ij} < 1$ if and only if (16.31) holds.
We have now produced a specification of the transition probabilities required for a chain having the target distribution P_∞ as stationary distribution: if (16.31) holds, set $P_{ij} = Q_{ij}\alpha_{ij}$, where α_{ij} are defined by (16.32), otherwise set $P_{ij} = Q_{ij}$. To create a chain with these transition probabilities is easy. Suppose the chain is in state i. If we accept the candidate (which is generated from the chain having transition probabilities $\{Q_{ij}\}$) with probability α_{ij}, then the probability of moving to state j will be $Q_{ij}\alpha_{ij}$. Thus, we use the following scheme:

define α_{ij} by (16.32)
if $\alpha_{ij} < 1$ then accept the candidate with probability α_{ij}
otherwise accept the candidate.

This is the Metropolis-Hastings algorithm. □

16.2 Latent Variables

When we introduced the concept of random variable (on p. 46) we were careful to distinguish the mathematical object from the data: we said that random variables and their probability distributions live in the theoretical world of mathematics while data live in the real world of observations. Random variables that are theoretical counterparts of observed data are sometimes called *observable*. But it is also possible to insert into a statistical model random variables that affect the distribution of the observable random variables without themselves representing data; instead they represent *unobserved*, hypothetical quantities. Such unobservable random variables are called[13] *latent variables*. Models that incorporate latent variables can be powerful

[13] The noise random variable ϵ_i in the regression model (12.1) is unobservable, but would not typically be called latent. To exclude such cases a random variable could be called latent only if it can not be written in terms of observable random variables. Thus, under this definition, because (12.1) implies $\epsilon_i = Y_i - f(x_i)$, ϵ_i would not be a latent variable. See Bollen (2002).

Fig. 16.1 First 3 s of spikes recorded over about 30 s in vitro, from a goldfish retinal ganglion neuron. Data from Levine (1991), furnished by Satish Iyengar; see Iyengar and Liao (1997). These data are discussed in Example 19.1.

and intuitive ways to describe variation in the data. We discussed the mixture-of-two-Gaussians model on p. 216, and we will return to mixtures of Gaussians in Section 17.4.3. Here is another example.

Example 16.3 Burst Detection from spike trains In many contexts neurons exhibit burstiness, meaning that action potentials (also called "spikes," see p. 3), appear across time in small clusters, or bursts. For instance, burstiness of dopamine neurons in the midbrain is believed to be a functionally relevant signal indicating reward and goal-directed behavior (see Grace et al. 2007). In the analysis of bursting neural spike trains, the epochs during which the neuron is bursting must somehow be inferred from the data. In Fig. 16.1, for example, due to the inherently erratic nature of the spiking, it is not always obvious whether the neuron is in a burst or not, or where a burst begins and ends.

To provide an algorithm together with statistical inferences, Tokdar et al. (2010) described bursty neurons by introducing a latent binary variable, which was 1 when the neuron was bursting and 0 when it was not bursting. Let us assume the recording time to occur in discrete steps $t = 1, 2, \ldots, T$, and define the random variable Y_t be 1 if a spike occurs at time t and 0 otherwise. Tokdar et al. discussed several alternative models. The simplest uses latent variables X_t that take the value 1 if the neuron is bursting at time t and 0 if non-bursting, and assumes that Y_t has a Bernoulli pdf with mean θ_1 if $X_t = 1$ and with mean θ_0 if $X_t = 0$. Here, θ_1 and θ_0 represent the firing rates of the neuron when bursting and when not bursting, and if θ_1 is much larger than θ_0 the neuron will tend to fire in rapid succession when $X_t = 1$, compared with its slower rate when $X_t = 0$. This describes the tendency to produce bursts. By introducing probability distributions for the latent variables X_t, and then estimating the value of each X_i, it is possible to infer where in time the bursts occurred.[14] □

In most statistical models the distribution of the random variables representing the data depends on some unknown parameters. In the model cited in Example 16.3 the distributions of the random variables representing the data depended on unknown parameters, but they also depended on the latent variables. The point is that the latent variables themselves followed probability distributions. In other words, one set of probability distributions—those describing the variation in the data—depended on random variables following another set of probability distributions, which described

[14] To speed computation Tokdar et al. chose to work with the inter-spike intervals instead of the variables Y_t we have defined here.

variation among certain theoretically interesting but unobserved quantities, namely the bursting or non-bursting status of the neuron within each ISI.

The parameters in statistical models are usually fixed coefficients (though they are typically unknown, and therefore estimated from the data). In Section 16.2.1 we describe models in which the parameters become random variables, and thus latent variables. We then briefly re-interpret penalized regression in Section 16.2.3 and return to the general structure underlying Example 16.3 in Section 16.2.4.

16.2.1 Hierarchical models produce estimates of related quantities that are pulled toward each other.

Nearly all the statistical models we have considered[15] begin with a parameterized family of probability densities $f(x|\theta)$, and the first statistical problem is to determine from the data x the likely values of the parameter θ. Sometimes there is an obvious source of variability among values of the parameter θ, as when θ could vary from subject-to-subject, or neuron-to-neuron, etc. In such cases we may introduce a second layer into the statistical model by considering a family of densities $f(\theta|\lambda)$. For generality, we will refer to individual subjects or individual neurons, etc., as *units*. In other words, we will say that we are interested in the variation of some parameter θ across units. In neuroimaging, for example, we might have task-related effects at particular voxels whose magnitude varies across subjects, and these could be assumed to follow some probability distribution. In analyzing neural responses, the way a particular measure of neural activity varies across neurons may be of interest, and might be assumed to follow a given probability distribution. In these situations we introduce both a probability density $f(x|\theta)$ for the data given a parameter vector θ and a probability density $f(\theta|\lambda)$ for θ that itself depends on a parameter λ. Such a specification is called a two-stage[16] *hierarchical model*.

Example 12.3 (continued from p. 331) As described previously, Behseta et al. considered spike counts from 54 neurons during performance of a serial-order eye-movement task, and the authors computed a rank order selectivity index

$$I_{\text{rank}} = \frac{(f_3 - f_1)}{(f_3 + f_1)}$$

where f_1 and f_3 were the mean firing rates measured at the times of the first and third saccades respectively, the mean being taken across trials. As part of the analysis, the rank selectivity indices across neurons were considered to follow a normal distribution. Let X_i represent I_{rank} for neuron i. Behseta et al. assumed a model of the

[15] Nonparametric methods (Section 13.3) are based on statistical models of a more general form that do not depend on a finite-dimensional parameter vector.

[16] In principle this process can continue, with λ distributed according to a family of densities, and so on, but they do not arise very often in practice.

form

$$X_i \sim N(\theta_i, \sigma_i^2) \tag{16.33}$$

$$\theta_i \sim N(\mu, \tau^2). \tag{16.34}$$

Here, θ_i is the theoretical mean of the the rank order selectivity index for neuron i and σ_i^2 is its variance. The value of θ_i becomes a quantity to be estimated, but is here considered to follow a distribution across the population of neurons, with population mean μ and variance τ^2. □

Equations (16.33) and (16.34) are an instance of a general structure. Let X_i be a random vector representing measurements made on unit i. For instance, in Eqs. (16.33) and (16.34) of Example 12.3, we took X_i to be the rank order selectivity index for neuron i. We assume the observations X_i, and the parameters θ_i, are *conditionally independent* across units, with variation being described by a two-stage hierarchical model:

Stage one: Conditionally on $(\theta_1, \ldots, \theta_k)$ and λ, the vectors X_i are independent with pdfs $f(x_i|\theta_i, \lambda)$, $i = 1, \ldots, k$, belonging to a family $\{f(x|\theta, \lambda)\}$.

Stage two: Conditionally on λ, the vectors θ_i are i.i.d. with pdf belonging to a family $\{f(\theta|\lambda)\}$.

In general, θ and λ are multidimensional, and we obtain what is sometimes called a *conditionally independent hierarchical model*. In (16.33) we would have $\lambda = (\mu, \tau)$.

In (16.33) and (16.34) let us pick a particular unit, with label h, so that $i = h$. If we were to consider the data $X_h = x_h$ in isolation, without reference to all the other data values x_i for $i \neq h$, we would estimate θ_h as x_h. The hierarchical model suggests something different: it assumes that the values of θ_i are related to each other, according to the second stage of the model, and the posterior therefore uses data *from the other units* in estimating θ_h. Tukey referred[17] to this as "borrowing strength." It would be especially valuable if σ_h^2 happened to be large, possibly due to a very small number of trials for that neuron; in this case the strength of the signal in x_h would be small, and additional strength for estimating θ_h would come from the other units.

Great insight is obtained by examining the formulas for the posterior distribution of θ_i in the normal hierarchical model in Eqs. (16.33) and (16.34), where we assume τ known but μ unknown, and we let μ have the improper uniform prior $\pi(\mu) = 1$. We will call this the *canonical normal hierarchical model*. In most practical cases τ is unknown and must be estimated, but we are ignoring that complication for the time being.

[17] See, for example, his 1973 article "Exploring data analysis as part of a larger whole," reprinted in Tukey (1987).

Result for the Canonical Normal Hierarchical Model
Suppose $X_i \sim N(\theta_i, \sigma_i^2)$, independently and $\theta_i \sim N(\mu, \tau^2)$ i.i.d. for $i = 1, \ldots, k$ with the σ_i's and τ known but μ unknown, and suppose μ has the improper uniform prior. The posterior distribution of θ_i given $x = (x_1, \ldots, x_k)$ is normal with

$$E(\theta_i|x) = \frac{\tau^2}{\sigma_i^2 + \tau^2} x_i + \frac{\sigma_i^2}{\sigma_i^2 + \tau^2} \bar{x}_\alpha \tag{16.35}$$

$$V(\theta_i|x) = \left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right)^{-1} + \left(\sum_i \frac{1}{\sigma_i^2 + \tau^2}\right)^{-1} \left(\frac{\sigma_i^2}{\sigma_i^2 + \tau^2}\right)^2 \tag{16.36}$$

where

$$\bar{x}_\alpha = (\sum_i \alpha_i x_i)/(\sum_i \alpha_i) \tag{16.37}$$

with $\alpha_i = (\sigma_i^2 + \tau^2)^{-1}$.

We postpone the derivation of the result until Section 16.4.

The expression for the posterior mean in Eq. (16.35) is beautifully simple. Note first that we may consider each random variable X_i to arise as a compound distribution, obtained by first drawing θ_i from a $N(\mu, \tau^2)$ distribution and then drawing X_i from a $N(\theta_i, \sigma_i^2)$. Conditionally on θ_i the random variable X_i has the same distribution as

$$Y_i = \theta_i + \epsilon_i \tag{16.38}$$

where $\epsilon_i \sim N(0, \sigma_i^2)$ and ϵ_i is independent of θ_i. In (16.38) we use $\theta_i \sim N(\mu, \tau^2)$ and, holding μ and τ fixed, compute the variance to get $V(X_i) = V(Y_i) = \sigma_i^2 + \tau^2$. The weights appearing in (16.37) are thus $\alpha_i = V(X_i)^{-1}$ and, therefore, $\bar{x}_\alpha$ is the usual weighted mean of Eq. (8.12). Keeping this in mind, we re-express the posterior mean as

$$E(\theta_i|x_i) = w_i x_i + (1 - w_i)\bar{x}_\alpha \tag{16.39}$$

where

$$w_i = \frac{\tau^2}{\sigma_i^2 + \tau^2}.$$

Comparing (16.39) with (16.12) we see that we have a very similar interpretation. In (16.12) the posterior mean was a weighted combination of the data value x and the prior mean μ, and the posterior mean resulted from shrinking the value x toward μ. In (16.39) the posterior mean is a weighted combination of the data value x_i and the data mean $\bar{x}_\alpha$, so posterior mean results from shrinking the value x_i toward $\bar{x}_\alpha$. Here the weight takes essentially the same form as in (16.12), with σ_i substituted for σ.

According to (16.34) the values θ_i are all related to each other and, from (16.33), every X_i will contribute some information about the value of θ_j. When there is large uncertainty about θ_i based on x_i alone, relative to the variability among the θ_i values so that τ/σ_i is small, w_i is small, the values of x_j for $j \neq i$ become very relevant to the estimation of θ_i, and the posterior mean is nearly equal to $\overline{x}_\alpha$. On the other hand, when x_i contributes a lot of knowledge about θ_i, relative to the variability among the θ_i values so that τ/σ_i is large, w_i is large and the posterior mean is nearly equal to x_i. As on p. 445, we say the posterior mean *shrinks* x_i toward $\overline{x}_\alpha$ with the amount of *shrinkage* determined by $1 - w_i$.

Example 12.3 (continued from p. 459) In the case of rank order selectivity index, each value of σ_i could be estimated directly from the data and was therefore taken to be known. But these values varied across neurons. Some neurons could have highly variable X_i, and thus poorly-determined values of θ_i, while other neurons could have less variable X_i and better-determined values of θ_i. The posterior mean (16.39) takes into account the diversity of precision in the selectivity index across neurons. We continue our analysis on p. 464. □

Example 16.4 Genetic Linkage Across Multiple Related Strains Bacterial meningitis is an inflammation of the meninges, the membranes that cover the brain and spinal cord. Prior to antibiotics it was usually fatal. In outbreaks of bacterial meningitis, the antibiotic rifampicin is highly effective (Gaunt and Lambert 1987). However, bacteria can mutate to become resistant to rifampicin. An experiment on mutation mechanisms in *E. Coli* concerned the genetic linkage between resistance to rifampicin and a neighboring gene known as uvrE which, when absent, produces acetate utilization deficiency. One result of this work was to show uvrE to be involved in DNA repair. Sklar and Strauss (1980) noted that if the acetate utilization deficiency mutation occurred during DNA replication then it would be linked to rifampicin resistance but if, instead, the mutation resulted from error-prone DNA repair there would be no such linkage. These investigators created two cell lines, one selected for rifampicin resistance and the other not selected. The absence of linkage, predicted by the error-prone repair hypothesis, would imply that the proportions p_1 and p_2 exhibiting acetate utilization deficiency in the selected and unselected cell lines would be equal. The authors looked at 13 closely-related strains of *E. coli*, all of which sometimes exhibited acetate utilization deficiency. We will return to these data in Section 16.3, where we describe evidence in favor of $H_0 : p_1 = p_2$ for the uvrE strain. Here we describe an analysis reported by Kass and Steffey (1989), who evaluated the difference between the proportions p_{i1} and p_{i2} on the logit scale (as used in logistic regression, p. 394). That is, for $i = 1, \ldots, 13$ they defined

$$\theta_i = \log \frac{p_{i1}}{1 - p_{i1}} - \log \frac{p_{i2}}{1 - p_{i2}} \tag{16.40}$$

and estimated $\theta_1, \ldots, \theta_{13}$. Because the strains were related, the data for strain j provided at least some relevant information about the value of θ_i, even when $i \neq j$.

Table 16.1 Observed difference in logits together with posterior means among 13 closely-related strains of *E. coli*. The values x_i are the observed differences of logits, σ_i are the corresponding standard errors, and $E(\theta_i|x)$ are the posterior means from (16.39), where τ^2 was set equal to .39, which was the MLE. The weighted mean was $\bar{x}_\alpha = 1.30$.

Strain	x_i	σ_i	$E(\theta_i\|x)$
1	1.36	.28	1.35
2	2.26	1.04	1.56
3	2.23	.75	1.68
4	1.32	.36	1.31
5	1.21	.38	1.24
6	1.27	.49	1.28
7	1.43	.57	1.37
8	1.85	.54	1.62
9	1.34	.56	1.30
10	3.44	.73	2.20
11	−0.42	.69	.53
12	−0.10	.31	.17
13	1.25	.39	1.27

The raw data for each strain were two pairs of sample sizes and corresponding proportions $(n_{i1}, \hat{p}_{i1})$ and $(n_{i2}, \hat{p}_{i2})$. Kass and Steffey assumed the data to be distributed as binomial proportions and transformed to the logit scale according to

$$X_i = \log \frac{\hat{p}_{i1}}{1 - \hat{p}_{i2}} - \log \frac{\hat{p}_{i2}}{1 - \hat{p}_{i1}} \tag{16.41}$$

taking σ_i^2 to be known and equal to the value obtained from the large-sample variance formula given on p. 231 (based on propagation of uncertainty) which, after simplifying, yields

$$\sigma_i^2 = \frac{1}{n_{i1}\hat{p}_{i1}} + \frac{1}{n_{i1}(1 - \hat{p}_{i1})} + \frac{1}{n_{i2}\hat{p}_{i2}} + \frac{1}{n_{i2}(1 - \hat{p}_{i2})}. \tag{16.42}$$

The transformed data, together with values of σ_i are shown in the first two columns of Table 16.1. From preliminary analysis, it appeared (as seen in Table 16.1) that p_{i1} was greater than p_{i2} in most strains. In two strains, however, $\hat{p}_{i1}$ was less than $\hat{p}_{i2}$ and the issue was whether this was due to sampling fluctuation (insufficiently large $ni1$ and n_{i2}) or a genuinely different phenomenon for either or both of the two strains in question. The results for the canonical normal hierarchical model (defined on p. 460) shed light on the issue. Model (16.33) may be applied to the variables X_i defined in (16.41) where the parameters θ_i and σ_i are defined by (16.40) and (16.42), and then (16.34) may be assumed. The value of τ was fixed with $\tau^2 = .39$ (which is the MLE, as we discuss below), and then (16.35) produces the posterior means, which are shown in the third column of Table 16.1.

The values in the table exhibit the effect of σ_i on the shrinkage behavior described above. For example, strain 2 has nearly the same value of x_i as strain 3 but it also has a larger standard error σ_i. This leads the posterior mean of strain 2 to shrink closer to $\bar{x}_\alpha = 1.30$ than that for strain 3 (1.56 is closer than 1.68). Similarly, x_i for strain 8 is quite a bit smaller than that for strain 2, but its standard error is also much smaller,

and the posterior mean for strain 2 shrinks closer to $\bar{x}_\alpha$ than that for strain 8 (1.56 is closer than 1.62). Similarly, because strain 11 has a much larger standard error than strain 12, strain 11 starts out at $x_i = -.42$ and shrinks to .53 while strain 12 starts out at $x_i = -.10$ but only shrinks to .17. These latter two strains are especially interesting because, under the repair hypothesis, the difference of the logits should be zero. In fact, as we discuss in Section 16.3.1, further analysis suggests that the repair hypothesis holds for strain 12 and not[18] for strain 11. □

The canonical normal hierarchical model on p. 460 assumes τ is known, and on p. 463 we said that we fixed its value using $\tau^2 = .39$. This value was obtained as the MLE. It is easy to write down the likelihood function $L(\lambda)$ on $\lambda = (\mu, \tau)$ after integrating out the parameters $\theta_1, \ldots, \theta_{13}$ (we have $X_i \sim N(\mu, \sigma_i^2 + \tau^2)$, independently, for $i = 1, \ldots, 13$), and $L(\lambda)$ may be maximized.[19] The MLE of μ is $\hat{\mu} = \bar{x}_\alpha$. This approach, using maximum likelihood for the second-stage parameter vector (the hyperparameter) λ, is often called *empirical Bayes*, to distinguish it from *fully Bayes*, which would instead introduce a prior on λ and then compute the posterior in a more elaborate model. In (16.33) and (16.34) this involves introducing a prior on τ as well as μ. With the fully Bayes approach the posterior means $E(\theta_i|x)$ would no longer have analytical expressions; instead, posterior simulation would typically be used to evaluate the posterior means, using methods outlined in Section 16.1.6. However, it may be shown that, for large samples, the empirical Bayes estimates are very nearly equal to the fully Bayes estimates (see Kass and Steffey 1989). In Example 16.4 the extent to which the estimate of θ_i is based on data from the other strains $j \neq i$ depends on the relative magnitudes of σ_i and τ, according to (16.39). Because τ was actually estimated from[20] the data, the data themselves determined the relevance of the information from the strains $j \neq i$ to the estimate of θ_i.

Example 12.3 (continued from p. 459) When we introduced this example, on p. 331, we said that Behseta et al. were concerned with two indices: the rank index $I_{\text{rank}} = \frac{(f_3 - f_1)}{(f_3 + f_1)}$, where f_1 and f_3 were the mean firing rates measured at the times of the first and third saccades respectively (the mean being taken across trials), and the reward index was $I_{\text{reward}} = \frac{(f_b - f_s)}{(f_b + f_s)}$ where f_b and f_s were the mean firing rates during the post-cue delay period on big-reward and small-reward trials respectively. The indices I_{rank} and I_{reward} were positively correlated, but the effect was smaller than expected, with $r = 0.49$ and the authors suspected that the correlation had been attenuated due to noise arising from trial-to-trial variation in the spike counts. We

[18] The data analyzed here were based on a pre-publication draft of the Sklar and Strauss paper and are slightly different than those reported in the final version. Because strain 11 had such a large uncertainty the authors replicated their experiment for strain 11 with a much larger sample and obtained results that were much more consistent with the other strains.

[19] Alternatively, this likelihood may be integrated over μ and then maximized over τ, which produces a slightly different and sometimes preferable estimate often known as the REML estimate of τ, for *restricted maximum likelihood* estimate.

[20] In the jargon of computer science we would say that the hyperparameter τ was *learned* from the data, as opposed to fixed within the estimation algorithm.

added that Behseta et al. developed a method to correct for the attenuation and when they applied it to these data the new estimate of correlation was .83, which was more reasonable. We now provide some details about the method.

On p. 459 we let X_i be the random variable representing I_{rank} for the ith neuron and we said that Behseta et al. used the normal hierarchical model (16.33) and (16.34). Let us reformulate (16.33) by writing

$$\begin{aligned} X_i &= \theta_i + \epsilon_i \\ \epsilon_i &\sim N(0, \sigma^2_{\epsilon_i}) \end{aligned}$$

and then let Y_i represent the value of I_{reward} for the ith neuron and write

$$\begin{aligned} Y_i &= \xi_i + \delta_i \\ \delta_i &\sim N(0, \sigma^2_{\delta_i}). \end{aligned}$$

Here θ_i and ξ_i represent the theoretical values of I_{rank} and I_{reward} for neuron i that would be obtained from noiseless measurements (or from infinitely many trials). The quantities σ_{ϵ_i} and σ_{δ_i} are the standard errors associated with x_i and y_i (they were obtained by propagation of uncertainty from the spike count means). Taking ϵ_i and δ_i to be independent we may combine the assumptions on X_i and Y_i by saying these random variables are bivariate normal according to

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N(m_i, V_i), \tag{16.43}$$

where

$$m_i = \begin{pmatrix} \theta_i \\ \xi_i \end{pmatrix} \text{ and } V_i = \begin{pmatrix} \sigma^2_{\epsilon_i} & 0 \\ 0 & \sigma^2_{\delta_i} \end{pmatrix}.$$

Equation (16.43) is the first stage of a bivariate normal hierarchical model. Behseta et al. wrote the second stage in the form

$$\begin{pmatrix} \theta_i \\ \xi_i \end{pmatrix} \sim N(\mu, \Sigma), \tag{16.44}$$

where

$$\mu = \begin{pmatrix} \mu_\theta \\ \mu_\xi \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \sigma^2_\theta & \rho_{\theta\xi}\sigma_\theta\sigma_\xi \\ \rho_{\theta\xi}\sigma_\theta\sigma_\xi & \sigma^2_\xi \end{pmatrix}$$

with $\mu_\theta, \mu_\xi, \sigma^2_\theta$, and σ^2_ξ being the means and the variances of θ_i and ξ_i respectively. The quantity of interest is $\rho_{\theta\xi}$, which represents the correlation between the theoretical values θ_i and ξ_i. Let us refer back to the theorem on attenuation of the correlation on p. 330. In the notation used here, that theorem says that if $\rho_{\theta\xi} > 0$ then

$$\rho_{XY} < \rho_{\theta\xi}.$$

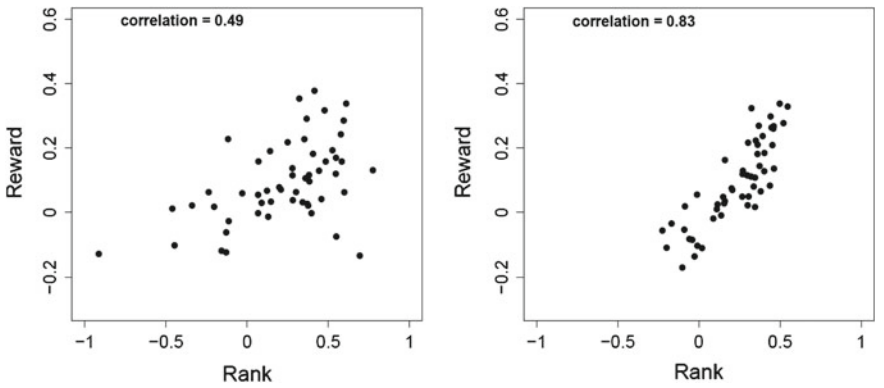

Fig. 16.2 Plots of reward-selective versus rank-selective indices, before and after Bayesian correction. *Left* uncorrected indices. The x-axis represents the index value for the the serial order saccade task. This is obtained through $I_{\text{rank}} = \frac{(f_3 - f_1)}{(f_3 + f_1)}$, where f_1 and f_3 were the mean firing rates measured at the times of the first and third saccades respectively. The y-axis indicates the index of selectivity for the size of the anticipated reward: $I_{\text{reward}} = \frac{(f_b - f_s)}{(f_b + f_s)}$ where f_b and f_s were the firing rates during the post-cue delay period on big-reward and small-reward trials respectively. *Right* plot of posterior means representing values after correction for noise. Adapted from Behseta et al. (2005).

In words, the correlation is attenuated (ρ_{XY} is smaller than $\rho_{\theta\xi}$) due to measurement noise. Because the model in (16.43) and (16.44) incorporates the uncertainty of the measurements (represented by σ_{ϵ_i} and σ_{δ_i}), a good estimate of $\rho_{\theta\xi}$ will adjust for measurement error and thereby correct the attenuated estimate (which, according to (16.43) and (16.44), mistakenly estimates ρ_{XY} rather than $\rho_{\theta\xi}$).

To get a good estimate of $\rho_{\theta\xi}$, ML estimation could be used, but Behseta et al. found it easiest to introduce prior distributions on μ and Σ and then apply MCMC, as outlined on p. 468. They obtained an estimate $\hat{\rho}_{\theta\xi} = .83$ with 95 % credible interval (.77, .88). The resulting shrinkage of the posterior means using (16.43) and (16.44), compared with the raw values of I_{rank} and I_{reward}, may be seen in Fig. 16.2. Behseta et al. provided simulations to show that this procedure produced estimates with small MSE and credible intervals with good coverage probability, especially compared with alternative methods that had appeared in the literature. □

16.2.2 For hierarchical models, posterior distributions are often computed by Gibbs sampling.

If random vectors X and Y have joint pdf $f(x, y)$, with $f(x, y) > 0$ for all x, y, then[21] $f_X(x) > 0$ for all x and $f_Y(y) > 0$ for all y. From Section 4.2.3, this implies

[21] In the discrete case $f_X(x) = \sum_y f(x, y)$ and the sum of positive quantities is positive. In the continuous case the integral of a positive function is positive.

there is a conditional distribution of X given Y with pdf $f_{X|Y}(x|y)$ and a conditional distribution of Y given X with pdf $f_{Y|X}(y|x)$. Furthermore (see Section 4.2.3), the two-step procedure

(a) draw a random variable $X = x$ from the marginal pdf $f_X(x)$, and then
(b) draw a random variable Y from the conditional pdf $f_{Y|X}(y|X = x)$

produces a draw (X, Y) from the distribution having joint pdf $f(x, y)$.

Now suppose $(X^{(g)}, Y^{(g)}) = (x^{(g)}, y^{(g)})$ is a draw from the joint distribution with pdf $f(x, y)$, for which $f(x, y) > 0$ for all x, y, and consider the following two-step process:

1. draw $X^{(g+1)} = x^{(g+1)}$ from the conditional distribution having pdf $f_{X|Y}(x|Y = y^{(g)})$;
2. draw $Y^{(g+1)} = y^{(g+1)}$ from the conditional distribution having pdf $f_{Y|X}(y|X = x^{(g+1)})$.

Because $Y^{(g)}$ has the marginal pdf $f_Y(y)$, step 1 corresponds to steps (a) and (b) above and produces a draw $(X^{(g+1)}, Y^{(g)}) = (x^{(g+1)}, y^{(g)})$ from the joint distribution having pdf $f(x, y)$. Therefore, $X^{(g+1)} = x^{(g+1)}$ is a draw from the distribution having marginal pdf $f_X(x)$ and then step 2 corresponds again to steps (a) and (b) above and produces a draw $(X^{(g+1)}, Y^{(g+1)}) = (x^{(g+1)}, y^{(g+1)})$ from the joint distribution having pdf $f(x, y)$. Thus, if we use steps 1 and 2 to define a Markov chain, then the pdf $f(x, y)$ is its stationary distribution. This version of MCMC is known as[22] *Gibbs sampling*.

What we have just described is called the *two-block* version of Gibbs sampling because the vector (X, Y) appears as two sets or "blocks," X and Y, of components. Specifically, the two-block Gibbs sampling algorithm initializes by setting a value of $(x^{(0)}, y^{(0)})$ then for $g = 0, \ldots, G - 1$ repeats steps 1 and 2 above.

To simulate from an m-dimensional posterior Gibbs sampling can involve cycling through k steps corresponding to k blocks of components, where $k \leq m$. In the case $k = m$, each component of the parameter vector $\theta = (\theta_1, \ldots, \theta_m)$ is updated using the conditional distribution conditionally on all other components. That is, letting

$$\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_m),$$

the ith component is updated using the conditional distribution having pdf $f_{\theta_i|\theta_{-i}}(\theta_i|\theta_{-i})$. More specifically, the first step of the m-step Gibbs sampling algorithm is

Step 1: draw $\theta^{(g+1)}$ from the conditional distribution having pdf $f_{\theta_1|\theta_{-1}}(\theta_1|\theta_{-1}^{(g)})$, and for $i = 2, \ldots, m$, the ith step in the m-step Gibbs sampling algorithm becomes

[22] Theoretical analysis of Gibbs sampling shows that the conditions mentioned in footnote 9 are satisfied (see Robert and Casella 2004). The name comes from Geman and Geman (1984), who applied it to image restoration, where there is a close analogy with the Gibbs distribution in statistical mechanics.

Step i: draw $\theta^{(g+1)}$ from the conditional distribution having pdf $f_{\theta_i|\theta_{-i}}(\theta_i|\theta_1^{(g+1)}, \ldots, \theta_{i-1}^{(g+1)}, \theta_{i+1}^{(g)}, \ldots, \theta_m^{(g)})$.

The distributions with pdfs $f_{\theta_i|\theta_{-i}}(\theta_i|\theta_{-i})$ are called *full conditional* distributions. Gibbs sampling is particularly convenient when the full conditional distributions have a standard form so that random draws may be obtained from existing software.

Example 12.3 (continued from p. 464) The unknown parameters in (16.43) and (16.44) are the vector $\mu = (\mu_\theta, \mu_\xi)$, the matrix Σ, which includes components $(\sigma_\theta^2, \sigma_\xi^2, \rho_{\theta\xi})$, and the vectors $m_i = (\theta_i, \xi_i)$, for $i = 1, \ldots, 54$. Behseta et al. put independent diffuse normal priors on μ_θ and μ_ξ, meaning normal priors with large variances, which approximate uniform priors. They used a particular conjugate prior,[23] following Kass and Natarajan (2006), for Σ. Gibbs sampling proceeds by, first, getting initial values of μ and Σ (e.g., using[24] method-of-moments estimators, see Section 7.2.1), and then iterating the steps

1. draw m_i, for $i = 1, \ldots, 54$, conditionally on the current values of μ and Σ;
2. draw μ conditionally on the current values of Σ and m_i, for $i = 1, \ldots, 54$;
3. draw Σ conditionally on the current values of μ and m_i, for $i = 1, \ldots, 54$.

Each of these steps is straightforward because of the conjugate structure of the model. In step 1, because μ and Σ are fixed (by conditioning) the second stage in Eq. (16.44) becomes analogous to the univariate $N(\mu_\pi, \sigma_\pi^2)$ prior leading to (16.8) and (16.9). The posterior on each m_i is thus the bivariate normal with mean and variance given by the bivariate extension of (16.8) and (16.9). In step 2, because Σ and all of the m_i are fixed, we use Eq. (16.44) together with the conjugate normal prior to get a bivariate normal posterior on μ. Step 3 is similar to step 2, except now μ is fixed and the conjugate prior on Σ is used to get a conjugate posterior. Because all three steps involve standard distributions, they may be carried out with existing software. Behseta et al. used the MCMC software package BUGS, which is freely available (see Lunn et al. 2012). □

The key to making Gibbs sampling easy in Example 12.3 is that the full conditional distributions become tractable when we consider not only the parameter vectors μ, Σ but also all of the parameter vectors (θ_i, ξ_i), for $i = 1, \ldots, 54$. One way to look at this is to consider the (θ_i, ξ_i) parameters to "augment" the data observations (x_i, y_i), for $i = 1, \ldots, 54$, in the sense of Section 8.4.5. In Section 8.4.5 we presented an illustration (a mixture of two Gaussians model) in which the data were augmented with latent variables, and then the EM algorithm could be implemented easily. In

[23] For a univariate normal variance parameter σ^2 the conjugate prior family is called *inverse-gamma* because σ^{-2} follows a gamma distribution. The multivariate extension is called *inverse-Wishart*. For a $p \times p$ variance matrix the inverse-Wishart itself has $1 + p(p+1)/2$ free parameters that must be selected. Kass and Natarajan (2006) suggested a method of doing so in the context of hierarchical models.

[24] A rough estimate of Σ may be obtained by setting $V_1 = \cdots = V_k = V^*$ in (16.43), where V^* is some kind of average of the V_i matrices (such as the inverse of the mean of the inverse matrices) and then applying the method of moments.

general, both Gibbs sampling and EM algorithms are easy to implement for problems in which data augmentation produces tractable conditional distributions.

16.2.3 Penalized regression may be viewed as Bayesian estimation.

The multiple regression model (12.44) is

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i \tag{12.44}$$

for $i = 1, \ldots, n$ where the ϵ_i random variables are assumed i.i.d. $N(0, \sigma^2)$. We can form a hierarchical model by assuming a second-stage distribution

$$\beta_j \sim N(0, \tau^2) \tag{16.45}$$

for $j = 1, \ldots, p$, independently. Calculations show that the posterior pdf on $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ conditionally on the data $y = (y_1, \ldots, y_n)$ and τ is

$$f(\beta|y, \lambda) \propto \exp\left(-\frac{1}{2\sigma^2}[\sum_{i=1}^{n}(y_i - y_i(\beta))^2 + \frac{\sigma^2}{\tau^2}\sum_{j=1}^{p}\beta_j^2]\right) \tag{16.46}$$

where

$$y_i(\beta) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$$

and

$$\lambda = \sigma^2/\tau^2.$$

Noticing that $y_i(\beta)$ in (16.46) is the same as y_i^* in Eq. (12.45), we see that maximizing the posterior in (16.46) is equivalent to finding the penalized least-squares solution in (12.71) with the L2 penalty (12.72).

The interpretation is this: the second-stage distribution (16.45) models the variation among the β_j coefficients as if they are normally distributed around 0; if this is more-or-less accurate, then the coefficients are likely to be closer to 0 than the least-squares solution would suggest (because the least-squares values are noisy) and, because the posterior shrinks the coefficients toward zero, the L2-penalized solution should provide a good estimate of β.

An alternative is to introduce a different second-stage distribution. It turns out that L1-penalized regression corresponds to using a hierarchical model of the form (16.45) but instead using the *Laplace distribution*, which has much thicker tails than the normal. In addition to suggesting a multitude of different procedures that correspond to different second-state distributions, the Bayesian formulation also opens the door to alternative computational methods. See Kyung et al. (2010) for additional discussion and references.

16.2.4 State-space models allow parameters to evolve dynamically.

The essential ideas in our discussion of Example 16.3, on p. 458, were that (i) the latent variables X_t represented the bursting or non-bursting status of the neuron at time t, and (ii) the value of this random variable (1 if bursting, 0 if non-bursting) could be estimated from the data. The bursting or non-bursting "status" is often called a *state*, and because it is represented by a latent variable it is often called *hidden*.

Example 16.3 (continued from p. 458) If we assume that the binary random variables X_t form a Markov chain (see p. 453) then we have a *hidden Markov model*. According to this model, the neuron[25] evolves from bursting to non-bursting states, and from non-bursting to bursting states, stochastically with fixed probabilities, as in Eq. (16.25). That is, the state variables X_t evolve as a Markov chain, and the probabilities for the observation variables Y_t are determined from the state variables. □

The structure illustrated above, in Example 16.3 has two stages: the observation random variables Y_t have distributions that depend on the state variables X_t, while the state variables themselves evolve stochastically in a relatively simple way. In Example 16.3 the state variables were binary and satisfied the Markov condition (16.24). We now allow both the observation variables and the state variables to be general. We assume the vector Y_t is p-dimensional and the vector X_t is m-dimensional. The Markov condition becomes

$$f(x_t|x_1, x_2, \ldots, x_{t-1}) = f(x_t|x_{t-1}) \tag{16.47}$$

and the pdf of Y_t may be written $f(y_t|x_t)$, which indicates its dependence on the value of the state variable X_t. This is the general form of a *state-space model*.

Because at time t we are considering a collection of variables

$$\{X_1, X_2, \ldots, X_t, Y_1, Y_2, \ldots, Y_t\}$$

we modify (16.47) to the more inclusive equation

$$f(x_t|x_1, x_2, \ldots, x_{t-1}, y_1, y_2, \ldots, y_{t-1}) = f(x_t|x_{t-1}) \tag{16.48}$$

and we similarly write

$$f(y_t|x_1, x_2, \ldots, x_t, y_1, y_2, \ldots, y_{t-1}) = f(y_t|x_t). \tag{16.49}$$

We also assume that at time $t = 1$ we have an initial state distribution with

$$X_1 \text{ has pdf } f(x_1). \tag{16.50}$$

[25] In fact, according to this assumption on the Y_t variables, the neuron's spike trains flip back and forth between two discrete-time versions of Poisson processes, a bursting Poisson process and a non-bursting Poisson process; see Chapter 19.

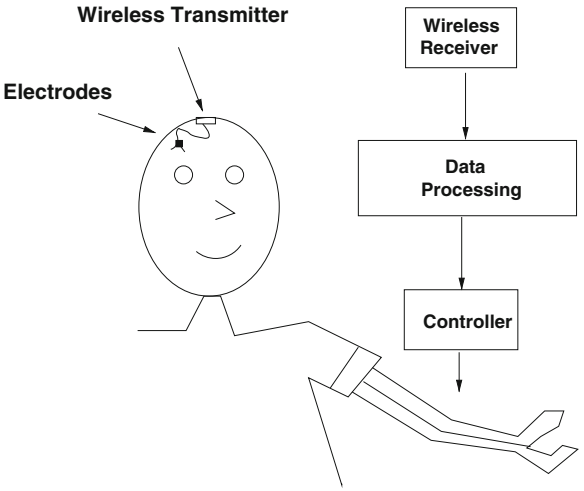

Fig. 16.3 Stick figure representation of brain-computer interface application.

Equations (16.48), (16.49), and (16.50) are standard assumptions for state-space models and for the remainder of our presentation we assume they hold.

Example 16.5 Decoding Hand Movement from Cortical Activity In Example 12.6 we described a statistical model for neural activity in primary motor cortex (M1) in terms of two-dimensional directional hand movement. Since the time of the original experiments by Georgopoulos, Schwartz, and colleagues, it has been recognized that simultaneous recordings from multiple M1 neurons could produce predictions of hand movement, and that this could furnish the basis for a brain-computer interface, which could assist severely disabled patients. See[26] Fig. 16.3. This is usually called *decoding* of hand movement, from the neural activity; see also Example 4.7 on p. 100. One way to perform the decoding is to introduce observation variables Y_t to represent neural activity, and state variables X_t to represent movement parameters, such as direction or velocity. More specifically, if we have N neurons and we let Y_t^i be the spike count for neuron i during an interval centered at time t, which we might for simplicity assume to be normally distributed, then we could take Y_t to be the vector $Y_t = (Y_t^1, Y_t^2, \ldots, Y_t^N)$ and we could let X_t be the hand movement direction[27] at time t and define a suitable distribution for the variables X_t, subject to the Markovian constraint (16.48). Estimation of the state variables X_t then produces the desired decoding prediction of hand movement. □

For a sequence of numbers or vectors $a_1, a_2, \ldots, a_t$ let us use the notation

[26] For a review of these ideas together with commentary on algorithms see Brockwell et al. (2007).

[27] Often movement velocity is used, and sometimes direction, velocity, and acceleration are all used as components of X_t.

$$a_{1:t} = (a_1, \ldots, a_t).$$

We give the basic result on sequential Bayesian estimation of the state variables $X_t = x_t$ via the posterior pdf $f(x_t|y_{1:t})$ based on the data $Y_{1:t} = y_{1:t}$.

Filtering and Prediction Equations Under the conditions (16.48), (16.49), and (16.50) the posterior pdf of X_t given $Y_{1:t} = y_{1:t}$ is given by the *filtering equation*,

$$f(x_t|y_{1:t}) \propto f(y_t|x_t) f(x_t|y_{1:t-1}) \tag{16.51}$$

where

$$f(x_t|y_{1:t-1}) = \int f(x_t|x_{t-1}) f(x_{t-1}|y_{1:t-1}) dx_{t-1}, \tag{16.52}$$

which is the *prediction equation*. The estimate of x_t is then the posterior mean

$$\hat{x}_t = E(X_t|Y_{1:t} = y_{1:t}). \tag{16.53}$$

The filtering and prediction equations provide a recursive prescription for finding the posterior pdf. Specifically, we initialize with the prior distribution for X_1 having pdf $f(x_1)$ (as in (16.50)) which we substitute for $f(x_t|y_{1:t-1})$ in (16.51) to get $f(x_1|y_1)$, then use this in (16.52) to get $f(x_2|y_1)$, then put $f(x_2|y_1)$ in (16.51) to get $f(x_2|y_{1:2})$, etc., repeatedly alternating between the filtering and prediction equations. Based on the successive posterior pdfs $f(x_1|y_1), f(x_2|y_2), \ldots$ we compute posterior means (as in (16.53)) $\hat{x}_1, \hat{x}_2, \ldots$, which become the sequential estimates of $x_1, x_2, \ldots$.

We provide the derivation in Section 16.4.

In (16.51), $f(x_t|y_{1:t-1})$ plays the role of the prior: it represents what is known about x_t prior to observing y_t. The factor $f(x_t|x_{t-1})$ in (16.52) is based on the modeling assumptions, which must conform to (16.48). This framework is very general. Here is an example in which (16.49) involved a detailed spike train model.

Example 16.6 Plasticity of hippocampal place fields Neural receptive fields are frequently plastic: a neural response to a stimulus can change over time as a result of experience. Frank et al. (2002) used a spike train model (along lines discussed in Section 19.3.4) to characterize spatial receptive fields of neurons from both the CA1 region of the hippocampus and the deep layers of the entorhinal cortex (EC) in awake, behaving rats. In this context, the spatial receptive fields are usually called *place fields*, because they indicate where the animal is currently located or moving. By then formulating a state-space model where features of the neural place fields were treated as state variables, the authors could describe the evolution of the place fields during the experiment. They found consistent but distinct patterns of plasticity in CA1 hippocampal neurons and deep entorhinal cortex (EC) neurons. We return to this example in Section 19.3.4. □

When the observation model (16.49) and the state model (16.48) both involve linear equations and normal (Gaussian) errors, the filtering and prediction equations simplify greatly, and the sequential estimation steps can be written analytically. We discuss this important special case in Section 16.2.5.

16.2.5 The Kalman filter may be used to estimate state variables for linear, Gaussian state-space models.

The state-space conception is stunning in its generality and simplicity of logical argument, but for its power to be realized computational methods are crucial. A special case, in which the recursions in (16.51) and (16.52) yield easily-computed algebraic expressions, has the form

$$X_t = AX_{t-1} + \epsilon_t \tag{16.54}$$

$$Y_t = BX_t + \eta_t \tag{16.55}$$

where A is an $m \times m$ matrix, B is a $p \times p$ matrix, and ϵ_t and η_t are multivariate normal for all t, all independently of each other, and we assume

$$V(\epsilon_t) = Q$$
$$V(\eta_t) = R$$

for all t. Equations (16.54) and (16.55) define what is usually called a *linear, Gaussian state-space model*. Sometimes the matrices A and B and/or the covariance matrices Q and R are allowed to vary with time, but we ignore that possibility here. The algorithm that implements the sequential estimates (16.53) given by the filtering and prediction Eqs. (16.51) and (16.52) under the linear, Gaussian assumptions (16.54) and (16.55) is called the *Kalman filter* (Kalman 1960). To completely specify the distributions given by the filtering and prediction equations we must also have an initial distribution for X_1, as in (16.50), which we assume to be m-dimensional normal

$$X_1 \sim N_m(\hat{x}_{0|0}, \hat{W}_{0|0}) \tag{16.56}$$

for some initializing mean vector $\hat{x}_{0|0}$ and variance matrix $\hat{W}_{0|0}$. We have introduced the subscripts to conform to those we use below. In particular, in the following, our notation for the posterior mean (from the pdf based on the filtering equation) replaces $\hat{x}_t$ in (16.53) with $\hat{x}_{t|t}$, which is supposed to indicate that we are estimating the state x_t at time t using all the data available at time t. We also use $\hat{W}_{t|t}$ to represent the corresponding posterior variance. The posterior mean from the pdf based on the prediction equation will be written $\hat{x}_{t|t-1}$.

The Kalman Filter
Under the model (16.54) and (16.55) with (16.56) the filtering and prediction equations yield posterior means (16.53) given by

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + \hat{W}_{t|t-1}B^T(B\hat{W}_{t|t-1}B^T + R)^{-1}(y_t - B\hat{x}_{t|t-1}) \tag{16.57}$$

where

$$\hat{x}_{t|t-1} = A\hat{x}_{t-1|t-1} \tag{16.58}$$
$$\hat{W}_{t|t-1} = A\hat{W}_{t-1|t-1}A^T + Q. \tag{16.59}$$
$$\hat{W}_{t|t} = \hat{W}_{t|t-1} + \hat{W}_{t|t-1}B^T(B\hat{W}_{t|t-1}B^T + R)^{-1}B\hat{W}_{t|t-1}. \tag{16.60}$$

Example 16.5 (continued from p. 471) Hand movement follows continuous paths which, at least for ordinary ballistic movements to a target, do not suddenly jump from one direction to another. A simple way to model such movement is to take

$$X_t = X_{t-1} + \delta_t \tag{16.61}$$

where the δ_t are assumed to be independently $N(0, \Sigma_\delta)$. When the magnitude of Σ_δ is small (e.g., if Σ_δ is diagonal and the diagonal elements, representing variances, are small) this constrains the path from X_1 to X_2 to $\cdots X_t$ to vary smoothly. Model (16.61) is called a normal (or Gaussian) *random walk* (confer p. 126 and 530), because each increment $X_t - X_{t-1}$ is analogous to taking a step of random length while walking. Random walk models in conjunction with normal observation assumptions produce analytically tractable filtering and prediction pdfs. They are often used in brain-computer interface applications. For instance, Wu et al. (2006) considered spike count data recorded from 42 neurons in primary motor cortex while a monkey performed a hand-movement task. These authors used a Kalman filter with (16.61) to reconstruct hand-movement position, velocity, and acceleration. Figure 16.4 shows some typical hand position data, together with the Kalman filter estimates, indicating that the Kalman filter does a very good job in reconstructing hand position. Such reconstructions are called "off-line" or "open loop" because they indicate the potential of the methodology, as opposed to "on-line" or "closed-loop" applications in which the decoding method is used by the subject to control the movement of a cursor or robotic device. See Koyama et al. (2010) for comparison of several decoding methods, and a discussion of performance differences in off-line and on-line control. □

We have not yet said how the parameters A, B, Q, R in the model (16.54) and (16.55) are estimated. Let us assume some preliminary *training* data are available to use for this purpose. If the state vectors X_t were known, then all of these parameters could be estimated by regression in (16.55) and auto-regression in (16.54). This suggests the following iterative algorithm.

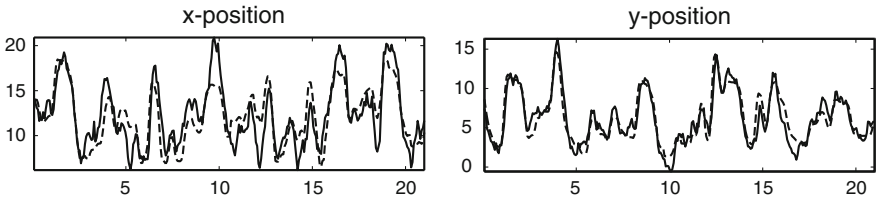

Fig. 16.4 Decoding performance of Kalman filter in a two-dimensional hand movement task. The *dashed traces* show x and y coordinates of hand position on a single trial, and the *solid traces* are the reconstructed movements using the Kalman filter. Adapted from Wu et al. (2006).

- Initialize A, B, Q, R.
- Iterate to convergence:

1. Run the Kalman filter[28] to get estimates of the state vectors $X_1, X_2, \ldots, X_T$ (where $t = T$ is the last value of t).
2. Using the estimates of $X_1, X_2, \ldots, X_T$ use regression and auto-regression to estimate A, B, Q, R.

This is an instance of the *EM algorithm*. As discussed in Section 8.4.5, the EM algorithm iteratively maximizes a likelihood by combining, at each iteration, an expectation step (here the estimation of the state vectors) with a maximization step (here maximizing the likelihood of A, B, Q, R given the state vectors).

16.3 Bayes Factors

As we said in Section 10.4.5, a common mistake among näive users of statistics is to commit the p-value fallacy, which is to interpret the p-value as $P(H_0|data)$. It is not surprising that this mistake occurs frequently: on the one hand, the logic of p-values is indirect, and correct statements interpreting them may seem convoluted while, on the other hand, $P(H_0|data)$ is a simple and intuitive summary of knowledge about the null hypothesis based on the data; it is easy to jump to the conclusion that the p-value must be delivering $P(H_0|data)$. Instead of erroneously interpreting the p-value as if it were $P(H_0|data)$, one can apply Bayes' theorem to calculate $P(H_0|data)$ according to Eq. (10.36), which we repeat here:

$$P(H_0|data) = \frac{P(data|H_0)P(H_0)}{P(data|H_0)P(H_0) + P(data|H_A)P(H_A)}, \tag{10.36}$$

[28] In practice one also runs a version of the Kalman filter backwards in time, beginning at the last time point $t = T$ and ending with $t = 1$. This pair of forward and backward algorithms is called the *Kalman smoother*.

where $P(H_A) = 1 - P(H_0)$. When this formula is applied to statistical models under H_0 and H_A, as in Section 11.1.6, $P(data|H_0)$ becomes a discrete or continuous pdf for a random vector X, so we substitute

$$f_0(x) = P(data|H_0) \text{ and } f_1(x) = P(data|H_A),$$

where we are using the subscript 1 to signify the alternative, and we write

$$P(H_0|x) = \frac{f_0(x)P(H_0)}{f_0(x)P(H_0) + f_1(x)P(H_A)}. \tag{16.62}$$

As we pointed out in Section 11.1.6, p. 297, the prior probability $P(H_0)$ may be removed by considering instead the *Bayes factor*, which is the ratio of posterior odds to prior odds,

$$BF_{01} = \frac{P(H_0|x)}{P(H_A|x)} \div \frac{P(H_0)}{P(H_A)} \tag{16.63}$$

and from

$$\frac{P(H_0|x)}{P(H_A|x)} = \frac{f_0(x)P(H_0)}{f_1(x)P(H_A)}$$

we have

$$BF_{01} = \frac{f_0(x)}{f_1(x)}. \tag{16.64}$$

The subscript on BF_{01} indicates that we are considering the Bayes factor in favor of H_0. Its reciprocal, BF_{10}, would be the Bayes factor in favor of H_A. In Section 16.3.1 we describe the way the Bayes factor quantifies evidence in favor of a hypothesis, in Section 16.3.2 we review briefly the contribution of Bayes factors and posterior probabilities to epistemology, in Section 16.3.3 we issue a note of caution concerning the strong dependence of Bayes factors on prior distributions, and in Section 16.3.4 we discuss their use in calibrating p-values.

Bayes factors were first discussed by Harold Jeffreys, who saw them as a way of evaluating the strength of evidence in favor of a new scientific theory (represented by H_A) that might replace an old one (H_0). A modern view was provided by Kass and Raftery (1995). Jeffreys (1961, Appendix B) suggested interpreting BF_{10} (the evidence in favor of the new theory) in half units on the $\log_{10}$ scale. Although probability itself provides a meaningful scale, as do the odds, Jeffreys felt it was useful to provide a rough statement about standards of evidence in scientific practice. Table 16.2 is a mildly modified version of his interpretive categories (taken from Kass and Raftery 1995). Interpretation may depend on context, but these categories remain useful. They are stated in terms of BF_{10} because weighing evidence *against* a null hypothesis is more familiar, but Bayes factors can equally well provide evidence *in favor of* a null hypothesis. Indeed, this is one of the strengths of the Bayesian approach. We illustrate by returning to Example 16.4 in Section 16.3.1.

Table 16.2 The interpretation of Bayes factors, in the form $BF_{10} = (BF_{01})^{-1}$, based on Jeffreys's recommendations.

$\log_{10}(BF_{10})$	BF_{10}	Evidence against H_0
0–$\frac{1}{2}$	1–3.2	Not worth more than a bare mention
$\frac{1}{2}$–1	3.2–10	Substantial
1–2	10–100	Strong
> 2	> 100	Decisive

16.3.1 Bayes factors can provide evidence in favor of hypotheses.

Equation (10.36) hides the important complication that the pdfs appearing in (16.64) typically contain unknown parameters. Let ω be the parameter vector corresponding to the pdf f_1 under H_0 and ξ the parameter vector corresponding to the pdf f_2 under H_A. As in Sections 11.1.3 and 11.1.6, these parameter vectors could have differing dimensionalities. For example, if $X \sim N(\mu, \sigma^2)$ and $H_0 : \mu = 0$, while under H_A μ is unrestricted, then $\omega = \sigma$ and $\xi = (\mu, \sigma)$. If the values of ω and ξ are unknown, they must be estimated. Within the Bayesian framework they then must have prior pdfs $\pi_0(\omega)$ and $\pi_1(\xi)$ and we then get the expressions

$$f_0(x) = \int f_0(x|\omega)\pi_0(\omega)d\omega$$
$$f_1(x) = \int f_1(x|\xi)\pi_1(\xi)d\xi$$

so that (16.64) becomes

$$BF_{01} = \frac{\int f_0(x|\omega)\pi_0(\omega)d\omega}{\int f_1(x|\xi)\pi_1(\xi)d\xi}. \tag{16.65}$$

Writing the normal pdf with mean m and variance v evaluated at x as $n(x; m, v)$, we have

$$BF_{01} = \frac{\int n(x; 0, \sigma^2)\pi_0(\sigma)d\sigma}{\int n(x; \mu, \sigma^2)\pi_1(\mu, \sigma)d\mu d\sigma}. \tag{16.66}$$

Equation (16.66) is an instance of (16.65) with $\omega = \mu$ and $\xi = (\mu, \sigma)$. If σ is known, formula (16.66) simplifies. Let us substitute $\theta = \mu$. We then have

$$BF_{01} = \frac{n(x; 0, \sigma^2)}{\int n(x; \theta, \sigma^2)\pi_1(\theta)d\theta} \tag{16.67}$$

which is in a form we can apply to Example 16.4.

Example 16.4 (continued) When we introduced this example on p. 462 we said that the investigators were interested in the possibility that, for a particular strain of *E. coli*, mutations (producing acetate utilization deficiency) might arise from error-prone DNA repair; if so, that strain would satisfy $H_0 : p_{i1} = p_{i2}$ or, from (16.40),

$H_0 : \theta_i = 0$. The results in Table 16.1 suggested that strain 12 might satisfy this hypothesis, i.e.,

$$H_0 : \theta_{12} = 0. \tag{16.68}$$

We now consider the evidence in favor of H_0 defined by (16.68), presenting results reported in Kass and Raftery (1995).

Under H_0 the data random variable X_{12} follows a normal distribution with mean 0 and known variance σ_{12}^2, and so the numerator of BF_{01} has the form of the numerator in (16.67). Under H_A we may assume $X_{12} \sim N(\theta_{12}, \sigma_{12}^2)$ and we now must choose $\pi_1(\theta_{12})$, which appears in the denominator of (16.67). Because, under H_A, strain 12 would be judged similar to all the other strains, we may use the second-stage normal distribution that appeared in the hierarchical model considered previously (p. 462), i.e., the pdf takes the form

$$\pi_1(\theta_{12}) = n(\theta_{12}; \mu, \tau^2) \tag{16.69}$$

where μ and τ are found from the data involving strains $j \neq 12$ (using ML estimation, as discussed in Section 16.2.1). Kass and Raftery reported that when this was done, the Bayes factor was

$$B_{01} = 15$$

indicating that these data produced[29] strong evidence in favor of H_0. □

Example 16.2 (continued) On p. 451 we described the way Solt et al. (2011) used the posterior probability $P(p_1 > p_2 | X_1 = 12, X_2 = 0)$ to judge their result that 12 out of 12 rats regained their upright position following a substantial dose of methylphenidate whereas 0 out of 6 did following a negligible dose. We may instead use Bayes factors.

We begin by considering the hypotheses to be tested. The data 0 out of 6 confirmed that the very small dose of methylphenidate left the rats unable to regain their upright position. If p is the probability of regaining upright position we might want to take $H_0 : p = 0$ and $H_A : p \neq 0$. Under H_0 the outcome 0 out of 6 has probability 1. Under H_A we may introduce a uniform prior on [0, 1] for the unknown value of p. Using $\binom{6}{0} = 1$, the Bayes factor in (16.65) becomes

$$BF_{01} = \frac{1}{\int_0^1 p^0 (1-p)^6 dp}$$

and, from Eq. (5.16), the denominator integral is equal to $6!/7! = 1/7$ and we get

[29] A possible issue is the extent to which strain 12 was selected *post hoc*, after the data had been examined. It is possible to correct the Bayes factor for such *post hoc* selection, analogously to (though differently than) the way p-values may be adjusted (see Section 11.3). The investigators repeated the experiment on strain 12 and found similar results, which provided strong confirmation of H_0.

$$BF_{01} = 7.$$

From this calculation, 0 out of 6 corresponds to substantial, but not strong evidence that rats will not regain an upright position.

Similarly, the data 12 out of 12 confirmed that at the high dose of methylphenidate, rats did regain their upright position. Here, we might take $H_0 : p = 1$ and $H_A : p \neq 1$. A similar calculation then gives

$$BF_{01} = \frac{1}{\int_0^1 p^{12}(1-p)^0 dp} = 13.$$

This would be considered strong evidence that rats regain their upright position for the high dose.

One might argue, however, that these two Bayes factors do not fully summarize the weight of evidence because they do not compare the 0 out of 6 and 12 out of 12 results to each other. The null hypothesis for the data from the higher dose of the drug could be formulated differently. If methylphenidate had no effect one might take $H_0 : p = 0$. However, under this null it is impossible for even one rat to regain its upright position. Therefore, if only one rat out of the whole sample were to regain its upright position there would be infinite odds against H_0. But, actually, if only one rat had regained upright position the results would not have been very convincing: experiments are not perfectly precise, and for reasons beyond the experimenter's control it might have happened that, even with a null effect, an occasional animal might, even if rarely, regain its upright position. Indeed, one could argue that the reason the authors demonstrated the null effect with six rats was to reassure readers that regaining an upright position, at a low dose, is rare. A more useful null hypothesis would be to take $H_0 : p \sim B(1, 7)$. That is, we use the 0 out of 6 data to form a null hypothesis for the high-dose data. With the alternative $H_A : p \sim B(1, 1)$, i.e., a uniform prior on p, we obtain (using $\binom{12}{12} = 1$)

$$BF_{01} = \frac{\int_0^1 p^{12}(1-p)^0 \frac{\Gamma(8)}{\Gamma(1)\Gamma(7)} p^0 (1-p)^6 dp}{\int_0^1 p^{12}(1-p)^0 dp}.$$

The numerator reduces to

$$7 \int_0^1 p^{12}(1-p)^6 dp = 7 \frac{12!6!}{19!} = 1.98 \times 10^{-5}$$

while the denominator is

$$\int_0^1 p^{12}(1-p)^0 dp = \frac{1}{13}.$$

We then get

$$BF_{01} = .00026$$

which conforms with the intuition that the evidence is overwhelmingly in favor of an effect of methylphenidate in enabling rats to regain an upright position. □

16.3.2 *Bayes factors provide an interpretation of scientific progress.*

At the end of Section 16.1.5 we said that the approximate $N(\hat{\theta}, I_{OBS}(\hat{\theta})^{-1})$ distribution of the posterior provides an expression of one of the guiding principles of science, namely that investigators with different knowledge or opinions will eventually come to agreement after taking into account sufficiently much data. This concerns the value of a parameter θ. An analogous statement can be made concerning the scientific law that describes a particular phenomenon. Following Eq. (11.12) we noted that BIC is a consistent model selection procedure in the sense that, for sufficiently large samples, the probability of BIC choosing the correct model will get arbitrarily close to one. By virtue of (11.12) the same is true of Bayes factors.[30] To re-phrase this fundamental result in terms of posterior probability, suppose we have a set of m candidate models M_k, with m being as large as we like, and suppose further that we place positive prior probabilities $P(M_k)$ on them. For sufficiently much data, the posterior probability on the correct model will get arbitrarily close to one. This means that investigators having different opinions about the merits of competing scientific laws (represented as statistical models) will eventually come to agreement after taking into account sufficiently much data.

The result was recognized by Jeffreys and Wrinch (1921), and was a primary motivation for Jeffreys' monumental treatise *Theory of Probability*. In the preface to the first edition of his book (in 1939) he wrote:

> In opposition to the statistical school, [physicists] and some other scientists are liable to say that a hypothesis is definitely proved by observation, which is certainly a logical fallacy; most statisticians appear to regard observations as a basis for possibly rejecting hypotheses, but in no case for supporting them. The latter attitude, if adopted consistently, would reduce all inductive inference to guesswork; the former, if adopted consistently, would make it impossible ever to alter the hypotheses, however badly they agreed with new evidence.... In the present book I ... maintain that the ordinary common-sense notion of probability is capable of precise and consistent treatment when once an adequate language is provided for it. It leads to the results that a precisely stated hypothesis may attain either a high or a negligible probability as a result of observational data.

[30] Mathematically the situation is reversed: an elegant theorem due to Doob establishes the consistency of the posterior distribution, and thus of Bayes factors, under weak conditions. Equation (11.12) then provides consistency of BIC. For precise statements see Schervish (1995, Section 7.2.1) and the references in Kass and Raftery (1995, Section 4.1.3).

From a philosophical perspective, Jeffreys' use of Bayes factors and posterior probability of hypotheses represented a huge advance in understanding the nature of scientific inductive reasoning.

16.3.3 Bayes factors can be difficult to use when there is little information about unknown parameters.

The appearance of unknown parameters in (16.65) introduces a non-trivial complication. In Example 16.4, on p. 478, the prior π_0 in (16.68) was completely specified by H_0 and the prior π_1 was found from related data. When there are no available relevant data, it can be difficult to know how to select such priors. As we reviewed in Section 16.1.5, in problems involving estimation of a parameter θ, for large samples the influence of the prior diminishes to the point that answers based on two distinct priors will be essentially the same. One way to say this is that within the peak of the posterior distribution, Eq. (16.22) holds, and so regardless of the prior chosen we get (16.17). However, when large-sample analysis is applied to the numerator and denominator of (16.65), and (16.22) is applied, a factor $\pi_0(\hat{\omega})$ remains in the numerator and $\pi_1(\hat{\xi})$ remains in the denominator. In other words, even for large samples the value of the Bayes factor depends on the choice of priors. In practice, this dependence of the Bayes factor on prior pdfs limits its applicability. Bayes factors are much more compelling when, as in Example 16.4, the prior pdfs are themselves determined by data.

Sometimes a wide range of plausible priors may be defined and sensitivity of conclusions within this range may be evaluated. See Application 5 in Kass and Raftery (1995). Another possibility is to use the BIC as if it were the log of the Bayes factor, according Eq. (11.12). In fact, Kass and Wasserman (1995) showed that BIC corresponds to the use of a particular prior they called the *unit information prior* because it injects the same amount of information as a single data value x_i within a sample $x_1, \ldots, x_n$. Sometimes the unit information prior is used in order to obtain a rough quantification of evidence based on the Bayes factor.

Example 16.7 Set Shifting in ADHD Wagenmakers et al. (2010) reanalyzed data from Geurts et al. (2004) who, among other things, compared the performance of children with ADHD with controls on the Wisconsin Card Sorting Test (WCST). The WCST requires subjects to learn to sort cards according to implicit rules that change during the course of the experiment; performance is thus thought to quantify cognitive flexibility or *set shifting* ability. The data came from 52 children with ADHD and 26 control children. Wagenmakers et al. used hierarchical models to describe the variation in ability across subjects within each of the two experimental groups. They introduced a unit information prior on the mean difference in abilities between the groups and then computed $BF_{01} = 4.0$. The authors concluded there was modest evidence in favor of the null hypothesis that the mean ability on the WCST was the same for ADHD and control subjects. □

16.3.4 *Bayes factors can be used to calibrate p-values.*

On p. 282 and 476 we distinguished between the p-value and the quantity $P(H_0|data)$, which is computed from Bayes' theorem. In order to compute $P(H_0|data)$ based on the data $X = x$, according to (16.62), we need $f_0(x)$ and $f_1(x)$. The pdf $f_0(x)$ is used in calculating[31] the p-value. If $f_1(x)$ is either known or assumed known, as in (16.69), the Bayes factor may be computed and if we further take $P(H_0) = P(H_A) = \frac{1}{2}$ then Eq. (16.62) gives

$$P(H_0|x) = \frac{BF_{01}}{1 + BF_{01}}. \tag{16.70}$$

By making assumptions about $f_1(x)$ it therefore becomes possible to compare the p-value with $P(H_0|x)$. This was done by Jeffreys (Jeffreys 1961, pp. 373–374), and subsequently by Edwards et al. (1963) and others. See Sellke et al. (2001) for a thorough discussion. The approach taken by Edwards et al. (1963) and by Sellke et al. (2001) was to assume that the pdf f_1 lies in some family of distributions, and for data x such that a given p-value occurred (such as $p = .05$) they then minimized BF_{01} over all possible members of that family. This minimum represents the strongest possible evidence against H_0 that the p-value could provide, under the given assumptions. Under assumptions considered reasonable[32] by Sellke, Bayarri, and Berger, the value $p = .05$ corresponds to a minimum of $BF_{01} = .41$. In other words, under those assumptions, using (16.70), the value $p = .05$ corresponds to $P(H_0|data) \geq .41/1.41 = .29$. Calculations of this sort lead to the general conclusion that $p = .05$ is relatively weak evidence against H_0.

16.4 Derivations of Results on Latent Variables

Derivation of the results for the normal hierarchical model:
Let us use the notations $x = (x_1, \ldots, x_k)$ and $\theta = (\theta_1, \ldots, \theta_k)$. We begin with

---

[31] In Eq. (10.24) the statistic Q could follow a standard distribution, such as a t_ν-distribution, in which case the calculation would be based on the distribution of Q. However, Eq. (10.24) may be rewritten as

$$p = \int_R f_0(x)dx$$

where $R = \{x : Q \geq q_{obs}\}$.

[32] For the normal testing problem of Section 10.3.1, one may consider the class of all pdfs that are symmetric around $\mu = \mu_0$, and also have their mode at $\mu = \mu_0$. Sellke et al. (2001) reported results based on this assumption. They also considered the distribution of the p-value. Under H_0 this distribution is uniform (see Section 10.4.1) and under H_A they assumed it to take the form $f(p) = \xi p^{\xi-1}$ for some ξ, which provided another way to formalize the family of alternatives and compute the minimum value of the Bayes factor.

$$f(\theta|x) = \int f(\theta, \mu|x)d\mu \tag{16.71}$$

and rewrite $f(\theta, \mu|x)$, first using Bayes' theorem to get

$$f(\theta, \mu|x) \propto f(x|\theta, \mu)f(\theta|\mu, \tau^2)$$

and then using

$$f(x|\theta, \mu) = \prod_i f(x_i|\theta_i)$$

to get

$$f(\theta, \mu|x) \propto \left(\prod_i f(x_i|\theta_i)\right) f(\theta|\mu, \tau^2). \tag{16.72}$$

Combining (16.71) and (16.72) we get

$$f(\theta|x) \propto \left(\prod_i f(x_i|\theta_i)\right) \int f(\theta|\mu, \tau^2)d\mu$$

and substituting the normal pdfs, while ignoring factors that do not involve θ, gives

$$f(\theta|x) \propto \exp\left(-\frac{1}{2}\sum_i \sigma_i^{-2}(x_i - \theta_i)^2\right) \int \exp\left(-\frac{1}{2}\tau^{-2}\sum_i (\theta_i - \mu)^2\right) d\mu. \tag{16.73}$$

We next need to evaluate the integral in (16.73), again retaining only the factors involving θ. Letting $\overline{\theta} = k^{-1}\sum_i \theta_i$ we expand and simplify:

$$\begin{aligned}
\sum_i (\theta_i - \mu)^2 &= \left(\sum_i \theta_i^2\right) - 2k\mu\overline{\theta} + k\mu^2 \\
&= \left(\sum_i \theta_i^2\right) + k(\mu^2 - 2\mu\overline{\theta}^2 + \overline{\theta}^2) - k\overline{\theta}^2 \\
&= k(\mu^2 - 2\mu\overline{\theta} + \overline{\theta}^2) + \left(\sum_i \theta_i^2\right) - k\overline{\theta}^2 \\
&= k(\mu^2 - 2\mu\overline{\theta} + \overline{\theta}^2) + \sum_i (\theta_i - \overline{\theta})^2.
\end{aligned} \tag{16.74}$$

From (16.74), the integral in (16.73) may be written, again retaining only terms involving θ, as

$$\int \exp\{-\frac{1}{2}\tau^{-2}\sum_i(\theta_i-\mu)^2\}d\mu \propto \exp\{-\frac{1}{2}\tau^{-2}\sum_i(\theta_i-\overline{\theta})^2\}$$

so that (16.73) becomes

$$f(\theta|x) \propto \exp\left(-(1/2)[\sum_i \sigma_i^{-2}(x_i-\theta_i)^2+\tau^{-2}\sum_i(\theta_i-\overline{\theta})^2]\right). \tag{16.75}$$

We next expand the exponent in (16.75) and collect terms:

$$\begin{aligned}
&\sum_i \sigma_i^{-2}(x_i-\theta_i)^2+\tau^{-2}\sum_i(\theta_i-\overline{\theta})^2\\
&=\sum_i(\sigma_i^{-2}+\tau^{-2})\theta_i^2-2\sum_i(\sigma_i^{-2}x_i+\tau^{-2}\overline{\theta})\theta_i+\tau^{-2}k\overline{\theta}^2+\sum_i\sigma_i^{-2}x_i^2\\
&=\sum_i(\sigma_i^{-2}+\tau^{-2})\theta_i^2-2\sum_i\sigma_i^{-2}x_i\theta_i-2k\tau^{-2}\overline{\theta}^2+\tau^{-2}k\overline{\theta}^2+\sum_i\sigma_i^{-2}x_i^2\\
&=\sum_i(\sigma_i^{-2}+\tau^{-2})\theta_i^2-2\sum_i\sigma_i^{-2}x_i\theta_i-k\tau^{-2}\overline{\theta}^2+\sum_i\sigma_i^{-2}x_i^2. \qquad (16.76)
\end{aligned}$$

We now note that

$$\sum_i\sum_j\theta_i\theta_j=k^2\overline{\theta}^2$$

so that

$$-k\tau^{-2}\overline{\theta}^2=-k^{-1}\tau^{-2}\sum_i\sum_j\theta_i\theta_j \tag{16.77}$$

and inserting (16.77) in (16.76) gives

$$\begin{aligned}
\sum_i\sigma_i^{-2}(x_i-\theta_i)^2+\tau^{-2}\sum_i(\theta_i-\overline{\theta})^2=&\sum_i(\sigma_i^{-2}+\tau^{-2})\theta_i^2\\
&-k^{-1}\tau^{-2}\sum_i\sum_j\theta_i\theta_j-2\sum_i\sigma_i^{-2}x_i\theta_i\\
&+\text{constant} \qquad (16.78)
\end{aligned}$$

which is quadratic in θ (and where "constant" stands for terms not involving θ). In general, for a matrix V and vector z we have

$$\theta^TV^{-1}\theta-2z^T\theta=(\theta-m)^TV^{-1}(\theta-m)-z^Tm$$

where $m=Vz$. We use this by setting $v^{ij}=(V^{-1})_{ij}$ and defining V^{-1} and z according to

$$v^{ij} = -k^{-1}\tau^{-2} + (\sigma_i^{-2} + \tau^{-2})\delta_{ij}$$
$$z_i = \sigma_i^{-2} x_i$$

where δ_{ij} is 1 if $i = j$ and 0 otherwise. We use these definitions in (16.78) to get

$$\sum_i \sigma_i^{-2}(x_i - \theta_i)^2 + \tau^{-2}\sum_i (\theta_i - \overline{\theta})^2 = (\theta - m)^T V^{-1}(\theta - m) + \text{ constant} \quad (16.79)$$

and putting (16.79) in (16.75) gives

$$f(\theta|x) \propto \exp\left(-(1/2)(\theta - m)^T V^{-1}(\theta - m)\right). \quad (16.80)$$

Therefore, the posterior distribution of θ is multivariate normal with expectation vector m and variance matrix V.

All that remains is to write down the explicit formulas for m and V. For this we use the following matrix identity: writing $c^{ij} = (C^{-1})_{ij}$, if

$$c_{ij} = -k^{-1}b + (a_i + b)\delta_{ij}$$

then

$$c^{ij} = \left(\sum_i \frac{a_i b}{a_i + b}\right)^{-1} \left(\frac{b}{a_i + b}\right)\left(\frac{b}{a_j + b}\right) + (a_i + b)^{-1}\delta_{ij}, \quad (16.81)$$

which may be verified by direct calculation. Here we will use

$$a_i = \sigma^{-2} \text{ and } b = \tau^{-2} \quad (16.82)$$

together with the identities

$$\frac{\sigma_i^{-2}\tau^{-2}}{\sigma_i^{-2} + \tau^{-2}} = \left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right)^{-1} \quad (16.83)$$

and

$$\frac{\tau^{-2}}{\sigma_i^{-2} + \tau^{-2}} = \frac{\sigma_i^2}{\sigma_i^2 + \tau^2}. \quad (16.84)$$

Putting (16.82) in (16.81) and applying (16.83) and (16.84), the components of the matrix V become

$$v_{ij} = (\sum_i (\sigma_i^2 + \tau^2)^{-1})^{-1} \left(\frac{\sigma_i^2}{\sigma_i^2 + \tau^2}\right)\left(\frac{\sigma_j^2}{\sigma_j^2 + \tau^2}\right) + \left(\frac{\sigma_i^2 \tau^2}{\sigma_i^2 + \tau^2}\right)\delta_{ij}$$

and the components v_{ii} give the posterior variances (16.36). The posterior means in (16.35) are the components of the vector m, which then become

$$\begin{aligned} m_i &= \sum_i v_{ij}\sigma_j^{-2}x_j \\ &= w_i\bar{x}_\alpha + (1-w_i)x_i \end{aligned}$$

where

$$\begin{aligned} w_i &= \sigma_i^2/(\sigma_i^2+\tau^2) \\ \alpha_i &= (\sigma_i^2+\tau^2)^{-1} \\ \bar{x}_\alpha &= (\sum_i \alpha_i x_i)/(\sum_i \alpha_i). \end{aligned}$$

□

Derivation of the filtering and prediction equations:
In general, to derive Bayes' Theorem for continuous random vectors X and Y we could combine

$$f(x,y) = f(x|y)f(y) \tag{16.85}$$

and its role-reversed counterpart

$$f(x,y) = f(y|x)f(x)$$

to write

$$f(x|y) = \frac{f(y|x)f(x)}{f(y)},$$

and then would use

$$f(y) = \int f(x,y)dx \tag{16.86}$$

to get

$$f(x|y) = \frac{f(y|x)f(x)}{\int f(y|x)f(x)dx},$$

which is often written as

$$f(x|y) \propto f(y|x)f(y). \tag{16.87}$$

Let us put $x = (u,z)$ and $y = w$ in (16.85). We have

$$f(u,z|w) = \frac{f(u,z,w)}{f(w)}$$

and dividing and multiplying the right-hand side by $f(z,w)$ gives

$$f(u, z|w) = \frac{f(u, z, w)}{f(z, w)} \frac{f(z, w)}{f(w)}.$$

Applying (16.85) to each of the fractions on the right-hand side now gives the variation on (16.85) after conditioning:

$$f(u, z|w) = f(u|z, w)f(z|w). \tag{16.88}$$

We may also derive a variation on (16.86): we write

$$f(u|w) = \frac{f(u, w)}{f(w)} = \frac{\int f(u, w, z)dz}{f(w)}$$

and bringing the denominator into the integral we get

$$f(u|w) = \int f(u, z|w)dz. \tag{16.89}$$

Equations (16.88) and (16.89), together with (16.48) and (16.49), are all that are needed to derive the filtering and prediction equations. We begin by writing

$$f(x_t|y_{1:t}) = f(x_t|y_t, y_{1:t-1})$$

and then set $z = y_t$, $w = y_{1:t-1}$ and $u = x_t$ in (16.88) to get

$$\begin{aligned} f(x_t|y_{1:t}) &= f(x_t|y_t, y_{1:t-1}) \\ &= \frac{f(y_t, x_t|y_{1:t-1})}{f(y_t|y_{1:t-1})}. \end{aligned}$$

We continue by applying (16.88) again with $u = y_t$, $z = x_t$ and $w = y_{1:t-1}$,

$$\begin{aligned} f(x_t|y_{1:t}) &= \frac{f(y_t, x_t|y_{1:t-1})}{f(y_t|y_{1:t-1})} \\ &= \frac{f(y_t|x_t, y_{1:t-1})f(x_t|y_{1:t-1})}{f(y_t|y_{1:t-1})} \end{aligned}$$

and then by (16.89),

$$\begin{aligned} f(x_t|y_{1:t}) &= \frac{f(y_t|x_t, y_{1:t-1})f(x_t|y_{1:t-1})}{f(y_t|y_{1:t-1})} \\ &= \frac{f(y_t|x_t, y_{1:t-1})f(x_t|y_{1:t-1})}{\int f(y_t|x_t, y_{1:t-1})f(x_t|y_{1:t-1})dx_t}. \end{aligned}$$

From (16.49), omitting the denominator as in (16.87), this gives (16.51). Now, putting $z = x_{t-1}$, $w = y_{1:t-1}$ and $u = x_t$ we apply (16.89) and then (16.88) to get

$$f(x_t|y_{1:t-1}) = \int f(x_t, x_{t-1}|y_{1:t-1})dx_{t-1}$$
$$= \int f(x_t|x_{t-1}, y_{1:t-1})f(x_{t-1}|y_{1:t-1})dx_{t-1}$$

and from (16.48) we then have (16.52). □

Derivation of the Kalman filter:
To obtain the posterior distribution of X_t conditionally on $Y_{1:t} = y_{1:t}$ we will use the theorem in Section 5.5.3 that gives the conditional distribution of one random vector given another when they are jointly multivariate normal. First, we observe that, conditionally on $Y_{1:t-1} = y_{1:t-1}$ the vector (Y_t, X_t) is multivariate normal. We will need the formulas for the mean and variance of X_t given $Y_{1:t-1} = y_{1:t-1}$. We begin with

$$\hat{x}_{t|t-1} = E(X_t|Y_{1:t-1} = y_{1:t-1})$$

and then rewrite:

$$\begin{aligned}\hat{x}_{t|t-1} &= E(AX_{t-1} + \epsilon_t|Y_{1:t-1} = y_{1:t-1}) \\ &= AE(X_{t-1}|Y_{1:t-1} = y_{1:t-1}) + E(\epsilon_t|Y_{1:t-1} = y_{1:t-1}) \\ &= A\hat{x}_{t-1|t-1}\end{aligned}$$

which is (16.58). Next we define

$$\hat{W}_{t|t-1} = V(X_t|Y_{1:t-1} = y_{1:t-1})$$

and have

$$\begin{aligned}\hat{W}_{t|t-1} &= V(AX_{t-1} + \epsilon_t|Y_{1:t-1} = y_{1:t-1}) \\ &= AV(X_{t-1}|Y_{1:t-1} = y_{1:t-1})A^T + V(\epsilon_t|Y_{1:t-1} = y_{1:t-1}) \\ &= A\hat{W}_{t-1|t-1}A^T + Q\end{aligned}$$

which is (16.59). Now let $\mu_b = E(X_t|Y_{1:t-1} = y_{1:t-1})$. By Eq. (16.58) we have

$$\mu_b = A\hat{x}_{t-1|t-1}.$$

We also let $\mu_a = E(Y_t|Y_{1:t-1} = y_{1:t-1})$ and then

$$\begin{aligned}\mu_a &= E(BX_t + \eta_t|Y_{1:t-1} = y_{1:t-1}) \\ &= BE(X_t|Y_{1:t-1} = y_{1:t-1}) + E(\eta_t|Y_{1:t-1} = y_{1:t-1}) \\ &= B\hat{x}_{t|t-1}.\end{aligned}$$

Next, we compute the covariance matrix. Let $\Sigma_{bb} = V(X_t|Y_{1:t-1} = y_{1:t-1})$. By Eq. (16.59) we may write $\Sigma_{bb} = A\hat{W}_{t-1|t-1}A^T + Q$. Let $\Sigma_{aa} = V(Y_t|Y_{1:t-1} = y_{1:t-1})$. We have

$$\begin{aligned}\Sigma_{aa} &= V(BX_t + \eta_t|Y_{1:t-1} = y_{1:t-1}) \\ &= BV(X_t|Y_{1:t-1} = y_{1:t-1})B^T + V(\eta_t|Y_{1:t-1} = y_{1:t-1}) \\ &= B\hat{W}_{t|t-1}B^T + R.\end{aligned}$$

Finally, let $\Sigma_{ab} = Cov(Y_t, X_t|Y_{1:t-1} = y_{1:t-1})$. We get

$$\begin{aligned}\Sigma_{ab} &= Cov(BX_t + \eta_t, X_t|Y_{1:t-1} = y_{1:t-1}) \\ &= BV(X_t|Y_{1:t-1} = y_{1:t-1}) + Cov[\eta_t, X_t|Y_{1:t-1} = y_{1:t-1}) \\ &= B\hat{W}_{t|t-1}.\end{aligned}$$

Because Σ is symmetric, $\Sigma_{ba} = \Sigma_{ab}$.

We have defined μ_a, μ_b, Σ_{ab} and Σ_{ba} to match the notation in Section 5.5.3. Using the theorem in that section, the joint multivariate normal distribution can be written

$$\begin{pmatrix} Y_t \\ X_t \end{pmatrix} \Big| Y_{1:t-1} = y_{1:t-1} \sim N_{2t}(\mu, \Sigma)$$

where

$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}$$

and Σ is given by Eq. (5.18), with μ_a, μ_b, Σ_{ab} and Σ_{ba} defined above.

To conclude the derivation we write $\hat{x}_{t|t} = \mu_{b|a}$ and $\hat{W}_{t|t} = \Sigma_{b|a}$ which, by Eqs. (5.19) and (5.20), are given by

$$\begin{aligned}\mu_{b|a} &= \mu_b + \Sigma_{ba}\Sigma_{aa}^{-1}(y_t - \mu_a) \\ &= \hat{x}_{t|t-1} + \hat{W}_{t|t-1}B^T(B\hat{W}_{t|t-1}B^T + R)^{-1}(y_t - B\hat{x}_{t|t-1})\end{aligned}$$

and

$$\begin{aligned}\Sigma_{b|a} &= \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab} \\ &= \hat{W}_{t|t-1} - \hat{W}_{t|t-1}B^T(B\hat{W}_{t|t-1}B^T + R)^{-1}B\hat{W}_{t|t-1}\end{aligned}$$

which are (16.57) and (16.60). □

Chapter 17
Multivariate Analysis

17.1 Introduction

Much of this book has been devoted to describing relationships among multiple noisy variables, yet we have until now managed to avoid a general discussion of multivariate co-variation. The regression and generalized regression models discussed in Chapters 12, 14, and 15 involved a response variable y that was related to one or more explanatory variables x and this asymmetry of response and explanatory variables allowed us, for the most part, to ignore the co-variation among the whole set of measured variables. In some contexts, however, there are advantages to analyzing multiple measurements together. For instance, in Example 4.7 (p. 100), which involved decoding of wrist movement from MEG signals, the signals came from 87 MEG sensors and it made sense to analyze these collectively, as an 87-dimensional vector at each time point. In this chapter we provide a short overview of methods that have been developed for such purposes, which fall under the heading of *multivariate analysis*, and we return to Example 4.7 on p. 494.

The starting point is the sample mean and sample variance matrix (see Section 4.3.1), while the theory is based largely on the theoretical mean and variance of a random vector (see Section 4.3.1) together with the multivariate normal distribution (see Section 5.5). Section 17.2 reviews the multivariate extensions of t-tests and one-way ANOVA, which are special cases of the general class of methods called *multivariate analysis of variance* (*MANOVA*). MANOVA balances two competing tendencies. On the one hand, when several variables respond similarly to a change in experimental conditions there is stronger evidence for differential response in their combined data than would be provided if each variable were considered separately. This was the idea behind the method of combining p-values from independent tests of the same null hypothesis, described in Section 11.3.1; in Example 11.2 we found that five separate p-values of .02 led to a combined p-value of 2.5×10^{-5}. On the other hand, if the multiple variables are correlated, the assessment must take account of the correlation, and this tends to decrease the effect: in the extreme case of perfect correlation, observing multiple variables becomes the same thing as observing a single

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_17,

variable. MANOVA incorporates correlation by comparing multivariate co-variation across conditions to that within conditions.

Section 17.3 reviews the main ideas behind dimensionality reduction. When the multiple variables are, collectively, so highly correlated that a variance matrix is no longer of full rank, i.e., no longer positive definite (see p. 618 of the Appendix), some formulas are voided. A solution to this problem is to define a smaller set of new variables that are linear combinations of the original variables, the process of which is called "dimensionality reduction" (though, in general, the combinations do not have to be linear). Dimensionality reduction is also useful for data simplification. For example, data are often displayed by plotting with x and y axes that are suitably defined by a reduction to 2 dimensions.

Section 17.4 returns to the problem of classification, introduced in Section 4.3.4. We first show how Bayes classifiers take a nice form when the classes are defined by multivariate normal distributions, and then go on to describe two commonly-applied alternative methods of classification. In Section 17.4.3 we discuss the concept of *clustering*, which involves putting observations into classes when the classes have not yet been defined and must be estimated or[1] *learned* from the data.

Multivariate analysis uses more advanced mathematics than univariate analysis, and many theoretically-inclined students find in the subject a majestic elegance. While nearly all the methods presented in our synopsis here were developed more than 50 years ago, it is a very active area of continuing research.

17.2 Multivariate Analysis of Variance

17.2.1 MANOVA provides a multivariate extension of ANOVA.

The one-way ANOVA model, given in Eq. (13.1), involves a set of random variables Y_{ij}. We repeat Eq. (13.1) here as Eq. (17.1). The model is

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \tag{17.1}$$

for $i = 1, \ldots, I$ and $j = 1, \ldots, n_i$ and the usual assumptions are

(i) the ANOVA model (13.1) holds;
(ii) the errors satisfy $E(\epsilon_i) = 0$ for all i;
(iii) the errors ϵ_i are independent of each other;
(iv-1D) $V(\epsilon_i) = \sigma^2$ for all i (homogeneity of error variances), and
(v-1D) $\epsilon_i \sim N(0, \sigma^2)$ (normality of the errors).

[1] The term "learning" tends to be used interchangeably with "estimation," i.e., the process of determining a parameter value from data. Because it may sometimes refer to significance testing, learning is somewhat broader, and it is often associated with techniques used heavily in the field of machine learning. See Hastie et al. (2009).

In Eq. (17.1) each ϵ_{ij} is a random variable, and μ and each α_i are numbers. If we instead take all Y_{ij} and ϵ_{ij} to be p-dimensional random vectors, and μ and all α_i to be vectors, then the model becomes

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \tag{17.2}$$

which is identical to (17.1) when $p = 1$. The usual assumptions (i–iii) have the same form for (17.2) as for (17.1) while the assumptions we labeled (iv-1D) and (v-1D) become

(iv) $V(\epsilon_i) = \Sigma$ for all i;
(v) $\epsilon_i \sim N(0, \Sigma)$.

Equation (17.2) together with these multivariate assumptions (i–v) then becomes a multivariate analysis of variance (MANOVA) model. Note that in this section we are using Y_{ij} to denote our generic random vector, while in the rest of this chapter we use X.

The idea behind one-way ANOVA is to test the null hypothesis

$$H_0 : \alpha_i = 0 \tag{17.3}$$

by, first, decomposing the total sum of squares

$$SST = \sum_{i,j} (y_{ij} - \bar{y}_{..})^2 \tag{17.4}$$

using the error sum of squares

$$SSE = \sum_{i,j} (y_{ij} - \bar{y}_{i.})^2 \tag{17.5}$$

as

$$SST = SS_{group} + SSE \tag{17.6}$$

where SS_{group} is defined from (17.6) by subtraction and, second, considering whether[2] SS_{group} is improbably large relative to SSE under H_0. The same idea may be applied in the multivariate case: formulas (17.4) and (17.5) become

$$SST = \sum_{i,j} (y_{ij} - \bar{y}_{..})(y_{ij} - \bar{y}_{..})^T \tag{17.7}$$

and

[2] In constructing the F-statistic, the values of SS_{group} and SSE are first standardized by dividing by their respective degrees of freedom, but that is for the convenience of judging the ratio relative to the number 1.

$$SSE = \sum_{i,j} (y_{ij} - \bar{y}_{i\cdot})(y_{ij} - \bar{y}_{i\cdot})^T \tag{17.8}$$

and then (17.6) may be applied. In addition, under the homogeneity of variance assumption (iv), an estimate of Σ is the *pooled sample variance matrix*

$$S_{pooled} = \frac{1}{N - I} SSE \tag{17.9}$$

where

$$N = \sum_{i=1}^{I} = n_{i\cdot}$$

On p. 367 we outlined the way the usual one-way ANOVA F-test arises as a likelihood ratio test. This suggests applying a likelihood ratio test in the multivariate setting. The result rejects the null hypothesis of (17.3) when *SST* is large relative to *SSE*, where "large" now refers to a matrix and is measured by the determinant (see the appendix, p. 616). Equivalently, the test rejects when the quantity

$$\Lambda = \frac{|SSE|}{|SST|} \tag{17.10}$$

is small. The test was derived by Wilks (1932) and the value Λ is usually called *Wilks' lambda*. An F statistic may be defined in terms of Λ (the expression is not very intuitive; we omit it) and this statistic has, approximately, an F distribution under H_0. The results are usually displayed in a table, much like the ANOVA table given as Table 13.4.

Example 17.1 Functional Specialization of Mouse Visual Areas Because of the potential for genetic manipulation, there is great interest in mouse models of brain function. Cortical areas in the primate visual system can be distinguished according to their differing neural responses. Marshel et al. (2011) sought to provide a similar characterization of mouse visual areas. Specifically, they examined the tuning properties of individual neurons with respect to direction, orientation, spatial frequency, and temporal frequency, across seven visual areas. For each tuning property they devised a measure of sensitivity, yielding a 4-dimensional vector for each neuron. The authors then applied MANOVA to look for differential neural responses in these 4-dimensional vectors across the seven areas. They found the seven areas to be distinguishable using MANOVA, and then proceeded to provide more detailed comparisons for each metric. □

Example 4.7 (continued from p. 100) In their study of decoding wrist movement from MEG sensor recordings, Wang et al. used Bayes classifiers to produce the results in Fig. 4.4. They also evaluated the classification accuracy after averaging the

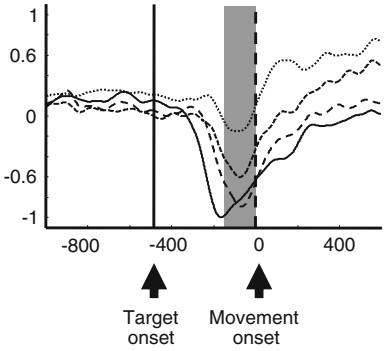

Fig. 17.1 Normalized MEG sensor signals from one subject in the Wang et al. study, averaged across trials. Four traces are shown for a single sensor, corresponding to the four directions of movement. The *shadowed gray region* is the optimal time window found by MANOVA. Adapted from Wang et al. (2010).

sensor recordings across 200 ms time windows. To compute classification accuracy, leave-one-out cross-validation was used. For each subject, and for each trial i, the movement direction on trial i was predicted after the remainder of the trials were used as training data. Using the training data, first an optimal time window for each subject was chosen and then a Bayes classifier was defined (it was assumed that sensor measurements were multivariate normal and the mean and variance parameters were estimated for each of the four directions of movement; see p. 506). The optimal time window of length 200 ms was chosen from 150 possible windows, centered at 150 time points spaced 10 ms apart. To select the optimal time window the authors applied MANOVA in each of the 150 windows, then found the window that produced the largest F statistic. See Fig. 17.1. □

In Section 13.1.3 we said that in the case of two groups, one-way ANOVA reduces to the usual t-test. Similarly, in the case of two groups, MANOVA may be reduced to a simpler form. Let us assume there are n_1 observations in group 1 and n_2 observations in group 2. The pooled sample variance matrix of Eq. (17.9) becomes

$$S_{pooled} = \frac{1}{n_1 + n_2 - 2}\left(\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)(y_{1j} - \bar{y}_1)^T + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)(y_{2j} - \bar{y}_2)^T\right) \tag{17.11}$$

which is analogous to the univariate S^2_{pooled} defined in Section 10.3.4. Let us change the notation x used in Section 10.3.4 to y as used here and then write the t-statistic (10.19) in the squared form

$$t^2_{obs} = (\bar{y}_1 - \bar{y}_2)\left((\frac{1}{n_1} + \frac{1}{n_2})s^2_{pooled}\right)^{-1}(\bar{y}_1 - \bar{y}_2). \tag{17.12}$$

The standard test statistic for testing $H_0 : \alpha_1 - \alpha_2 = 0$ in the multivariate case is

$$T^2 = (\bar{y}_1 - \bar{y}_2)^T \left((\frac{1}{n_1} + \frac{1}{n_2}) S_{pooled} \right)^{-1} (\bar{y}_1 - \bar{y}_2), \tag{17.13}$$

where S_{pooled} is defined above, which is a generalization of (17.12). The statistic T^2 is usually called *Hotelling's* T^2. In this case, under H_0 and the assumptions following Eq. (17.2), including the normality assumption (v), the approximate F distribution of the MANOVA F statistic found by the likelihood ratio test becomes exact and[3] we have

$$\frac{n_1 + n_2 - p}{(n_1 + n_2 - 1)p} T^2 \sim F_{p, n_1 + n_2 - p}.$$

We have discussed one-way MANOVA here, but similar ideas apply to multivariate extensions of two-way ANOVA and more complicated ANOVA designs.

17.2.2 When the variance matrices across conditions are unequal, the likelihood ratio test may be applied.

It sometimes happens that the homogeneity assumption (iv) in the multivariate model (17.2) is violated. The likelihood ratio test may still be used, and p-values may be obtained by simulation.

Example 17.2 Testing Equality of Time-Varying Firing Rates One way to compare the responses of a neuron across two or more experimental conditions is to pick a window of time, compute the spike counts within that window for each of many trials, and then apply a t-test or ANOVA or, possibly, a generalized version of these as in Table 14.7. Sometimes, however, the firing rate may fluctuate across the recorded time interval and it may not be clear what time window would be most appropriate.

Behseta and Kass (2005) and Behseta et al. (2007) suggested, instead, testing the null hypothesis that the firing rate, as a varying function of time, remains the same across the two or more conditions. The situation is illustrated in Fig. 17.2. In the two upper left panels are PSTHs for a motor cortical neuron under two experimental conditions together with smoothed versions of the PSTHs, obtained by methods similar to those of Example 1.1 on p. 422.

The smooth curves in Fig. 17.2 may be considered estimated firing-rate functions, which vary across time. Section 19.3.3 spells this out by defining what is called the *marginal intensity* function $\lambda(t)$ (Eq. (19.23), which is the trial-averaged firing

[3] Here we are using T^2 both as an observed value of a statistic based on data and as a random variable that has a probability distribution. To be consistent with earlier notation, in using T^2 as a random variable we should replace $\bar{y}_1$ and $\bar{y}_2$ in (17.13) and (17.11) with $\bar{Y}_1$ and $\bar{Y}_2$.

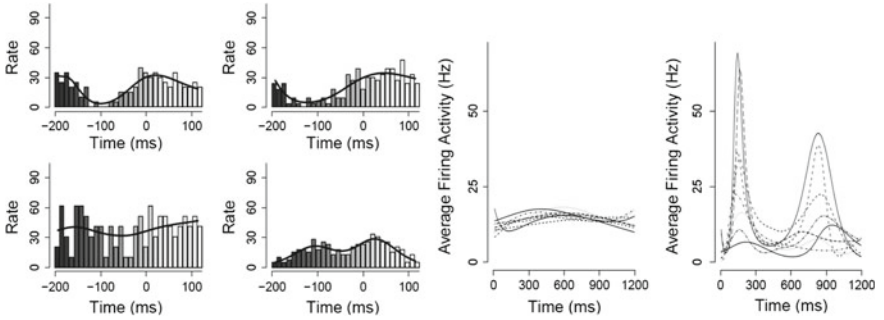

Fig. 17.2 *Left* Responses of two motor cortical neurons. Shown are PSTHs together with *smoothed* versions (*black curves*) obtained from *BARS* (Section 15.2.6). In the two *upper panels* are the estimated firing-rate functions of neuron 1 under two different experimental conditions; for this neuron the firing-rate functions look very similar. In the *lower two panels* are the corresponding estimated firing-rate functions of neuron 2, which look clearly different. Adapted from Behseta and Kass (2005). *Right* Responses of two neurons from the supplementary eye field during eye movements in eight different directions. The first neuron has nearly flat firing-rate functions in all directions, while the second neuron has modulated firing-rate functions which look clearly different. Adapted from Behseta et al. (2007).

rate function (Eq. (19.25)) and, as explained there, the PSTH may be considered an estimate of $\lambda(t)$. To avoid confusion with our use, in this chapter, of λ to denote an eigenvalue, we will here write the trial-averaged firing-rate function instead as $g(t)$. In the case of two firing-rate functions $g_1(t)$ and $g_2(t)$ under two experimental conditions, the null hypothesis becomes $H_0\colon\ g_1(t) = g_2(t)$ for all t.The smooth curves in the left panels of Fig. 17.2 become estimates $\hat{g}_1(t)$ and $\hat{g}_2(t)$. Behseta and Kass (2005) showed how a version of the T^2 test in (17.13) could be defined from the smooth curves $\hat{g}_1(t)$ and $\hat{g}_2(t)$, together with their estimated variance matrices that come from the smoothing algorithm. As would be expected from Fig. 17.2, the test was not significant for the firing-rate curves in the two upper left panels but was highly significant for the firing-rate curves in the two lower left panels.

Behseta et al. (2007) went on to derive a likelihood ratio test for the more general case in which there are I conditions ($I \geq 2$) and the null hypothesis becomes H_0 : $g_1(t) = g_2(t) = \cdots = g_I(t)$ for all t. This applies to the right-hand panels of Fig. 17.2, which display smoothed firing-rate functions from a pair of supplementary eye field neurons for eye movements in eight directions ($I = 8$). To treat this situation, Behseta et al. 2007 had to allow for the possibility that the variance matrices in each group might be different. Again, the test was not significant for the curves shown for the first neuron but was highly significant for the curves shown for the second neuron. □

17.3 Dimensionality Reduction

17.3.1 A variance matrix may be decomposed into principal components.

The variability of an m-dimensional random vector X is summarized by[4] its variance matrix Σ. According to the spectral decomposition (see p. 617 of the Appendix), we may decompose Σ in the form

$$\Sigma = PDP^T \tag{17.14}$$

where D is an $m \times m$ diagonal matrix and P is an $m \times m$ orthogonal matrix. As discussed on p. 618, the equation $x^T \Sigma x = 1$ defines an m-dimensional ellipse (or *ellipsoid*) the axes of which are defined by the columns of P, which are eigenvectors of Σ. The lengths of these axes are twice the square-root of the corresponding eigenvalues, which are the diagonal elements of D.

Using (12.59) together with the orthogonality relationships $P^T P = PP^T = I_m$, where I_m is the m-dimensional identity matrix, the transformed random vector

$$Y = P^T X \tag{17.15}$$

has variance matrix

$$V(Y) = P^T(PDP^T)P = D. \tag{17.16}$$

Let us assume that the columns of P and diagonal elements of D have been ordered so that $D_{11} \geq D_{22} \geq \cdots \geq D_{mm}$. These diagonal elements, which are eigenvalues of Σ, are usually written $\lambda_j = D_{jj}$, so that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m.$$

Then, if $\text{col}_j(P)$ is the jth column of P (the jth eigenvector of Σ) the jth component Y_j of Y is given by

$$Y_j = \text{col}_j(P)^T X \tag{17.17}$$

and its variance is

$$V(Y_j) = \text{col}_j(P)^T X = \lambda_j. \tag{17.18}$$

Also, when $i \neq j$, Y_i and Y_j are uncorrelated. If X is multivariate normal, then Y_i and Y_j are independent.

Now for any unit vector u we have

[4] This assumes that the variance matrix is well-defined in the sense that every linear combination $a^T X$ has finite variance. There exist multivariate distributions for which nonzero linear combinations $a^T X$ have infinite variance. We do not consider these here.

$$V(u^T X) \leq \lambda_1. \tag{17.19}$$

Details: Let $w = P^T u$. Notice that

$$w^T w = u^T P P^T u = u^T u = 1$$

so that w is also a unit vector. We compute $V(u^T X)$:

$$V(u^T X) = u^T P D P^T u = w^T D w = \sum_{j=1}^{m} w_j^2 \lambda_j$$

and because $\lambda_j \leq \lambda_1$, we get

$$\sum_{j=1}^{m} w_j^2 \lambda_j \leq \lambda_1 \sum_{j=1}^{m} w_j^2 = \lambda_1.$$

□

Meanwhile, from (17.18) we have that the special case $u = \text{col}_1(P)$ gives

$$V(Y_1) = \lambda_1. \tag{17.20}$$

Together, (17.19) and (17.20) show that Y_1 is the linear combination of components of X that maximizes the variance, among all linear combinations scaled so that the coefficients define a unit vector. In this sense, $\text{col}_j(P)$, the first eigenvector of Σ, defines the *direction of maximal variation* of the random vector X. The linear combination Y_1 is called the *first principal component* of Σ or, more loosely, the first principal component of the distribution of X. Sometimes the term "first principal component" is applied to the first eigenvector $\text{col}_1(P)$.

A similar argument shows that Y_m is the linear combination of components of X that minimizes the variance, among all linear combinations scaled so that the coefficients define a unit vector. With a little more algebra it may also be shown that among all unit vectors u that are perpendicular to $\text{col}_1(P)$, the variance $V(u^T X)$ is maximized by $u = \text{col}_2(P)$. Similarly, $\text{col}_j(P)$ maximizes the variance $V(u^T X)$ among all unit vectors u that are perpendicular to all of $\text{col}_1(P), \text{col}_2(P), \ldots, \text{col}_k(P)$, where $k = j-1$. The linear combination Y_j is called the jth principal component of Σ.

To summarize, the transformation (17.15), based on the eigenvectors of Σ, produces a new version of X consisting of its principal components. The principal components, given by (17.17), are rotated versions of the components of X that are uncorrelated. If X is multivariate normal, then the principal components are mutually independent. Furthermore, the principal components indicate directions of maximal variation of X in the sense outlined above: the first principal component is in the direction of maximal variation of X, the second principal component is in the direction of maximal variation of X subject to being orthogonal to the first principal component,

the third principal component is in the direction of maximal variation of X subject to being orthogonal to the first two principal components, and so on.

Similar analysis may be applied to the sample variance matrix S, defined on p. 90. In this case, we speak of the principal components of S, or of the data vector. This assumes S is of full rank m, i.e. it is positive definite (see p. 617 of the Appendix).

On p. 131 we noted that when a variance matrix Σ is less than full rank, some of its eigenvalues are equal to 0. Suppose there are k positive eigenvalues. Then, as noted on p. 131, Σ may be decomposed instead in terms of the first k eigenvectors, corresponding only to the k positive eigenvalues. These eigenvectors define a k-dimensional subspace in which the variation of X is concentrated. In the case of a sample variance matrix S, which may be considered a noisy estimate of a theoretical variance matrix Σ, the smallest eigenvalues may not be numerically equal to 0 but several may be very close to 0. If we choose a suitable cutoff value c, below which we will say that the smallest eigenvalues are, for practical purposes, the same as 0, then we have effectively determined that there are k positive eigenvalues and the data vector lies in a k-dimensional space. This is the starting point for the idea of dimensionality reduction via principal components: to reduce the dimensionality of a random vector we consider the subspace (the set of linear combinations of its components) corresponding to the positive eigenvalues of its covariance matrix.

Example 17.2 (continued from p. 496) The analysis of Behseta and Kass (2005) involved picking a grid of time values $t_1, \ldots, t_m$ at which to evaluate $\hat{g}_1(t)$ and $\hat{g}_2(t)$. This produced m-dimensional data vectors $(\hat{g}_1(t_1), \ldots, \hat{g}_1(t_m))$ and $(\hat{g}_2(t_1), \ldots, \hat{g}_2(t_m))$ that could be compared based on estimated variance matrices S_1 and S_2 that came from the smoothing method. The authors showed how a statistic similar to T^2 could be defined by replacing the matrix representing the variance of the difference of means, $(\frac{1}{n_1} + \frac{1}{n_2})S_{pooled}$, with $W = S_1 + S_2$, where

$$S_1 = V\left((\hat{g}_1(t_1), \ldots, \hat{g}_1(t_m))\right)$$
$$S_2 = V\left((\hat{g}_2(t_1), \ldots, \hat{g}_2(t_m))\right)$$

which, by independence of the data under the two conditions, satisfies

$$W = V\left((\hat{g}_1(t_1), \ldots, \hat{g}_1(t_m)) - (\hat{g}_2(t_1), \ldots, \hat{g}_2(t_m))\right).$$

Specifically, letting

$$U_1 = (\hat{g}_1(t_1), \ldots, \hat{g}_1(t_m))$$
$$U_2 = (\hat{g}_2(t_1), \ldots, \hat{g}_2(t_m))$$

they wished to use a statistic T^2_{curves} given by

$$T^2_{curves} = (U_1 - U_2)^T W^{-1}(U_1 - U_2). \tag{17.21}$$

However, because the grid comprised many time points (m was relatively large), the matrix W was less than full rank, so that (17.21) could not be applied. The authors therefore reduced dimensionality by choosing a suitable small positive number c and retained only the eigenvalues λ_j of W for which $\lambda_j > c$. (The value of c will be discussed below p. 501.) Let us suppose there were k retained eigenvalues, let D_k be the $k \times k$ diagonal matrix having λ_j as its jth diagonal element, and let P_k be the corresponding matrix of the first k eigenvectors of W. Although the matrix $W = PDP^T$ was not of full rank, the $k \times k$ matrix W_k defined by

$$W_k = P_k D_k P_k^T$$

was of full rank k and the new version of the statistic

$$T^2_{curves} = (U_1 - U_2)^T W_k^{-1} (U_1 - U_2)$$

was well-defined. □

The choice of the cutoff c, below which the remaining eigenvalues are treated as equal to 0, is important. As c increases, additional eigenvalues are set to 0 and dimensionality is further reduced. For a given theoretical variance matrix Σ we may identify the eigenvalues that are zero and then consider the subspace corresponding to the positive eigenvalues. But if all we have is a sample variance matrix S, which we view as a noisy estimate of Σ, it may be difficult to determine how many of the corresponding theoretical eigenvalues of Σ are 0. This gives rise to a dramatic extension of the idea of dimensionality reduction: instead of finding a cutoff for which the remaining eigenvalues are nearly 0, the value c could represent a cutoff for which "most of the variation" in the data occurs in the remaining subspace. For this purpose, a standard procedure is to compute the eigenvalues $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_m$ of S (which are considered to be estimates of $\lambda_1, \lambda_2, \ldots, \lambda_m$) and to declare that the subspace corresponding to the first k eigenvalues *contains a proportion q of the variability* in the data, where q is defined by

$$q = \frac{\hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_k}{\hat{\lambda}_1 + \hat{\lambda}_2 + \cdots \hat{\lambda}_k + \hat{\lambda}_{k+1} + \cdots + \hat{\lambda}_m}.$$

Data analysts often pick k such that 90 or 95 % of the variability is, in this sense, contained in the subspace defined by the first k principal components.

Example 17.3 Postural Hand Synergies Santello et al. (1998) asked subjects to shape their hand as if grasping and using many familiar objects. The authors defined hand shape using 15 joint angles formed when the subjects were in a static grasp position. The authors reported that roughly 90 % of the variability in these hand shape vectors was accounted for by the first three principal components. They interpreted the 3-dimensional representation to be defined by "synergies," meaning shape combinations resulting from the redundancies in hand movement. □

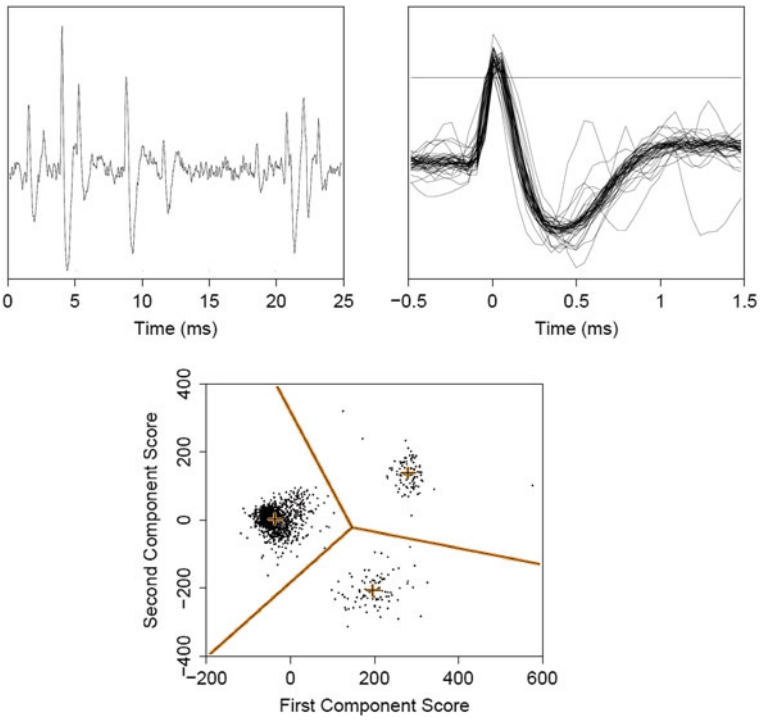

Fig. 17.3 *Top left* A segment of an extracellular electrode voltage recording. *Top right* A plot overlaying the many waveforms from a well-isolated neuron. *Bottom* Three clusters, with cluster boundaries, plotted using axes defined by the first two principal components. The boundaries separating the clusters are defined by $K-$means clustering (see Section 17.4.3). Adapted from Lewicki (1998).

Principal components are also used to visualize data. Let us write the data vectors as $x^1, x^2, \ldots, x^N$. Typically, plots are made of the first two principal components, that is, of the data pairs $(\text{col}_1(P)^T x^i, \text{col}_2(P)^T x^i)$, for $i = 1, \ldots, N$.

Example 17.4 Spike sorting Forebrain Recordings In Example 4.1 we described the problem of spike sorting. Lewicki (1998) reviewed methods and issues and, to illustrate, used a recording from a Zebra finch forebrain. An extracellular electrode records voltage impulses from many different neurons, but each neuron contributes waveforms that are very similar in shape. Several waveforms, apparently from the same neuron, are overlaid in the left panel of Fig. 17.3. Spike sorting attempts to put similar waveforms together into groups or *clusters*, under the assumption that those within a given cluster are likely to emanate from a particular neuron. This poses the statistical machine learning problem of *clustering*, which we discuss in Section 17.4.3.

A spike waveform has a duration of roughly 1.5 ms. If voltage is sampled at 40 kHz (kilohertz) each waveform is a vector of length 60. The data are then all of the wave-

forms in a recording session, represented as vectors of length 60. Some methods of clustering (including the mixture-of-Gaussians method discussed in Section 17.4.3) have difficulty in high dimensions and it is advantageous to reduce dimensionality. In addition, it can be useful to visualize the data in a two-dimensional space. Principal components may be used for these purposes. The bottom panel of Fig. 17.3 displays a set of the Zebra finch forebrain data plotted using the first two principal components. Three distinct clusters appear, corresponding to waveforms that become identified as coming from three distinct neurons. □

The use of principal components for any purpose is usually called *principal component analysis* (*PCA*).

17.3.2 Methods other than PCA may be used to reduce dimensionality.

Principal component analysis can be very effective in reducing dimensionality of multivariate data that are more-or-less normally distributed. The assumption is that a substantial fraction of the variation lies in a linear subspace, which may be obtained from the principal components corresponding to the large eigenvalues of the variance matrix. Alternatives include methods that attempt to find latent factors, possibly while assuming the data to be non-normal, and methods that assume variation is concentrated in nonlinear subspaces (concentrated in subspaces known[5] as smooth manifolds). We do not discuss methods aimed at finding smooth manifolds on which the variation of X is concentrated, which come under the rubric *manifold learning*. We very briefly describe two other approaches to dimensionality reduction.

The usual *factor analysis* model for an m-dimensional random vector X is given in terms of an $m \times p$ matrix A and a p-dimensional random vector S, with $p < m$, by

$$X = AS + \epsilon$$

where the components of S and ϵ satisfy $S_i \sim N(0, 1)$ and $\epsilon_i \sim N(0, \sigma_i^2)$, all independently, for $i = 1, \ldots, m$. (In this section we are using S to stand for a vector "source" of variation, rather than a sample variance matrix.) The intuition is that the variation of X is driven by a set of p *latent factors*, which are the unobserved (thus, latent, as in Section 16.2) components of S, plus independent noise, and the rows of the matrix A contain the coefficients, called *factor loadings*, that define the combination of factors determining each component of X. Because a fit of the model to data will produce latent factors, and the factor loadings become interpretable, this conception is very appealing. It suffers, however, from a serious difficulty: the

[5] A subspace N of R^m is a smooth manifold if at every point $x \in N$ there is a local coordinate representation in which all points near x in N have the form (u, v) where $v = 0$. In other words, everywhere in N there is a local coordinate system that makes N look like a linear subspace. See Appendix A of Kass and Vos (1997).

unknown parameters are the components of the variance matrix $V(X) = \Sigma$ and for any orthogonal matrix P, if we define $B = AP$, using (12.59) and $PP^T = I_m$ we have

$$\begin{aligned} V(BS + \epsilon) = BV(S)B^T + I_m &= API_mP^TA^T + I_m \\ &= AA^T + I_m \\ &= \Sigma. \end{aligned}$$

In other words, we obtain the same variance matrix using both B and A, so an interpretation of factor loadings based on B would be neither more or less valid than an interpretation based on A. There are thus infinitely many equivalent interpretations. Various methods have been used to specify a unique factor loading matrix, but there often remains a degree of arbitrariness that leaves many practitioners wary of resulting interpretations.[6]

A related, but different approach is to begin by allowing the latent vector S to be non-normal, but with independent components, in the linear latent variable model

$$X = AS,$$

where S and X are both m-dimensional and A is taken to be orthogonal. The idea is that the independent components in S would drive the vector X through the linear combinations in A. If S is assumed to be normally distributed, then so is X, and the solution is given by PCA, i.e., S consists of the principal components. However, if S is allowed to be non-normal it can be quite different.

Let us assume the data vector $X = x$ has been standardized (or *pre-whitened*, see p. 557) so that its sample variance matrix is the m-dimensional identity. We wish to find A and s such that $x = As$. By orthogonality $A^TA = I_m$ so that $A^Tx = s$. The matrix A may be defined to minimize the mutual information among the components of $s = A^Tx$, where mutual information is the Kullback-Leibler divergence between the joint pdf and the independence pdf (estimated from the data), as in (4.28). That is, the components of s are taken to be as close to independent as possible, in the sense of mutual information. The resulting procedure is called *independent components analysis* (*ICA*). It turns out that minimizing mutual information in A^Ts has the effect of making the distribution of s as far from normal as possible (measured in terms of entropy).

Example 17.4 Efficient coding of natural sounds Lewicki (2002) used ICA to find components of auditory signals. Some of the components he found from human speech are shown in Fig. 17.4. For comparison, response properties of cochlear neurons are also displayed. There is a qualitative resemblance between the ICA components and the neural response functions. Lewicki argued that ICA may capture an efficient representation of auditory input. □

[6] The most famous example is Spearman's general intelligence index g, which is obtained from factor analysis. See, e.g., Gould (1996); Devlin et al. (1997).

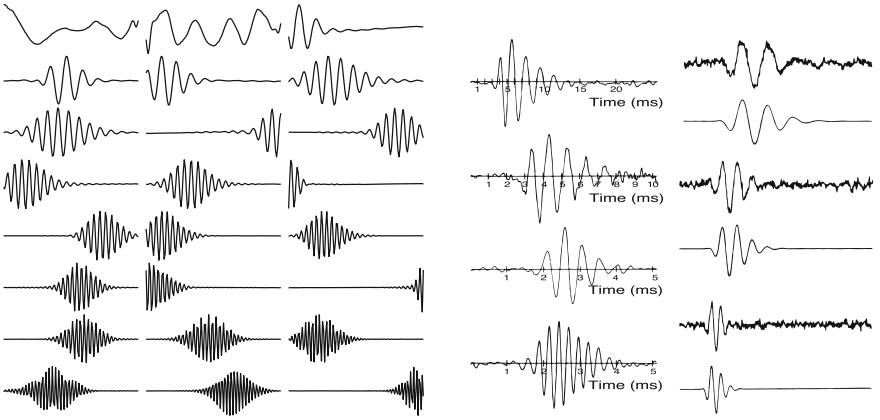

Fig. 17.4 *Left panel* components determined by ICA from human speech. *Right panel* response functions from cochlear neurons. The latter used linear regression of the binary spike train (see Chapter 19) on the input signal at multiple time lags (see p. 530). Adapted from Lewicki (2002).

17.4 Classification and Clustering

17.4.1 Bayes classifiers for multivariate normal distributions take a simple form.

Suppose each of many m-dimensional observation vectors $X = x$ comes from one of K classes $C_1, C_2, \ldots, C_K$, and when it comes from class k the random vector X has pdf $f_k(x)$, for $k = 1, \ldots, K$. The problem of classification (see Section 4.3.4) is to determine, for each observation $X = x$, the class to which x belongs. As we showed in Section 4.3.4, the expected number of classification errors is minimized by using a Bayes classifier. For each x the Bayes classifier finds the class C_k that maximizes the posterior probability given by Eq. (4.38), which we repeat here:

$$P(C = C_k|X = x) = \frac{f_k(x)\pi_k}{\sum_{i=1}^{m} f_i(x)\pi_i}. \tag{17.22}$$

In the special case where, for each class k, we have $X \sim N_m(\mu_k, \Sigma)$ for some μ_k and Σ, the solution takes a simple form. If we write the ratio of posterior probabilities for two classes j and k by plugging the pdfs given by Eq. (5.17) into (17.22), and take logs, after some algebra we obtain

$$\begin{aligned} \log \frac{P(C = C_j|X = x)}{P(C = C_k|X = x)} &= \log \frac{f_j(x)}{f_k(x)} + \log \frac{\pi_j}{\pi_k} \\ &= \delta_j(x) - \delta_k(x) \end{aligned} \tag{17.23}$$

where, for $i = j, k$

$$\delta_i = x^T \Sigma^{-1} \mu_i - \frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \log \pi_i. \tag{17.24}$$

In other words, we have $P(C = C_j | X = x) > P(C = C_k | X = x)$ if and only if $\delta_j(x) > \delta_k(x)$, so that the posterior probability is maximized by selecting the class k that maximizes $\delta_i(x)$. The function $\delta_i(x)$ is a linear function of x. It is called the *linear discriminant function*. Classification based on the linear discriminant function is optimal when the classes are defined by multivariate normal distributions all having the same variance matrix.

A similar argument may be applied to the case in which the classes continue to be defined by multivariate normal distributions but the variance matrices are allowed to be different. In this case the linear discriminant functions $\delta_i(x)$ are replaced by quadratic functions of x, which are then called *quadratic discriminant functions*.

In practice, we do not know π_k, μ_k or Σ_k, even when the latter is assumed to satisfy $\Sigma_1 = \cdots = \Sigma_K = \Sigma$. Assuming we have preliminary data arising from known classes from which to *train* the classifier (such data being called *training data*), each prior probability π_k may be estimated by the proportion of training data vectors that fall in class k, i.e., number of training vectors within class k divided by the total number of training data vectors; and we may replace the theoretical means and variance matrices μ_k and Σ_k by the corresponding sample mean and variance calculated within class k. When, for simplicity, it is assumed that $\Sigma_1 = \cdots = \Sigma_K = \Sigma$ the sample variance matrix is pooled across classes as in MANOVA, i.e., the matrix S_{pooled} defined in (17.9) is used, where the groups become the classes. The resulting classification method is called *linear discriminant analysis* (LDA).

Example 4.7 (continued from p. 494) To classify movement directions based on the MEG sensor signals within a 200 ms time window (see Fig. 17.1), Wang et al. used LDA. With this approach the authors reported 4-direction classification accuracies (with chance being 25 %), among nine subjects, ranging from 51.3 to 88.6 % (with a mean of 67 %) for overt movement and 39.6–95 % (with a mean of 62.5 %) for imagined movement. □

LDA often performs well for noisy data, even when the variation is strikingly non-normal. However, for highly structured data alternative methods can do better. See Section 17.4.2.

17.4.2 Bayes classifiers are not always practical.

The optimal performance of Bayes classifiers depends on the use of the pdf $f_k(x)$ that generates the m-dimensional random vector X when it comes from class k. In practice, $f_k(x)$ must be estimated from training data which, as m increases, becomes a hard problem unless strong assumptions are made, such as multivariate normality. Even with multivariate normality there are $m(m+1)/2$ parameters to be estimated in

the variance matrix Σ, and for large m the data may be insufficient to get good estimates. Sometimes Σ is assumed to be diagonal, so that the components of X become independent. The resulting Bayesian classification procedure is then called *näive Bayes*, which is fast and sometimes effective but it excludes potentially important correlation among the components of X. In general, as the match of the estimated pdfs to the variation in the data deteriorates, the performance of any Bayes classifier may decline. This leads to the problem of designing alternative methods of classification. We describe two popular approaches.

When the data vector satisfies $X \sim N(\mu_k, \Sigma)$ for each class C_k, with $k = 1, \ldots, K$, Eqs. (17.23) and (17.24) give the form of the Bayes classifier in terms of the linear discriminant function. Let us consider, first, the case of binary classification, where $k = 1, 2$. Examining (17.23) and (17.24), if we combine the terms that do not depend on x we may write (17.23) in the alternative form

$$\log \frac{P(C = C_1|X = x)}{P(C = C_2|X = x)} = \alpha_0 + x^T \alpha$$

where $\alpha = \Sigma^{-1}(\mu_1 - \mu_2)$ is an m-dimensional vector. Because, in this binary case, $P(C = C_2|X = x) = 1 - P(C = C_1|X = x)$, we have

$$\log \frac{P(C = C_1|X = x)}{1 - P(C = C_1|X = x)} = \alpha_0 + x^T \alpha. \tag{17.25}$$

Equation (17.25) puts the linear discriminant function in the form of a logistic regression model for binary data, as given by Eq. (14.6), i.e., we could rewrite (17.25) as

$$\log \frac{P(C = C_1|X = x)}{1 - P(C = C_1|X = x)} = \beta_0 + x^T \beta \tag{17.26}$$

and this suggests solving the binary classification problem using logistic regression. More specifically, given training data, the parameters β_0 and β may be estimated using logistic regression applied to the training data to get ML estimates $\hat{\beta}_0$ and $\hat{\beta}$ (as outlined in Section 14.1.2) and then observations may be classified by replacing β_0 and β with $\hat{\beta}_0$ and $\hat{\beta}$ in (17.26) and then assigning an observation to class 1 whenever the function in (17.26) is positive. This method is called a *logistic regression classifier*. The method may be extended to multiple classes using a multi-category generalization of logistic regression, often called *polytomous regression* or *multinomial logistic regression*.

The model in (17.25) looks the same as the model in (17.26) but according to Section 17.4.1, in applying LDA using (17.25) we would estimate the parameters using the sample means and pooled variance matrix. On the other hand, logistic regression would estimate the parameters using maximum likelihood, which is different. The distinction is that logistic regression does not make the assumption of multivariate normality and, instead, treats the x values as fixed.

The general wisdom is that logistic regression classifiers often perform similarly to LDA classifiers. See Hastie et al. (2009) for additional discussion. Although the form of the right-hand side of (17.26) is linear, logistic regression can accommodate complicated nonlinear relationships using the methods discussed in Chapter 15.

A different idea lies behind the *support vector machine* (*SVM*) *classifier*, which we explain briefly by first describing the *perceptron neural network* model. A perceptron model is a function that takes a set of input variables $x_1, \ldots, x_m$ and performs a linear computation followed by binary thresholding:

$$\begin{aligned} \nu &= \phi(u) \\ u &= \left(\sum_{i=1}^{m} w_i x_i \right) - b \end{aligned} \tag{17.27}$$

where $w_1, \ldots, w_m$ are a set of weights associated with that specific perceptron, and $\phi(u) = 1$ when $u \geq 0$ and $\phi(u) = -1$ when $u < 0$. This is a binary classifier in the sense that a vector $x = (x_1, \ldots, x_m)$ is put into class 1 when $\phi(x) = 1$ and into class 2 when $\phi(x) = -1$.

Let us now consider the performance of the perceptron classifier when the data may be separated cleanly into two classes.

Suppose w is an m-dimensional vector. The set $\{x \in R^m : \langle x, w \rangle = 0\}$ is the $(m-1)$-dimensional plane perpendicular to the vector w. It separates two halves of R^m, namely the sets $\{x \in R^m : \langle x, w \rangle > 0\}$ and $\{x \in R^m : \langle x, w \rangle < 0\}$. It is thus called a *separating hyperplane*. The hyperplane $S_0 = \{x \in R^m : \langle x, w \rangle = 0\}$ passes through the origin, i.e., the m-dimensional 0 vector is in this hyperplane ($0 \in S_0$). If $v \in R^m$ we can define $S_v = v + S_0$ to be the set of all vectors in S_0 added to v. This S_v is another separating hyperplane: it may be written

$$S_v = \{x \in R^m : \langle x - v, w \rangle = 0\} = \{x \in R^m : \langle x, w \rangle = b\}$$

where $b = \langle v, w \rangle$ and it separates the sets $\{x \in R^m : \langle x, w \rangle > b\}$ and $\{x \in R^m : \langle x, w \rangle < b\}$.

The separating hyperplane concept applies to data when one set of data vectors lies in a set $\{x \in R^m : \langle x, w \rangle > b\}$ and another set of data lies in a set $\{x \in R^m : \langle x, w \rangle < b\}$. See Fig. 17.5. If two such sets of data come from two distinct classes, then the classifier defined by (17.27) would perfectly classify such data.

The original *perceptron learning rule* attempted to estimate or "learn" the weights $w_1, \ldots, w_m$ from data in order to perform classification. The simple method we have described would be considered ineffective for general-purpose classification, partly because data are not usually perfectly separated in this way and partly because there is not a unique solution: as seen in Fig. 17.5, there are infinitely many separating hyperplanes that fall in the shaded region.

Both of these problems are overcome by classifiers known as *support vector machines* (*SVMs*). Lack of uniqueness is solved by finding the separating hyperplane that maximizes the distance to the closest point in each class. This is found in terms

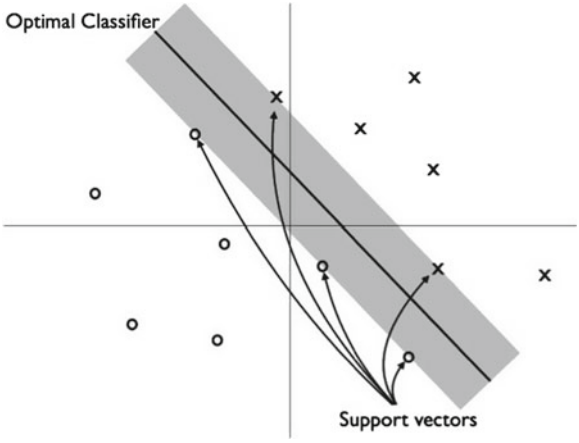

Fig. 17.5 Optimal classification boundary and support vectors for a problem with separable classes. Hypothetical data from two classes are indicated by x and o. The *dark black line* is defined by an optimal classifier that separates the two classes of data. However, any parallel line falling within the gray region would produce the same classification of the given data. The points labeled "support vectors" lie on the boundary of this gray region. The optimal classifier is then determined by maximizing the distance from the separating line to each of the two boundaries of the gray region, which are determined by the support vectors.

of the *support vectors*, which are illustrated in Fig. 17.5. Separation of data vectors is improved by using transformations to higher-dimensional spaces, analogously to what is done in regression when one transforms a single variable x to a polynomial (see Section 12.5.4) or a spline (see Section 15.2). Such transformations take the form $h(x) = (h_1(x), h_2(x), \ldots, h_M(x))$. As the space gets larger, it becomes easier to separate the data vectors from the two classes. One might expect difficulties in implementation, and problems with over-fitting, but there is a so-called *kernel trick* that makes the method[7] practical. It turns out that all of the required computations can be carried out in terms of a *kernel function* $K(u, v)$ that specifies an inner product between m-dimensional vectors u and v,

$$K(u, v) = \langle h(u), h(v) \rangle. \tag{17.28}$$

For example, if we assume $m = 2$, so that $u = (u_1, u_2)$ and $v = (v_1, v_2)$, and we define

$$K(u, v) = (\langle u, v \rangle)^2$$

then (17.28) is satisfied when $h(x)$ (for $x = (x_1, x_2)$) is defined by

$$h(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2).$$

This simplification allows theory and implementation to be developed.

[7] This use of "kernel" is different than that in Section 15.3.1.

Example 17.5 Predicting Reading Improvement in Dyslexic Children from fMRI To see whether fMRI or diffusion tensor imaging (DTI) might predict future gains in reading ability among dyslexic children, Hoeft et al. (2011) followed 20 such subjects for 2.5 years. The authors split the subjects into two groups based on their improvement in single-word reading skill across the period of observation (high improvement vs. low improvement). They then applied SVM to whole-brain fMRI, and also DTI, to see whether these imaging modalities could be used to predict outcome. They reported 92 % classification accuracy from leave-one-out cross-validation, based on the fMRI data. □

In many situations SVM classifiers behave similarly to logistic regression classifiers, but they are in principle very flexible and sometimes outperform other methods. See Hastie et al. (2009) for additional discussion.

17.4.3 Multivariate observations may be clustered into groups.

In Section 17.4.1 we showed that when a data vector X in class k satisfies $X \sim N_m(\mu_k, \Sigma)$, the Bayes classifier takes the simple form of linear discriminant analysis, given in (17.23) and (17.24). Under the multivariate normality assumption, together with homogeneity of the variance matrices, linear discriminant analysis solves the problem of optimally assigning observations to classes. This, however, requires that the class parameters are known—or that they can be estimated from training data and then treated as known. Estimating parameters from training data is an instance of *supervised learning* because the knowledge of class membership in the training data could be considered a form of supervision. The corresponding *unsupervised* problem of putting data into classes with no prior knowledge of class structure is called *clustering*, and the resulting empirically-defined classes are called *clusters*. We provided an illustration of clustering in Example 17.4 on p. 502.

To discuss the problem in generality, let us assume there are K classes, that X is drawn from class k with probability π_k, and that, conditionally on X being drawn from class k, X follows an m-dimensional multivariate normal distribution with mean μ_k and variance matrix Σ_k. We could write this latter statement as $X|C = k \sim N_m(\mu_k, \Sigma_k)$. We then have a two-stage distribution for X, the first stage involving the distribution of class membership C and the second stage involving the multivariate normal distribution. Taking account of both of these, the marginal distribution of X (after marginalizing over the distribution of C) has pdf found by averaging over C:

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x; \mu_k, \Sigma_k) \tag{17.29}$$

where $f_k(x; \mu_k, \Sigma_k)$ is the $N_m(\mu_k, \Sigma_k)$ pdf given by (5.17). This is a *mixture model* in the sense that the K multivariate normal distributions are "mixed" according to

the prior probabilities $\pi_1, \ldots, \pi_K$. The distribution defined by (17.29) is a *mixture of Gaussians model*, as in the illustration in Section 8.4.5. *Mixture of Gaussians clustering* applies ML estimation to the collection of observations x to estimate the parameters $\mu_1, \ldots, \mu_K$, $\Sigma_1, \ldots, \Sigma_K$ and the prior probabilities $\pi_1, \ldots, \pi_K$, and then uses the resulting Bayes classifier to assign each observation x to a cluster. As discussed in Section 8.4.5, ML estimation in mixture of Gaussian models is often implemented using the EM algorithm. Because, in practice, the number of clusters is not known in advance, the model is typically fitted for several different values of K and then a model selection procedure such as AIC or BIC is used (see Section 12.5.7).

In mixture of Gaussians clustering the variance matrices $\Sigma_1, \ldots, \Sigma_K$ must be estimated from the data, and sometimes the data are too sparse to get good estimates of the many variance and covariance parameters. In this case the variance matrices are often assumed[8] equal, $\Sigma_1 = \cdots = \Sigma_K$. A more extreme assumption is to take $\Sigma_1 = \cdots = \Sigma_K = \sigma^2 I_m$ for some σ, i.e., to assume all the variance matrices are equal to a multiple of the m-dimensional identity matrix. This turns out to be closely related to another method, known as K-means clustering.

In K-means clustering it is assumed there are K clusters, with the kth cluster having a mean μ_k. The idea is to put the data vector x into the cluster having its mean closest to x. Thus, after the procedure is applied, so that the clusters are determined and the means μ_k are fixed (by setting them equal to estimated values), every data vector x in cluster j will satisfy

$$||x - \mu_j|| = \min_{k=1,\ldots,K} ||x - \mu_k||. \tag{17.30}$$

However, initially the clusters are not known. They are determined iteratively. After an arbitrary initialization that assigns each data vector to one of K clusters, the following steps are iterated:

1. For $k = 1, \ldots, K$, the mean vectors μ_k is set equal to the sample mean $\bar{x}^k$ of the vectors assigned to cluster k;
2. Each x is assigned to the cluster that minimizes distance as in (17.30).

At each iteration, this algorithm will reduce the sum of squared distances $||x - \mu_j||^2$, summed over all data values, with μ_j being the mean of the cluster to which x is assigned. The algorithm converges to a local minimum of the sum of squared distances (it may not be the global minimum).

Example 17.4 (continued from p. 502) The three clusters in the bottom panel of Fig. 17.3 were identified by K-means clustering (here, with $K = 3$). Three boundary lines are also drawn in Fig. 17.3. Each line is equally distant from the sample means in two of the clusters. □

The relationship of K-means clustering to mixture-of-Gaussian clustering is spelled out in many sources (e.g., Hastie et al. 2009). If it is assumed that $\Sigma_1 =$

[8] Each matrix Σ_k has $m(m+1)/2$ parameters so there are $Km(m+1)/2$ parameters when the matrices are allowed to be different and only $m(m+1)/2$ if they are assumed to be equal.

$\cdots = \Sigma_K = \sigma^2 I_m$ for some σ, and we write the ith data vector as x^i, for $i = 1, \ldots, n$, then the maximum likelihood estimate of μ_k in the mixture-of-Gaussians model, for $k = 1, \ldots, K$, is given by

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n} \gamma_{ik} x^i}{\sum_{i=1}^{n} \gamma_{ik}} \tag{17.31}$$

where γ_{ik} is the posterior probability that observation x^i is in class k (see Eq. (8.48)), and is estimated from the data (see p. 217). This is not the same estimate as the sample mean $\bar{x}^k$ over the observations within cluster k. However, when the posterior probabilities become close to 0 and 1 we get

$$\hat{\mu}_k \approx \bar{x}^k.$$

This occurs when the data form highly distinct clusters or, equivalently, when σ is close to 0 relative to the distance between the means of the clusters.

Chapter 18
Time Series

18.1 Introduction

In the analysis of neural data, time is important. We experience life as evolving, and neurophysiological investigations focus increasingly on dynamic features of brain activity. If we wish to understand the signals produced by nervous system processes we must use an analytical framework that is built for time-varying observations.

From a mathematical point of view, time is a number with an arbitrarily-chosen origin, the value $t = 0$ typically representing an experimental or behavioral marker such as the onset of a visual cue. We may work backward in time by taking t to be negative. Although measurements are always made with some resolution of temporal accuracy, often determined by a sampling rate (such as 20 KHz, giving a precision of $\Delta t = .05$ ms), mathematically we allow t to be any real number, such as $t = \frac{\pi}{2}$ s. When measurements depend on time we may think of them as functions of time, as in $y = f(t)$, and when we acknowledge that the measurements are noisy we might write

$$Y = f(t) + \varepsilon$$

where ε is a random variable representing noise and Y is written as a capital letter to emphasize that it, too, is a random variable. Given n observation pairs $(t_1, y_1), \ldots, (t_n, y_n)$ we might write

$$Y_i = f(t_i) + \varepsilon_i, \tag{18.1}$$

and this returns us to the usual nonparametric regression model of Chapter 15, in which the variables $\varepsilon_1, \ldots, \varepsilon_n$ are assumed independent. While at first glance (18.1) may seem natural, this kind of formulation does not yet go far enough in dealing with measurements that vary across time because it does not take account of the sequential nature of the argument t. In (18.1) the values $i = 1, 2, \ldots, n$ are generally no longer arbitrary labels but rather important and meaningful indications of temporal ordering with $t_1 < t_2 < \cdots < t_n$. If time matters, then even the noise variables $\varepsilon_1, \ldots, \varepsilon_n$ may

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_18,

be related to one another, and thus no longer independent. In this case, specialized methods can produce powerful results. The term *time series*, refers both to data collected across time and to the large body of theory and methods for analyzing such data.

Let us switch over to the general notation for random variables and write a theoretical sequence of measurements as $X_1, X_2, \ldots$, and a generic random variable in the sequence as X_t. Another way to say the X_t variables are dependent is that knowing $X_1, X_2, \ldots, X_{t-1}$ should allow us to predict, at least up to some uncertainty, X_t. Predictability plays an important role in time series analysis.

Example 2.2 (continued from p. 27) On p. 27 we displayed several EEG spectrograms taken under different stages of anesthesia. We noted earlier that both the roughly 10 Hz alpha rhythm and the 1–4 Hz delta rhythm are visible in the time series plot. In this scenario we can say a lot about the variation among the EEG values based on their sequence along time: in the time bin at time t the EEG voltage is likely to be close to that at time $t - 1$ and from the voltage in multiple time bins preceding time t we could produce a good prediction of the value at time t. □

The spectrograms in Example 2.2 display the rhythmic, wave-like features of the EEG signals contrasting them across phases of anesthesia. They do so by decomposing the signal into components of various frequencies, using one of the chief techniques of time series analysis. The decompositions are possible in this context because the EEGs may be described with relatively simple and standard time series models, but this is not true of all time series. The EEG series are, in a sense, very special because their variation occurs on a time scale that is substantially smaller than the observation interval. By contrast, if we go back to Fig. 1.5 of Example 1.6 we see another time series where the variation is on a longer time scale. The EPSC signal drops suddenly, and only once, shortly after the beginning of the series, then recovers slowly throughout the remainder of the series. In other words, the variation in the EPSC takes place on a time scale roughly equal to the length of the observation interval. Another way to put this is that the EEG at time x_t may be predicted reasonably well using only the preceding EEG values $x_{t-1}, x_{t-2}, \ldots, x_{t-h}$, going back h time bins, where h is some fairly small integer, but a prediction of the EPSC at x_t based on earlier observations would require nearly the entire previous series and still might not be very good. The most common time series methods, those we describe here, assume predictability on relatively short time scales.

So far we have said that the EEG at time x_t may be predicted using the preceding EEG values $x_{t-1}, x_{t-2}, \ldots, x_{t-h}$, but we did not specify which value of t we were referring to. Part of the point is that it doesn't much matter. In other words, it is possible to predict almost *any* x_t using the preceding h observations. (We say "almost" any x_t because we have to exclude the first few x_t observations, with $t \leq h$, where there do not exist h preceding observations from which to predict.) Furthermore, the formula we concoct to combine $x_{t-1}, x_{t-2}, \ldots, x_{t-h}$ in order to predict x_t may be chosen independently of t. This is a very strong kind of predictability, one that is stable across time, or *time-invariant*. The notion of time invariance is at the heart of time series analysis.

We now begin to formalize these ideas. Let X_t be the measurement of a series at time t, with $t = 1, \ldots, n$. Let $\mu_t = E(X_t)$ and $\Sigma_{ij} = Cov(X_i, X_j)$. As soon as we contemplate estimation of this mean vector and covariance matrix we are faced with a serious difficulty. For simplicity consider time t and the problem of estimating μ_t and $\sigma_t^2 = \Sigma_{tt}$. If we have many replications of the measurements at time t (as is usually the case, for example, with evoked potentials) we can collect all the observations across replications at time t and compute their sample mean and sample variance. However, if we have only one time series, and therefore one observation at t, we do not have a sample from which to compute the sample mean and variance. The only way to apply any kind of averaging is by using observations at other values of time. Thus, we can only get meaningful estimates of mean and covariance by making assumptions about the way X_t varies across time. Let us introduce a theoretical time series, or *discrete-time stochastic process* $\{X_t; t \in \mathcal{Z}\}$, $\mathcal{Z}$ being the set of all integers. We are now in a position to define the kinds of time invariance we will need. We say that the series X_t is *strictly stationary* if it is time-invariant in the sense that the joint distribution of each set of variables $\{X_t, X_{t+1}, \ldots, X_{t+h}\}$ is the same as that of the variables $\{X_s, X_{s+1}, \ldots, X_{s+h}\}$ for all t, s, h. Because the time index takes all possible integer values it is an abstraction (no experiment runs indefinitely far into the past and future) but it is an extremely useful one. A standard notation in the time series context is $\gamma(s, t) = \Sigma_{st}$. The function $\gamma(s, t)$ is called the *autocovariance function* and the *autocorrelation function* (ACF) is defined by

$$\rho(s, t) = \frac{\gamma(s, t)}{\sqrt{\gamma(s, s)\gamma(t, t)}}.$$

The prefix "auto," which signifies here that we are considering dependence of the time series on itself, is a hint that one might instead consider dependence across multiple time series, where we would instead have "cross-covariance" and "cross-correlation" functions (which we discuss in Section 18.5). A time series is said to be *weakly stationary* or *covariance stationary* if (i) μ_t is constant for all t and (ii) $\gamma(s, t)$ depends on s and t only through the magnitude of their difference $|s-t|$. This weaker sense of stationarity is all that is needed for many theoretical arguments. Under either form of stationarity we follow the convention of writing the autocovariance function in terms of a single argument, $h = t - s$, in the form $\gamma(h) = \gamma(t - h, t)$. Note that $\gamma(0) = V(X_t)$. It is not hard to show that $\gamma(0) \geq |\gamma(h)|$ for all h, and $\gamma(h) = \gamma(-h)$. In the stationary case the autocorrelation function becomes

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}. \tag{18.2}$$

Illustration: The 3-point moving average process

$$X_t = \frac{1}{3}(U_t + U_{t-1} + U_{t-2})$$

where the U_t variables are independent, with $E(U_t) = 0$ and $V(U_i) = \sigma_U^2$, is a stationary process with autocovariance and autocorrelation

$$\begin{aligned}
\gamma(0) &= \frac{\sigma_U^2}{3} \\
\gamma(\pm 1) &= \frac{2\sigma_U^2}{9} \\
\rho(\pm 1) &= \frac{2}{3} \\
\gamma(\pm 2) &= \frac{\sigma_U^2}{9} \\
\rho(\pm 2) &= \frac{1}{3} \\
\gamma(\pm h) &= \rho(h) = 0, \text{ for } |h| \geq 3.
\end{aligned}$$

□

Having defined what it means for a process to be stationary, and also having defined the autocorrelation function, let us return to the distinction we were trying to draw between the EEG and EPSC time series. The EEG series may be modeled as stationary, and furthermore its variation is consistent with what is called *short-range dependence*. A theoretical time series exhibits short-range dependence when its correlation function $\rho(h)$ vanishes quickly as h becomes infinite. For the most common time series models the correlation function vanishes exponentially fast (i.e., there is a positive number a for which $\rho(h)e^{a|h|} \to 0$ as $h \to \pm\infty$). On the other hand, it is questionable whether one would want to model the EPSC time series as stationary and, if so, it would be necessary to use a model that assumes long-range dependence, where the correlation function dies out slowly as h becomes infinite. Time series analysis is concerned with variation across time while being cognizant of the role of stationarity. Much time series theory explicitly assumes stationarity. There is also considerable interest in non-stationary series, but the theoretical developments involve particular kinds of non-stationarity or modifications of methods that apply to stationary series. In contrast, nonparametric regression does not consider time-invariance arguments at all. In (18.1) the usual nonparametric assumption is $E(\varepsilon_t) = 0$, and we have $\mu_t = E(Y_t) = f(t)$. In other words, instead of a constant mean required by stationarity, the nonparametric problem focuses on the evolution of the mean as a function of time. In fact, many investigations involve a mix of these two possibilities: there is a stimulus that produces a time-varying mean component of the response, but there is also a wave-like time-invariant component of the response. From a practical point of view, it is very important to consider these components separately.

Example 15.2 (continued) For illustrative purposes we analyze here a small record of an LFP, which was recorded for 30 s (seconds) and sampled at 1 KHz as part of the experiment described briefly on p. 421. We confine our attention to the first second and the last second (each consisting of 1,000 observations), and will consider whether the signal appears consistent across these two time periods in the sense of containing

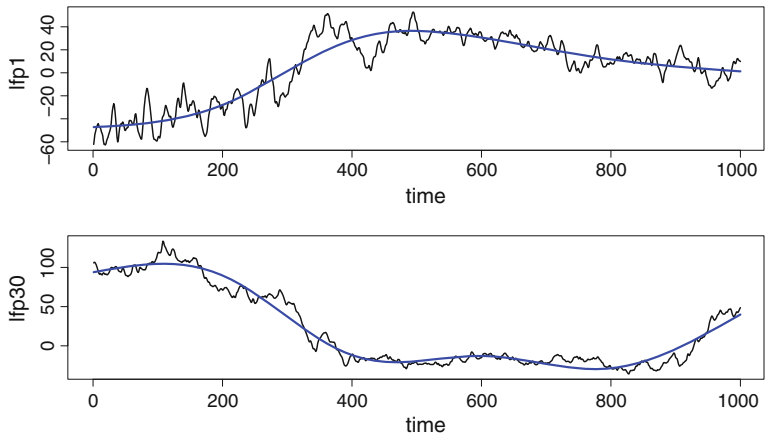

Fig. 18.1 LFP and smoothed versions representing slowly-varying trends. *Top* First second of average LFP. *Bottom* Last (thirtieth) second of average LFP. Smoothing was performed using regression splines with a small number of knots, as described on p. 421.

the same delta-wave content. Figure 18.1 displays these two time series, together with smoothed versions of the average LFP in these two periods. When we focus on a single second of observation time, the slow-wave activity shows up as slowly-varying mean signals, or trends, represented by the smoothed versions of the two LFP traces in the figure. Even though the slowly-varying trends could be considered roughly oscillatory on a longer time scale, at this time scale they can not be represented as oscillatory and are, instead, sources of long-range dependence or non-stationarity akin to that in Fig. 1.5. In order to capture the higher-frequency, stationary activity in these plots (with short-range dependence) we must first remove the slow trends. We analyze these data further in subsequent sections. □

In motivating stationarity we brought up the problem of estimating the mean and covariance functions, pointing out that in the absence of replications some assumptions must be made. Under stationarity the value of the constant mean $\mu_t = \mu$ may be estimated by the sample mean and an obvious estimator of the autocovariance function is the *sample autocovariance function*

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h} - \overline{x})(x_t - \overline{x}) \tag{18.3}$$

for $h = 0, 1, \ldots, n-1$ and then $\hat{\gamma}(-h) = \hat{\gamma}(h)$.We then have the *sample autocorrelation function* (sample ACF),

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} \tag{18.4}$$

which is an estimator of the autocorrelation function (18.2).

In this chapter we provide an overview of key concepts in time series analysis. Section 18.2 describes the two major approaches to time series analysis. Section 18.3 gives some details on methods used to decompose time series into frequencies, as in Example 2.2. There are several important subtleties, and we discuss these as well. Section 18.4 discusses assessing uncertainty about frequency components, and Section 18.5 reviews the way these methods are adapted to assess dependence between pairs of simultaneous time series.

18.2 Time Domain and Frequency Domain

In discussing Example 2.2, on p. 514, we alluded to the decomposition of the signal into frequency-based components. In general, time series analysis relies on two complementary classes of methods. As the name indicates, *time domain* methods view the signal as a function of time and use statistical models that describe temporal dependence. *Frequency domain* methods decompose the signal into frequency-based components, and describe the relative contribution of these in making up the signal. In this section we provide a brief introduction to these two approaches, starting with frequency-based analysis. Here are two examples.

Example 18.1 Gamma oscillations in MEG during learning Cortical oscillatory activity in the gamma band (roughly 30–120 Hz) has been associated with many cognitive functions. Chaumon et al. (2009) used MEG imaging to investigate the role of gamma oscillations during unconscious learning. They used a paradigm in which subjects were to find the letter "T" within a set of distractors and determine its orientation. On some trials, which they called "predictive," the distractors were repeated and the location of the "T" remained the same. On other trials, which they called "nonpredictive," the distractors changed configurations and the location of the "T" changed. The subjects were shown many blocks containing 12 trials of each type. Although they remained unaware of the information provided by the configuration type, their reaction time decreased faster across blocks for the predictive trials than for the nonpredictive trials. The authors were interested in whether this unconscious learning was associated with changes in gamma band activity recorded with MEG. □

Example 18.2 fMRI BOLD signal and neural activity To investigate the neural basis of the fMRI BOLD signal, Logothetis et al. (2001) recorded local field potential (LFP) and multi-unit activity (MUA) together with fMRI from a region in primary visual cortex across 29 experimental sessions using 10 macaque monkeys. The stimulus involved rotating checkerboard patterns. In examining the relationship between LFP and BOLD, the authors focused on gamma band activity from 40 to 130 Hz. □

We now introduce another example, which we will use repeatedly in several parts of this chapter to demonstrate analytical techniques.

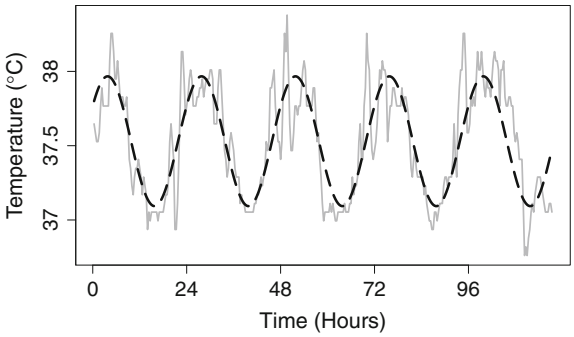

Fig. 18.2 Core temperature on a human subject, recordings taken every 20 min; *y*-axis in units of degrees Celsius (data shown with a *solid line*). Overlaid on the data is the least-squares fit of a cosine (shown with a *dashed line*), having a period of 24 h (hours).

Example 18.3 The circadian rhythm in core temperature Human physiology, like that of other organisms, has adapted to the cycle of changing environmental conditions, and resulting levels of activity, across each day and night. The result is a clear day/night pattern in hormone levels in the blood, and other indicators of the body's attempt to maintain homeostasis. In a study of methodology used to characterize circadian rhythms, Greenhouse et al. (1987) analyzed core temperatures of a human subject measured every 20 min across several days. Figure 18.2 displays the data. There is an obvious daily cycle in the temperatures. Figure 18.2 also shows a cosine curve, with a 24 h period, that has been fitted to the data using ordinary least-squares regression. □

The cosine curve in Fig. 18.2 was obtained by applying linear regression. We discussed fitting a cosine curve previously, in Example 12.6, in the context of directional tuning. Here, we begin with a cosine function $\cos(2\pi\omega_1 t)$, where ω_1 is the frequency (in cycles per unit time), then introduce an amplitude R_{amp}, an offset average value μ_{avg}, and a phase ϕ to put it in the functional form

$$f(t) = \mu_{avg} + R_{amp}\cos(2\pi(\omega_1 t - \phi)). \tag{18.5}$$

Details: The function $R_{amp}\cos(2\pi\omega_1 t)$ varies between a minimum of $-R_{amp}$ and a maximum of R_{amp}, and its average on $[0, 1]$ is 0. Adding the constant μ_{avg} makes the cosine oscillate around μ_{avg} with minimum $\mu_{avg} - R_{amp}$ and maximum $\mu_{avg} + R_{amp}$. It is also worth mentioning that the regression in Example 12.6 was set up slightly differently because the explanatory variable of interest was not time but rather the angle $\theta = 2\pi(\omega t - \phi)$. □

Based on (18.5) the statistical model for observations $y_1, \ldots, y_n$ at time points $t_1, \ldots, t_n$ is then

$$Y_i = \mu_{avg} + R_{amp} \cos(2\pi(\omega_1 t_i - \phi)) + \varepsilon_i$$

where, for the core temperature data, $\omega_1 = 1/72$ cycles per 20 min is the frequency corresponding to 1 cycle per day (a 24 h period). To simplify fitting, this model may be converted to a linear form, i.e., a form that is linear in the unknown parameters. Using

$$\cos(u - v) = \cos u \cos v + \sin u \sin v \tag{18.6}$$

with $u = 2\pi\omega_1 t_i$ and $v = 2\pi\phi$ we have

$$R_{amp} \cos(2\pi(\omega_1 t_i - \phi)) = A \cos(2\pi\omega_1 t_i) + B \sin(2\pi\omega_1 t_i) \tag{18.7}$$

where $A = R_{amp} \cos(2\pi\phi)$ and $B = R_{amp} \sin(2\pi\phi)$. We may therefore rewrite the statistical model as

$$Y_i = \mu_{avg} + A \cos(2\pi\omega_1 t_i) + B \sin(2\pi\omega_1 t_i) + \varepsilon_i, \tag{18.8}$$

which has the form of a linear regression model, and may be fitted using ordinary linear regression. Specifically, we do the following:

1. Assume the data $(t_1, \ldots, t_n)$ and $(y_1, \ldots, y_n)$ are in respective variables `time` and `temp`.
2. Define

$$\texttt{cosine} = cos(2\pi\texttt{time}/72)$$
$$\texttt{sine} = sin(2\pi\texttt{time}/72).$$

3. Regress `temp` on `cosine` and `sine`.

For future reference we note that the squared amplitude of the cosine function in (18.7) is

$$R_{amp}^2 = A^2 + B^2 \tag{18.9}$$

and the phase is

$$\phi = \frac{1}{2\pi} \arctan(\frac{B}{A}). \tag{18.10}$$

In the core temperature data of Example 18.3 there is a clear, dominant periodicity, which is easily described by a cosine function using linear regression. We may do a bit better if we allow the fitted curve to flatten out a little, compared to the cosine function. This is accomplished by introducing a second frequency, $\omega_2 = 2\omega_1$ to produce the model

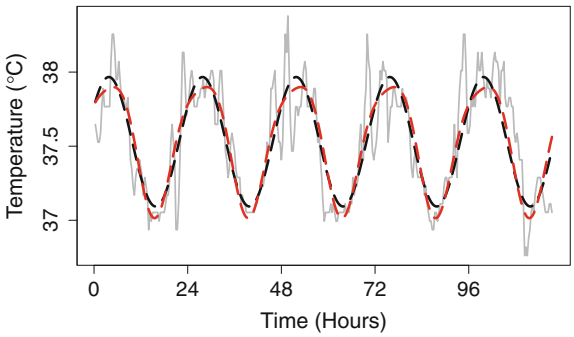

Fig. 18.3 Plot of core temperature, as in Fig. 18.2, together with fit of (18.8), shown in the *black dashed line*, using the fundamental frequency $\omega_1 = 1/72$ (one oscillation every 72 data points, i.e., every 24 h), and fit of (18.11), shown in *red dashed line*. The latter improves the fit somewhat in the peaks and troughs.

$$Y_i = \mu_{avg} + A_1 \cos(2\pi\omega_1 t_i) + B_1 \sin(2\pi\omega_1 t_i) + A_2 \cos(2\pi\omega_2 t_i) + B_2 \sin(2\pi\omega_2 t_i) + \varepsilon_i. \tag{18.11}$$

Example 18.3 (continued from p. 519) Least-squares regression using model (18.11) yields a highly significant effect for the second cosine–sine pair ($p < 10^{-6}$) and Fig. 18.3 displays a modest improvement in fit. □

Model (18.8) was modified in (18.11) by introducing the additiona1 cosine–sine pair corresponding to the frequency ω_2. In principle this process could be continued by introducing frequencies of the form $\omega_k = k\omega_1$ for $k = 3, 4, \ldots$. Here, ω_1 is called the *fundamental frequency*, the additional frequencies ω_k are *harmonic frequencies*, and the resulting regression model is often called *harmonic regression*. For the core temperature data it turns out that $k = 2$ is a satisfactory choice (see Greenhouse et al. 1987) but, in general, one might use linear regression to fit many harmonics and ask how much variation in the data is explained by each cosine–sine pair. For this purpose one might use contributions to R^2, which is the germ of the idea behind one of the main topics in time series, *spectral analysis*. Spectral analysis can be a very effective way to describe wave-like behavior, as seen in the EEG signals of Example 2.2.

18.2.1 *Fourier analysis is one of the great achievements of mathematical science.*

Spectral analysis, otherwise known as Fourier analysis,[1] decomposes an oscillatory signal into trigonometric components. Because many physical phenomena may be described by applying this technique (and it is at the heart of quantum mechanics), the physicist Richard Feynman called[2] the ability to create such decompositions "probably the most far-reaching principle in mathematical physics." From a practical point of view, our world has been changed dramatically by applications of Fourier analysis.

The argument may be broken into several steps.

1. The signal may be represented by a smoothly varying function $f(t)$, for values of t (usually thought of as time) in a suitable interval $[a, b]$, which, for convenience, we may take[3] to be $[0, 1]$.
2. If we pick n values of t spaced evenly across the interval, say, $t_1, t_2, \ldots, t_n$, then $f(t)$ may be determined to a close approximation by its values at these points, i.e., by $f(t_1), f(t_2), \ldots, f(t_n)$, for sufficiently large n. That is, if $f(t)$ varies smoothly then, for practical purposes, interpolation will suffice to reproduce it from its values $f(t_1), f(t_2), \ldots, f(t_n)$.
3. The cosine and sine functions $\cos(2\pi t)$ and $\sin(2\pi t)$ are periodic, completing a single cycle on $[0, 1]$, and thus having frequency 1 (per unit time). This is the fundamental frequency and the corresponding harmonic frequencies are $2, 3, 4, \ldots$. The cosine and sine functions at harmonic frequencies may be considered primitive functions—meaning building blocks of other functions—on $[0, 1]$. When we evaluate a sufficiently large number of primitive functions at $t_1, t_2, \ldots, t_n$, and take linear combinations of them, we are able to reproduce $f(t)$ at the values $t_1, t_2, \ldots, t_n$, which, according to step 2, suffices for reconstructing $f(t)$ throughout $[0, 1]$. That is, we can decompose $f(t)$ into harmonic trigonometric components. This has the potential to provide the appealing interpretation that $f(t)$ is "made up" of particular harmonic components in particular amounts, according to the linear combinations.
4. In order to have this interpretation make sense, the "particular amount" of each component given by the decomposition in step 3 must not depend on the number of components being considered, for that would make the interpretation self-contradictory. In non-orthogonal decompositions the amount, or weight, given to a particular component *does* depend on the other components being considered, but for orthogonal decompositions it does not. (See the discussion in Chapter 12,

[1] The term "spectral analysis" sometimes connotes statistical analysis, rather than purely mathematical analysis, but for now we are ignoring any noise considerations.

[2] Feynman et al. (1963 Volume I, p. 49–1).

[3] The argument we sketch here makes the most sense for functions that are periodic on $[0, 1]$, meaning that they satisfy $f(0) = f(1)$. In Section 18.3.6 we discuss what happens when this condition fails to hold.

p. 351.) Harmonic trigonometric functions are orthogonal, so the interpretation is internally consistent.

These steps all involved major conceptual breakthroughs for mathematics.[4] Taken together they suggest that a signal represented by a smoothly varying function $f(t)$ may be decomposed into cosine and sine harmonic components. This is what Fourier analysis accomplishes.

To be a little more specific, suppose that $f(t)$ is a function on the interval [0, 1] and let us consider time points $t_j = \frac{j}{n}$ for $j = 1, 2, \ldots, n$ where, for simplicity, we assume n is odd so that $(n-1)/2$ is an integer. If we evaluate $f(t)$ at the time points t_j we get an n-dimensional vector

$$y = (f(t_1), f(t_2), \ldots, f(t_n))^T. \tag{18.12}$$

Now define the harmonic trigonometric functions $f_k(t) = \cos(2\pi kt)$ and $g_k(t) = \sin(2\pi kt)$, for $k = 1, 2, \ldots, (n-1)/2$. By evaluating these functions at $t_1, t_2, \ldots, t_n$ we form vectors $f_k = (f_k(t_1), f_k(t_2), \ldots, f_k(t_n))^T$ and $g_k = (g_k(t_1), g_k(t_2), \ldots, g_k(t_n))^T$ and, it turns out, the collection of vectors

$$1_{vec}, f_1, \ldots, f_{(n-1)/2}, g_1, \ldots, g_{(n-1)/2}$$

are orthogonal, where $1_{vec} = (1, 1, \ldots, 1)^T$. (This follows from straightforward algebraic manipulation, together with properties of sines and cosines, see Bloomfield 2000). They therefore form an orthogonal basis for R^n (see Section A.9), which means that any vector y, such as in (18.12), may be written in the form

$$\begin{aligned} y = \mu_{avg} 1_{vec} + A_1 f_1 + \cdots + A_{(n-1)/2} f_{(n-1)/2} \\ + B_1 g_1 + \cdots + B_{(n-1)/2} g_{(n-1)/2}. \end{aligned} \tag{18.13}$$

If we define

$$\begin{aligned} p_n(t) = \mu_{avg} + A_1 f_1(t) + \cdots + A_{(n-1)/2} f_{(n-1)/2}(t) \\ + B_1 g_1(t) + \cdots + B_{(n-1)/2} g_{(n-1)/2}(t) \end{aligned} \tag{18.14}$$

[4] The first requires the notion of function, which emerged roughly in the 1700s, especially in the work of Euler (the notation $f(x)$ apparently being introduced in 1735). The second may be considered intuitively obvious, but a detailed rigorous understanding of the situation did not come until the 1800s, particularly in the work of Cauchy (represented by a publication in 1821) and Weierstrass (in 1872). The notion of harmonics was one of the greatest discoveries of antiquity, and is associated with Pythagoras. The third and fourth steps emerged in work by D'Alembert in the mid-1700s, and by Fourier in 1807. Along the way, representations using complex numbers were used by Euler (his famous formula, given below, appeared in 1748), but they were considered quite mysterious until their geometric interpretation was given by Wessel, Argand, and Gauss, the latter in an influential 1832 exposition. A complete understanding of basic Fourier analysis was achieved by the early 1900s with the development of the Lebesgue integral. Recommended general discussions may be found in Courant and Robbins (1996), Lanczos (1966), and Hawkins (2001).

then we have

$$f(t) = p_n(t) \tag{18.15}$$

for $t = t_j$ for $j = 1, \ldots, n$ and, by interpolation we get the approximation

$$f(t) \approx p_n(t), \tag{18.16}$$

for all $t \in [0, 1]$, which may be considered a decomposition of $f(t)$ into trigonometric components based on the n data values $f(t_1), f(t_2), \ldots, f(t_n)$. The constants $\mu_{avg}, A_1, \ldots, A_k, B_1, \ldots, B_k$ are called the *Fourier coefficients* of $f(t)$. By analogy with the approximate representation of functions by polynomials, the expression $p_n(t)$ in (18.14) is often called a *trigonometric polynomial*. With reference to (18.7), we may say that $A_k f_k$ and $B_k g_k$ together determine the component of $f(t)$ having frequency k.

We now consider the magnitude of y. Using the orthogonality of the component vectors, Eq. (18.13) gives

$$\begin{aligned} ||y||^2 = ||\mu_{avg} 1_{vec}||^2 + ||A_1 f_1||^2 + \cdots + ||A_{(n-1)/2} f_{(n-1)/2}||^2 \\ + ||B_1 g_1||^2 + \cdots + ||B_{(n-1)/2} g_{(n-1)/2}||^2 \end{aligned}$$

and re-writing this we get

$$||y||^2 = ||\mu_{avg} 1_{vec}||^2 + \sum_{k=1}^{(n-1)/2} ||A_k f_k||^2 + ||B_k g_k||^2. \tag{18.17}$$

Equation (18.17) decomposes the squared magnitude of y into magnitudes corresponding to its trigonometric components. Using (18.15) we say that any vector of function evaluations may be written in terms of the trigonometric basis vectors, and its squared length is equal to the sum of squares of its trigonometric components. From (18.16) we see that an analogous statement should hold for functions on [0,1].

We can also use (18.17) to give a nice interpretation of the Fourier decomposition in terms of least-squares regression. We begin by considering (18.13) to be a noiseless regression equation. If we regress y on the variables $f_1, \ldots, f_{(n-1)/2}, g_1, \ldots, g_{(n-1)/2}$ we obtain the coefficients $A_1, B_1, \ldots, A_{(n-1)/2}, B_{(n-1)/2}$. Furthermore, because the trigonometric vectors are orthogonal, the coefficient found by regressing y on all the variables $f_1, \ldots, f_{(n-1)/2}, g_1, \ldots, g_{(n-1)/2}$ is the same as the coefficient of f_k (or g_k) in the regression of y on f_k (or g_k) alone. Thus, it makes sense to say that $A_k f_k$ and $B_k g_k$ together uniquely represent the component of y corresponding to frequency k. Because (18.13) provides an exact fit of y, if we regress y on all the variables $f_1, \ldots, f_{(n-1)/2}, g_1, \ldots, g_{(n-1)/2}$ we get $R^2 = 1$. The regression of y on 1_{vec} gives $\mu_{avg} = \bar{y}$ and $\mu_{avg} 1_{vec} = \bar{y} 1_{vec}$ has squared length $n\bar{y}^2$ so that (18.17) may be rewritten in terms of the total sum of squares

$$||y||^2 - n\bar{y}^2 = \sum_{k=1}^{(n-1)/2} ||A_k f_k||^2 + ||B_k g_k||^2$$

and, dividing both sides by $||y||^2 - n\bar{y}^2$ while using $R^2 = 1$ we get

$$R^2 = \sum_{k=1}^{(n-1)/2} R_k^2, \tag{18.18}$$

where

$$R_k^2 = \frac{||A_k f_k||^2 + ||B_k g_k||^2}{||y||^2 - n\bar{y}^2}, \tag{18.19}$$

which is the proportion of variation in y, and therefore $f(t)$, at frequency k. In other words, this trigonometric representation, using sines and cosines at harmonic frequencies, has the wonderful property that it decomposes the variability of the function $f(t)$ into frequency-based components, the magnitudes of which add to the total variation in $f(t)$. The decomposition (18.18) into components (18.19) is the starting point for spectral analysis.

18.2.2 The periodogram is both a scaled representation of contributions to R^2 from harmonic regression and a scaled power function associated with the discrete Fourier transform of a data set.

We now apply to data $x_1, x_2, \ldots, x_n$ the spectral analysis decomposition discussed in Section 18.2.1. We write $y = (x_1, x_2, \ldots, x_n)$ and use (18.13). We may get a rough idea of the relative contributions to the variability in the data due to the harmonic frequency components simply by plotting R_k^2, defined in Eq. (18.19), against the frequency k. A scaled plot of R_k^2 against frequency is known as the *periodogram*, with the precise definition appearing in Eq. (18.25). The periodogram, together with some important modifications of it, is enormously useful in practice.

Example 18.3 (continued from 521) The periodogram for the core temperature data (introduced on p. 519) is shown in Fig. 18.4. Note the dominant contribution to R^2 corresponding to the roughly daily cycle. □

The coefficients A_k and B_k in (18.13) and (18.19) turn out to be

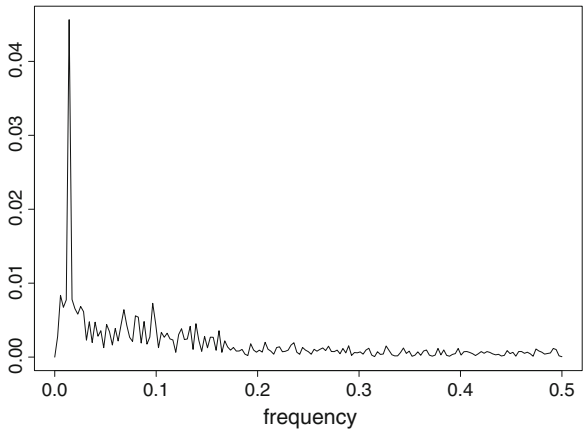

Fig. 18.4 Periodogram of core body temperature data. There is a peak at the frequency representing, very nearly, daily oscillation and this peak is much higher than the remainder of the periodogram.

$$\mu_{avg} = \frac{1}{n}\sum_{j=1}^{n} x_j$$

$$A_k = \frac{2}{n}\sum_{j=1}^{n} x_j \cos(2k\pi j/n) \tag{18.20}$$

$$B_k = \frac{2}{n}\sum_{j=1}^{n} x_j \sin(2k\pi j/n) \tag{18.21}$$

for $k = 1, \ldots, (n-1)/2$. Because the cosine and sine terms always occur in pairs, it is often simpler to represent expressions (18.20) and (18.21) instead in exponential form via Euler's formula,

$$e^{i\theta} = \cos\theta + i\sin\theta, \tag{18.22}$$

which is also Eq. (A.31) in the Appendix. This formula is extremely helpful in Fourier analysis. On the one hand, it provides a kind of "book-keeping" of cosine and sine terms within a complex exponential while, on the other hand, it simplifies many manipulations because multiplication becomes addition of exponents. Applying Euler's formula (18.22), we have

$$\sum_{j=1}^{n} x_j \cos(2k\pi j/n) + i\sum_{j=1}^{n} x_j \sin(2k\pi j/n) = \sum_{j=1}^{n} x_j e^{2k\pi ij/n}$$

and then (18.20) and (18.21) may be replaced with

$$A_k + iB_k = \frac{2}{n}\sum_{j=1}^{n} x_j e^{2\pi ikj/n}$$

for $k = 1, \ldots, (n-1)/2$. By convention the equivalent form

$$A_k - iB_k = \frac{2}{n}\sum_{j=1}^{n} x_j e^{-2\pi ikj/n} \tag{18.23}$$

for $k = 1, \ldots, (n-1)/2$, is used instead. Aside from the multiplier, the right-hand side of (18.23) is the *discrete Fourier transform*. Specifically, for a data sequence $x_1, \ldots, x_n$, we let

$$\omega_j = j/n$$

denote frequency, for $j = 0, \ldots, n-1$. Then the discrete Fourier transform (DFT) is given by

$$d(\omega_j) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} x_t e^{-2\pi i\omega_j t} \tag{18.24}$$

and the periodogram is

$$I(\omega_j) = |d(\omega_j)|^2. \tag{18.25}$$

Here we are interested only in the first $(n-1)/2$ frequencies (if n is odd; otherwise, the first $n/2$ frequencies). From (18.23) we have $d(\omega_j) = \frac{\sqrt{n}}{2}(A_j - iB_j)$, and because $||A_j + iB_j||^2 = A_j^2 + B_j^2$, we get

$$|d(\omega_j)|^2 = \frac{n}{4}(A_j^2 + B_j^2).$$

According to the definition in Eq. (18.19), $A_j^2 + B_j^2$ is proportional to R_j^2 (meaning that the constant multiple does not depend on j) and so we arrive at

$$I(\omega_j) \propto R_j^2,$$

which justifies the interpretation of the periodogram we gave on p. 525. Algorithms for computing the DFT are based on the *fast Fourier transform*, which had a huge impact on signal processing following a 1965 publication of the method by James Cooley and John Tukey. The DFT also has an interpretation using the terminology of signal processing. If we return to the interpretation of $x_1, \ldots, x_n$ as function values $f(t_1), \ldots, f(t_n)$ as in Eq. (18.16), then $||y||^2 = ||(f(t_1), \ldots, f(t_n))||^2$ is

(approximately, by (18.16)), the *power* of the function $f(t)$ on $[0, 1]$ and $I(\omega_j)$ is (approximately[5]) proportional to the power of $f(t)$ at frequency ω_j.

Unfortunately, in spectral analysis, the various notational conventions that get invoked are not consistent across authors. In particular, we have introduced the *Fourier frequencies* $\omega_j = j/n$ for $j = 0, 1, \ldots, n-1$. Because we divided the harmonic integers by n, the Fourier frequencies are restricted to the interval $[0, 1]$. In fact, because we use only the first $(n-1)/2$ frequencies (if n is odd and the first $n/2$ frequencies if n is even) they are restricted to $[0, \frac{1}{2}]$. In some texts $j = 1, \ldots, n$ is used. Furthermore, the multiplier of the complex exponential sum we used in (18.24) to define the DFT is also not universal. For some purposes one must pay attention to the definitions being used by a particular book or piece of software.

It is also important to notice that the Fourier frequencies we have defined on $[0, 1]$ (or $[0, \frac{1}{2}]$) have units of cycles per observation. If the units of time (such as seconds) involve m observations (such as m observations per second) then $m\omega_j$ will be in cycles per unit time. See the legend to Fig. 18.6.

With some additional mathematics, these concepts carry over to infinite-dimensional vector spaces with inner products. The infinite-dimensional representation is analogous: periodic functions (actually, square-integrable periodic functions) form a vector space for which the harmonic trigonometric functions provide an orthogonal basis. The resulting infinite-dimensional harmonic trigonometric expansion is called a Fourier expansion, and the coefficients are the Fourier coefficients.[6] In mathematics, Fourier analysis concerns infinite-dimensional function spaces, but in statistics and engineering these terms are also applied, as here, to the finite-dimensional setting involving data.

The DFT and its inverse are finite versions of the usual Fourier transform and its inverse, which is used extensively in mathematical analysis and signal processing, including theoretical studies of stationary time series. We discuss stationary time series in Section 18.3.1. We also discuss, in the remainder of Section 18.3, several practical issues that arise when using and interpreting the periodogram. We have already mentioned one of these in our discussion of Example 15.2.

Example 15.2 (continued from p. 421) Fig. 18.5 displays the log periodogram for the first second of average LFP, which was plotted previously in the top portion of Fig. 18.1. In Section 18.3.6 we explain why the log transform is used. The point, for now, is that the periodogram does not have a peak corresponding to delta range or other frequencies. This is quite common in series that have slowly varying trends. In contrast, after we remove the trends seen in Fig. 18.1 from the two series (by subtraction, so that the residuals are analyzed instead) the peaks of interest become visible, as seen in Fig. 18.6. □

---

[5] The approximation becomes exact when $f(t)$ is periodic, $f(t)^2$ has a finite integral, and the expansion involves all of the infinitely many harmonics.

[6] With appropriate mathematics (especially the theory of Lebesgue integration) it may be shown that every square-integrable function on [0,1] may be represented, equivalently, by its set of Fourier coefficients, and its integrated squared magnitude is equal to the sum of squares of the coefficients.

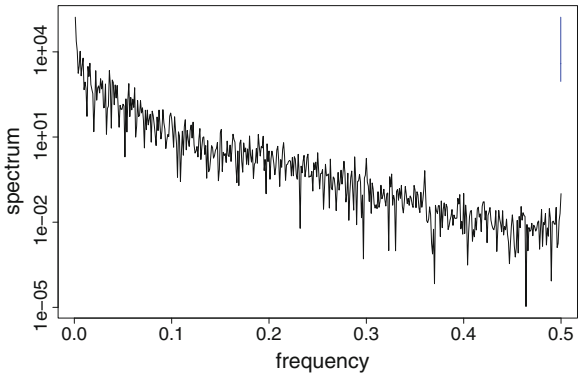

Fig. 18.5 Log periodogram for the first second of average LFP data in Example 15.2.

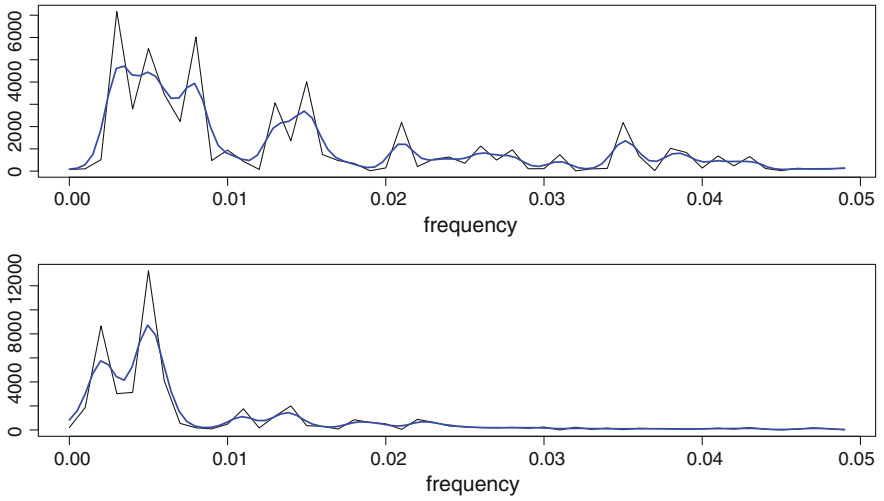

Fig. 18.6 Periodograms and smoothed periodograms from LFP detrended series. *Top* First second of average LFP. *Bottom* Last second of average LFP. Notice that the frequency units are cycles per observation. To get cycles per second (Hz) we must multiply by the number of observations per second, which is 1,000. Thus, the first peak of power in these plots is centered roughly at .005, which corresponds to 5 Hz.

The contrast between Figs. 18.5 and 18.6 illustrates the importance of checking time series for slowly-varying trends, and removing them from the data before performing spectral analysis. This is often called *detrending* the series.

18.2.3 Autoregressive models may be fitted by lagged regression.

As we have indicated, time series are special among kinds of data because of their serial dependence, e.g., the value of X_t is likely to depend on the value of X_{t-1}. The simplest form of dependence is linear dependence, as in the *autoregressive model* given by

$$X_t = \phi X_{t-1} + \epsilon_t.$$

This says that X_t has a regression on X_{t-1}, and otherwise is determined by noise. For consistency with later notation let us write the noise variables as[7] W_t:

$$X_t = \phi X_{t-1} + W_t. \tag{18.26}$$

The natural generalization,

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + W_t, \tag{18.27}$$

is called an *autoregressive model of order p*, written $AR(p)$. The W_t variables are usually assumed to be i.i.d. $N(0, \sigma^2)$. Model (18.26) then becomes the standard $AR(1)$ model. The parameter ϕ in (18.26) is usually assumed to satisfy $|\phi| < 1$, and analogous, but more complicated constraints are assumed for the parameters in (18.27).

> *Some details:* It may be shown that the case of (18.26) with $\phi = 1$, known as a *random walk* model (confer p. 126), is non-stationary. This makes it unsuitable for most auto-regressive modeling methodology. $\phi = -1$ is also non-stationary. The case $|\phi| > 1$ is somewhat more subtle, and it turns out to be non-causal in the sense that X_t depends on W_{t+i} for $i > 0$. The condition $|\phi| < 1$ restricts the $AR(1)$ so that it is neither non-stationary nor non-causal. Additional explanation is provided in time series texts such as Shumway and Stoffer (2006). □

Because the $AR(p)$ model (18.27) has the form of an ordinary linear regression model, we may apply it to data $x = (x_1, \ldots, x_n)$ using ordinary least squares regression after first defining suitable *lagged* variables. In the simplest case, with $p = 1$, we begin by defining a pair of variables y and x_{B1}, each of length $n - 1$:

[7] W is often used to represent time series noise out of deference to Norbert Wiener, a major figure in the development of time series theory.

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \end{pmatrix} = \begin{pmatrix} x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix}$$

$$x_{B1} = \begin{pmatrix} x_{B1,1} \\ x_{B1,2} \\ \vdots \\ x_{B1,n-1} \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \end{pmatrix}.$$

We use the subscript $B1$ for "back 1" because $x_{B1,t} = y_{t-1}$ (x_{B1} "lags" behind y and is often called the lag-1 version of y). We then fit the $AR(1)$ model (18.26) by performing least-squares regression of y on x_{B1}, without using an intercept. The resulting regression coefficient becomes the estimate $\hat{\phi}$ of the $AR(1)$ parameter ϕ.

More generally, to fit an $AR(p)$ model using ordinary least squares we begin by defining $y_{n-p} = x_n$, $y_{n-p-1} = x_{n-1}, \ldots, y_1 = x_{n-p+1}$ and then also defining x_{B1} to be the lag-1 version of y, x_{B2} to be the analogous lag-2 version of y, etc., until we reach x_{Bp}. We then regress y on the variables $x_{B1}, x_{B2}, \ldots, x_{Bp}$.

It is often unclear what order p should be used in the $AR(p)$ model. Sometimes the model selection criteria AIC or BIC are used (see Section 11.1.6). One simple idea is to pick a relatively large value of p, perform the regression, and examine the coefficients from first to last to see when they become non-significant. A similar idea is to use the sample autocorrelation function (ACF), which was defined in (18.4), and the partial autocorrelation function (PACF). Under fairly general conditions, if $X_1, \ldots, X_n$ are i.i.d. with finite variance, and the sample ACF is computed for the random variables X_t, then

$$\sqrt{n}\hat{\rho}(h) \xrightarrow{\mathcal{D}} N(0, 1).$$

Based on this result, the sample ACF is usually plotted together with horizontal lines drawn at $\pm 2/\sqrt{n}$. If the series were i.i.d., then roughly 95 % of the sample autocorrelation coefficients would fall between theses lines. The ACF coefficients outside these lines are considered significant, with $p < .05$, approximately, for large n. This is illustrated for Example 18.3 below.

A difficulty with the sample ACF plot, however, is that it is based on the individual correlations of each lagged variable with the original data. That is, its results come from many single-variable regressions, of y on x_{Bk} for various values of k. A significant regression of y on x_{B2}, for example, could be based on the correlation between x_{B1} and x_{B2} and may reflect a relationship between y and x_{B1}. An alternative is to perform the multiple linear regression of y on *both* x_{B1} and x_{B2} and examine whether the coefficient of x_{B2} is significant, which assesses the explanatory power of x_{B2} after including x_{B1} in the model. The sample PACF at lag h is the sample partial correlation, defined by (5.22), between the time series and itself at lag-h given the lag-1 through lag-$h-1$ series. The lag-h partial autocorrelation coefficient measures

the lag-h correlation after adjusting for the effects of lags 1 through $h-1$, adjusting as in multiple linear regression. It may be computed as the normalized lag-h regression coefficient found from an $AR(h)$ model, normalized by dividing the series by the sample variance $\hat{\gamma}(0)$.

A detail: Suppose X_t is a mean-zero stationary Gaussian series. Then the theoretical PACF is given by $\phi_{11} = Cor(X_t, X_{t+1})$ and for $h \geq 2$,

$$\phi_{hh} = Cor(X_t, X_{t+h} | X_{t+1}, X_{t+2}, \ldots, X_{t+h-1}).$$

More generally, for any mean-zero stationary series let $X_t^{h-1} = \sum_{j=1}^{h-1} \beta_j X_{t-j}$ where the coefficients $\beta_1, \ldots, \beta_{h-1}$ minimize $E((X_t - \sum_{j=1}^{h-1} \alpha_j X_{t-j})^2)$ over the α_js. Then, for $h \geq 2$,

$$\phi_{hh} = Cor(X_t - X_t^{h-1}, X_{t+h} - X_{t+h}^{h-1}). \qquad \square$$

Once again, using large-sample theory, horizontal lines may be drawn on the sample PACF to indicate where the coefficients stop being significant. The sample PACF is often used to choose the order of the autoregressive model.

Example 18.3 (continued from p. 525) Let us consider an $AR(p)$ model for the core temperature residuals following the cosine regression reported on p. 519, and then detrending (using BARS, see Section 15.2.6). We take $p = 22$. The fitted coefficients are plotted in Fig. 18.7. Here is an abbreviated table of coefficients:

Variable	Coefficient	Std. Err.	t-ratio	p-value
x_{B1}	.906	.057	15.9	$< 10^{-15}$
x_{B2}	−.205	.077	−2.7	.008
x_{B3}	−.147	.078	−1.9	.06
x_{B4}	.005	.078	.1	.95
x_{B5}	−0.154	.078	−1.9	.05
x_{B6}	.115	.078	.9	.35
...				
x_{B21}	−.031	.076	−.4	.69
x_{B22}	.011	.057	−.2	.84

Only the first two lagged variables have large t statistics, so it appears that only the first two lagged variables are likely to be helpful in predicting the response variable. Also shown in Fig. 18.7 is the sample ACF, together with horizontal lines drawn at $\pm 2/\sqrt{n}$. The PACF in Fig. 18.7 has nonzero lag-1 and lag-2 coefficients, but the remaining coefficients are not distinctly different from zero relative to statistical uncertainty. Using an $AR(2)$ fit to the residuals added to the fitted 24 h cycle produces the overall fit to the temperature data shown in Fig. 18.8. $\square$

In general, autoregressive models may be fit by maximum likelihood. We now connect ML estimation with lagged least-squares regression (p. 531), by writing down the

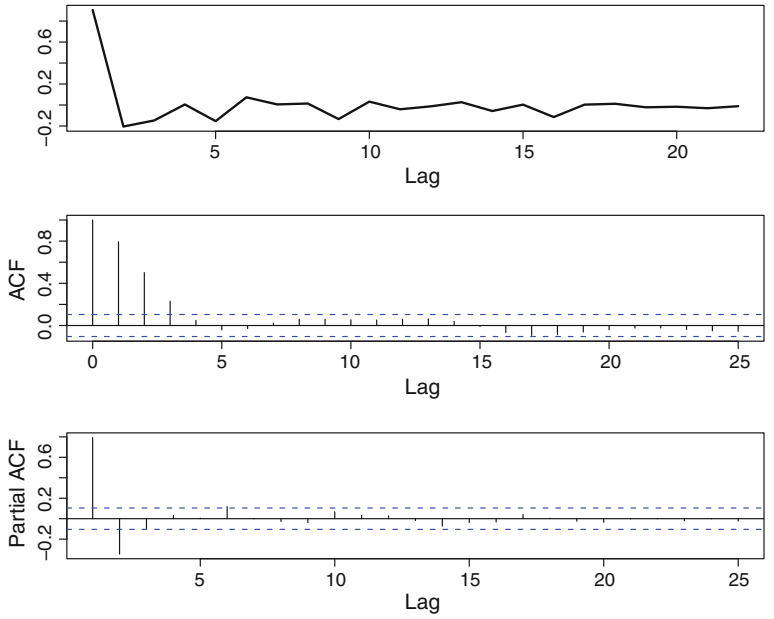

Fig. 18.7 Autoregressive model of order $p = 22$ for core temperature residuals. *Top* Coefficients $\hat{\phi}_i$ as a function of lag i. *Middle* The sample autocorrelation function. *Bottom* The sample partial autocorrelation function.

likelihood function for the $AR(1)$ model, assuming X_t is Gaussian with mean zero and $|\phi| < 1$. We have $X_1 \sim N(0, \sigma_1^2)$ where

$$\sigma_1^2 = \sigma_W^2/(1-\phi^2). \tag{18.28}$$

We also have $X_t|X_{t-1} = x_{t-1} \sim N(\phi x_{t-1}, \sigma_W^2)$ for $t = 2, \ldots, n$. The joint pdf is

$$\begin{aligned} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) &= f_{X_1}(x_1) f_{X_2|X_1}(x_2|X_1 = x_1) \cdots f_{X_n|X_{n-1}}(x_n|X_{n-1} = x_{n-1}) \\ &= \frac{1}{\sigma_1} f_Z(\frac{x_1}{\sigma_1}) \prod_{t=2}^{n} \frac{1}{\sigma_W} f_Z(\frac{x_t - \phi x_{t-1}}{\sigma_W}) \end{aligned}$$

where $f_Z(z)$ is the $N(0, 1)$ pdf. The factors in the product above may be written

$$\begin{aligned} \frac{1}{\sigma_W} f_Z(\frac{x_t - \phi x_{t-1}}{\sigma_W}) &= \frac{1}{\sqrt{2\pi}\sigma_W} \exp(-\frac{(x_t - \phi x_{t-1})^2}{2\sigma_W^2}) \\ &= \frac{1}{\sqrt{2\pi}\sigma_W} \exp(-\frac{(y_{t-1} - \phi x_{B1,t-1})^2}{2\sigma_W^2}). \end{aligned}$$

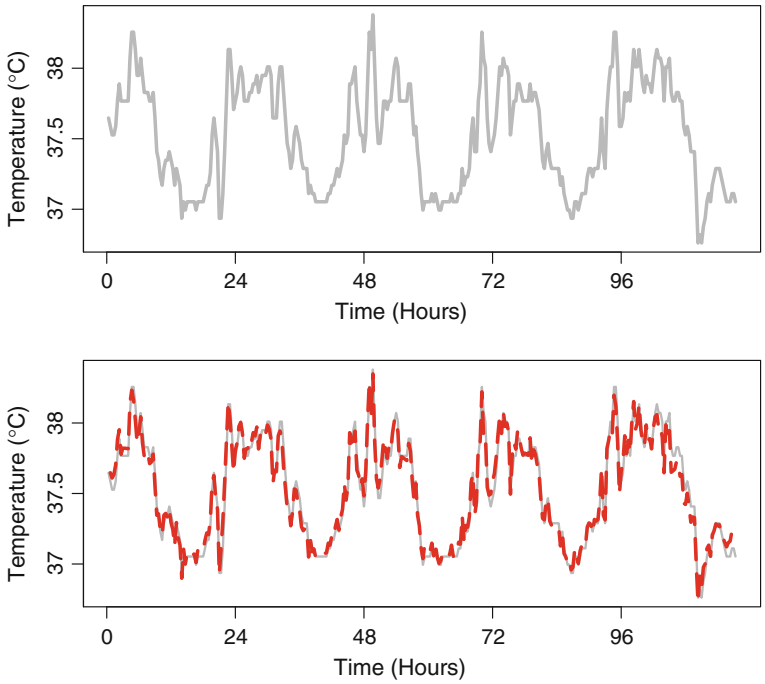

Fig. 18.8 Core temperature data together with fit. *Top* plot of temperature data. *Bottom* Plot of temperature data together with fit (*in red*) based on the sum of an $AR(2)$ fit to residuals and the fitted 24 h cycle.

This final form of each factor is the same as would appear in the likelihood for the regression of y on x_{B1}, with no intercept. Thus, if we ignore x_1, maximizing the likelihood $L(\phi, \sigma_W)$ amounts to solving the ordinary least-squares problem in the regression of y on x_{B1}. This maximization is called *conditional maximum likelihood* because we act as if the distribution of X_1 is given, i.e., it involves no unknown parameters. Because σ_1 in (18.28) is a function of ϕ and σ_W, when we include the factor due to X_1, which is $f_Z(x_1/\sigma_1)/\sigma_1$, the maximization problem changes and it is no longer solvable by least squares. Thus, the MLE must be found by an iterative method, but it is likely to be very close to the conditional MLE. Similar considerations hold also for $AR(p)$ models: the likelihood is nonlinear in the autoregressive parameters, but if we condition on the first p values then ML estimation reduces to ordinary least squares lagged regression. Statistical software for fitting autoregressive models typically either uses ML estimation, or a method that is very nearly equivalent. (The Kalman filter, described in Section 16.2.5, is sometimes used to obtain ML estimates in time series models.) For large samples, the fitted coefficients are essentially the same as those obtained using lagged regression.

The fit to the core temperature data in the bottom panel of Fig. 18.8 combines the fitted 24 h cycle and the $AR(2)$ fit to the residuals. This is an example of *regression*

with time series errors. As mentioned on p. 346, a general approach to regression with time series errors may be based on weighted least squares. Specifically, the model (12.64) may be used with the variance matrix R defined by the $AR(p)$ process and a fit, together with confidence intervals and significance tests, may be obtained[8] from the following steps:

1. Fit the regression variables X to the response variable Y using ordinary least squares;
2. Fit an $AR(p)$ model to the residuals from step 1;
3. Re-fit the regression variables X to the response variable Y using weighted least squares (see p. 345), based on the estimated R matrix found from the fitted auto-regressive model in step 2.

In practice, steps 1-3 may be adequate but, in addition, steps 2 and 3 could be iterated, or ML estimation could be applied once the $AR(p)$ model is determined in Step 2 (e.g., Greenhouse et al. 1987). Statistical software for regression with time series errors is usually based on ML estimation.

18.3 The Periodogram for Stationary Processes

18.3.1 The periodogram may be considered an estimate of the spectral density function.

The DFT is relatively easy to use without thinking about its continuous analogue. However, to understand the way the DFT behaves, and to derive statistical assessments of uncertainty, we must consider the analogous object defined for a theoretical stationary time series $\{X_t; t \in \mathcal{Z}\}$.

Assume $\sigma_t^2 = V(X_t) < \infty$ and let $\mu_t = E(X_t)$. Recall that the autocovariance function is given by

$$\gamma(h) = E((X_t - \mu_t)(X_{t+h} - \mu_{t+h})).$$

Under the summability condition

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty \tag{18.29}$$

general results give the existence of a *spectral density function* $f(\omega)$ for which

[8] The fit in Fig. 18.8 avoided step 3, and would not change very much if step 3 were included, but the statistical inferences involving confidence intervals and significance tests do require step 3.

$$\gamma(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i\omega h} f(\omega) d\omega \tag{18.30}$$

and

$$f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i\omega h}. \tag{18.31}$$

From (18.31) it follows immediately that the spectral density is positive, $f(\omega) = f(-\omega)$, $f(\omega)$ is periodic with period 1, and

$$\gamma(0) = \int_{-\frac{1}{2}}^{\frac{1}{2}} f(\omega) d\omega. \tag{18.32}$$

Equation (18.32) says that the total variability $V(X_t)$ is the integral of the spectral density function. This is a continuous analogue of the discrete decomposition (18.18).

Note that (18.29) rules out pure sinusoids. Signals that have purely periodic (composite sinusoidal) components have "mixed" spectra consisting of "line spectra" representing the pure sinusoids and spectral densities representing everything else.

Returning to the periodogram, defined in Equation (18.25), some manipulations (which we omit) show that it may be written in the form

$$I(\omega_j) = \sum_{h=-(n-1)}^{n-1} \hat{\gamma}(h) e^{-2\pi i\omega_j h} \tag{18.33}$$

where $\hat{\gamma}(h)$ is the sample autocovariance function defined in (18.3). Comparing (18.33) with (18.31), we see that the periodogram may be considered an estimator of the spectral density. In addition, using $\hat{\gamma}(-h) = \hat{\gamma}(h)$, Equation (18.33) shows that the periodogram is proportional to the DFT of the sample covariance function.

Further manipulations show that the periodogram may also be written as

$$I(\omega_j) = \frac{1}{n} \sum_{h=-(n-1)}^{n-1} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \mu)(x_t - \mu) e^{-2\pi i\omega_j h}$$

for $j \neq 0$ and if we replace x_t and $x_{t+|h|}$ with their theoretical counterparts X_t and $X_{t+|h|}$, and then take the expectation, we get

$$E(I(\omega_j)) = \sum_{h=-(n-1)}^{n-1} \left(\frac{n - |h|}{n}\right) \gamma(h) e^{-2\pi i\omega_j h}.$$

Let us consider what happens[9] when $\omega_j \to \omega$ as $n \to \infty$. Assuming the summability condition (18.29) holds we get

$$E(I(\omega_{j_n})) \to \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i \omega h},$$

that is,

$$E(I(\omega_{j_n})) \to f(\omega). \tag{18.34}$$

This result forms a connection between the data-based periodogram and the theoretical spectral density: when the periodogram is considered an estimator of the spectral density, for large samples it is approximately unbiased. However, as we will see in Section 18.3.3, the periodogram only becomes a reasonable estimator after smoothing is applied.

18.3.2 For large samples, the periodogram ordinates computed from a stationary time series are approximately independent of one another and chi-squared distributed.

In Section 18.3.1 we showed that the periodogram may be considered an estimator of the spectral density function, but we ended with the remark that it only becomes reasonable after smoothing. We develop this important observation in Section 18.3.3. Here we first review some basic results on the large-sample distribution of the DFT and periodogram. These allow us to get confidence intervals for quantities based on the periodogram, including smoothed periodograms.

The starting point is to imbed the data $x_1, \ldots, x_t$ in a hypothetical infinite sequence of random variables X_t, where t is taken to run through all integers, including negative integers. The assumptions needed for the distributional results are (1) the time series $\{X_t\}$ is stationary; (2) for sufficiently large h, the variables $\{X_t, t < t_0\}$ are nearly independent of the variables $\{X_t, t > t_0 + h\}$ (for any, and therefore—under stationarity—every, t_0); and (3) the spectral density $f(\omega)$ exists. These conditions allow application of the Central Limit Theorem(CLT) to the sum that defines the DFT. We are being deliberately vague in the statement of (2). For technical discussion see Lahiri (2003a).

To get asymptotic variances and covariances, and the asymptotic distribution of the periodogram, let us replace x_t by X_t in (18.20) and (18.21) and consider the large-sample distribution of the coefficients

[9] To get a sequence of Fourier frequencies ω_j that converge to ω, define $\omega_{j_n} = j_n/n$ with j_n a sequence of integers for which $j_n/n \to \omega$.

$$A_k = \frac{2}{n}\sum_{j=1}^{n} X_j \cos(2k\pi j/n)$$
$$B_k = \frac{2}{n}\sum_{j=1}^{n} X_j \sin(2k\pi j/n).$$

To simplify a little, let us write

$$d_c(\omega_k) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} X_j \cos(2k\pi j/n)$$
$$d_s(\omega_k) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} X_j \sin(2k\pi j/n).$$

We assume that the expectation of X_t is zero (if not, we can subtract $E(X_t)$ from each variable). By the CLT, $d_c(\omega_j)$ and $d_s(\omega_j)$ are approximately normally distributed. In addition, we have $E(d_c(\omega_k)) = E(d_s(\omega_k)) = 0$ and, it turns out, for the large-sample variances we have

$$V(d_c(\omega_k)) \approx \frac{1}{2} f(\omega_k) \tag{18.35}$$
$$V(d_s(\omega_k)) \approx \frac{1}{2} f(\omega_k) \tag{18.36}$$

while the covariances are approximately zero: for $j \neq k$,

$$Cov(d_c(\omega_j), d_c(\omega_k)) \approx 0 \tag{18.37}$$
$$Cov(d_s(\omega_j), d_s(\omega_k)) \approx 0 \tag{18.38}$$

and for all j, k,

$$Cov(d_c(\omega_j), d_s(\omega_k)) \approx 0. \tag{18.39}$$

The asymptotic independence in (18.37)–(18.39) greatly simplifies statistical inference based on the DFT.

The periodogram is related to $d_c(\omega_k)$ and $d_s(\omega_k)$ by

$$I(\omega_k) = d_c(\omega_k)^2 + d_s(\omega_k)^2.$$

From the CLT together with (18.35) and (18.36), we have

$$\sqrt{\frac{2}{f(\omega_k)}} d_c(\omega_k) \xrightarrow{\mathcal{D}} N(0,1)$$
$$\sqrt{\frac{2}{f(\omega_k)}} d_s(\omega_k) \xrightarrow{\mathcal{D}} N(0,1).$$

By (18.39) these two random variables are approximately independent. Recalling that if $Z_1 \sim N(0,1)$ and $Z_2 \sim N(0,1)$, independently, then $Z_1^2 + Z_2^2 \sim \chi_2^2$ we therefore have

$$\frac{2I(\omega_k)}{f(\omega_k)} \text{is approximately } \chi_2^2 \tag{18.40}$$

which we may also write as

$$I(\omega_k) \text{ is approximately } \frac{f(\omega_k)}{2}\chi_2^2.$$

Furthermore, from (18.37)–(18.39), we have that $I(\omega_j)$ and $I(\omega_k)$ are approximately independent for $j \neq k$.

The limiting distribution in (18.40) is a beautifully convenient result, making it relatively easy to get confidence intervals for quantities derived from the periodogram. We describe the methods in Section 18.4.1.

18.3.3 Consistent estimators of the spectral density function result from smoothing the periodogram.

As we discussed in Chapter 8, in large samples the distribution of an estimator T should become concentrated near the quantity θ it is estimating. While (18.40) gives a nice way to assess uncertainty about the periodogram, it also shows that the large-sample distribution of the periodogram does *not* become concentrated around the spectral density: its variance does not decrease with the sample size. In statistical parlance, the periodogram is not a consistent estimator. However, under conditions analogous to those used for consistency of linear smoothers in nonparametric regression, as discussed in Section 15.3.3, smoothed versions of the periodogram will be consistent. This is strong theoretical motivation for smoothing the periodogram.

In the statistical and neuroscientific literatures there are five main approaches to smoothing the periodogram. The first is to apply a smoother, such as a Gaussian kernel smoother to the sequence of values $I(\omega_k)$. Kernel smoothers were discussed in Section 15.3.1 in the context of nonparametric regression and Section 15.4.1 in the context of density estimation. Because kernel smoothers compute linear combinations of the data they are linear smoothers or *linear filters*. We make some further comments about linear filters in Section 18.3.4. When applied to time series Gaussian kernel smoothers are usually called *Gaussian filters*.

The second method of smoothing a periodogram is to split the time domain into a set of many long intervals (long enough to capture low frequencies of potential interest), estimate the spectral density within each interval, and average the resulting estimates. With this method it may be shown that it is advantageous to allow the intervals to have some overlap (Welch 1967). The estimator based on such averaging is sometimes known by the acronym WOSA for *weighted overlapping segment averaging* or *Welch's method*.

The third approach applies a simple generalized linear model based on the asymptotic distribution of the periodogram in (18.40). Recall that the $\chi^2_2/2$ distribution is the same as the standard exponential distribution $Exp(1)$. We may then write

$$I(\omega_k) \stackrel{\cdot}{\sim} f(\omega_k) Exp(1) \tag{18.41}$$

or

$$I(\omega_k) \stackrel{\cdot}{\sim} Exp(\lambda_k) \tag{18.42}$$

where

$$\lambda_k = \frac{1}{f(\omega_k)}.$$

This says that the periodogram ordinates form, approximately, a generalized linear model and therefore may be smoothed using the technology in Section 15.2.3, adapted for exponential regression. The likelihood function based on (18.42) is called the *Whittle likelihood*.

The fourth class of methods for smoothing the periodogram again uses the asymptotic distribution in the form of (18.41) but instead deals with the log ordinates. Letting $Y_k = \log I(\omega_k)$, (18.41) may be written

$$Y_k \approx \log f(\omega_k) + \epsilon_k \tag{18.43}$$

where the ϵ_k variables are independently distributed as $\log X$ where $X \sim Exp(1)$. This provides a standard nonparametric regression model, and the log of an exponential random variable is reasonably close to being normal. However, $E(\epsilon_k) \neq 0$, so there is some bias introduced into the estimation process. Nonetheless, in many cases the bias is small relative to the variation in the log periodogram.

The fifth way to smooth a periodogram is to assume the data follow an autoregressive model, and then use the resulting form of the spectral density. Specifically, calculations show that the $AR(p)$ model (18.27) has spectral density

$$f_X(\omega) = \frac{\sigma_W^2}{|1 - \phi_1 e^{-2\pi i\omega} - \phi_2 e^{-4\pi i\omega} - \cdots - \phi_p e^{-2p\pi i\omega}|^2}.$$

In addition, a more concise class of models, known as *autoregressive moving average* or *ARMA* models, is often used, and these too have closed-form expressions for their spectral densities.

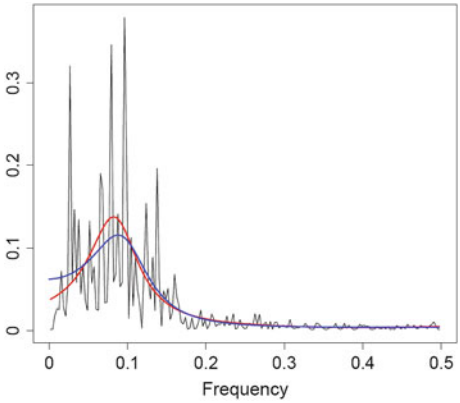

Fig. 18.9 Spectral density estimates for the BARS-detrended residuals from the core body temperature data, after removing the fitted 24 h cycle. The tapered periodogram is highly variable; the Whittle smoothed version is overlaid in *blue*; and the estimate from the $AR(3)$ model is overlaid in *red*.

Example 18.3 (continued from p. 525) We obtained smooth versions of the periodogram for the core temperature data after first removing the trend. (Recall our discussion of Example 15.2 on p. 528; to fit the trend we used the nonparametric regression methodas described briefly in Chapter 15). The $AR(3)$ spectral density estimate is shown in Fig. 18.9. Note that it is very smooth. (An $AR(2)$ based estimate gives similar results.) The Whittle smoothed periodogram is shown for comparison, and agrees reasonably well. There appears to be a peak near $\omega_j = .1$. To interpret this, we need units. The temperature was sampled every 20 min, and there were 352 observations. If $\omega_j = .1$, then the frequency is .1 per time unit (or 35.2 per 352 time units). To get frequency per day we multiply by 72 and get roughly 7. There appears to be a roughly oscillatory component with a period of about 3.5 h. □

We elaborate briefly on linear smoothing in Section 18.3.4 but otherwise omit details on smoothing periodograms.[10] Smoothing is typically handled in spectral analysis software. Regardless of the method used, the most important point is that *some* smoothing is essential.

18.3.4 Linear filters can be fast and effective.

We indicated in Section 18.3.3 that kernel smoothers are linear filters. In this section we say what we mean by a linear filter, and indicate why linear filters are widely applied.

[10] A reference advocating methods three and four, above, is Fan and Kreutzberger (1998).

Suppose we have time series data $x_1, \ldots, x_n$. A linear filter is a set of numbers (coefficients) $\{a_r, a_{r+1}, \ldots, a_s\}$ and its application to the series x_t results in the filtered series

$$y_t = \sum_{h=r}^{s} a_h x_{t-h} \tag{18.44}$$

where, typically, $s - r$ is much less than n. For example, the result of applying the five-point filter with coefficients $(1, 2, 3, 2, 1)/9$ would be

$$y_t = \frac{1}{9}(x_{t-2} + 2x_{t-1} + 3x_t + 2x_{t+1} + x_{t+2}) \tag{18.45}$$

for $t = 3, \ldots, n - 2$. A Gaussian filter would be similar but would instead use a normal (Gaussian) pdf to define the coefficients.

It may be shown that the DFT of $\{y_t\}$ is related to the DFT of $\{x_t\}$ according to

$$d_y(\omega) = \sqrt{n} d_a(\omega) d_x(\omega) \tag{18.46}$$

where $d_a(\omega)$ is the Fourier transform of $\{a_r, a_{r+1}, \ldots, a_s, 0, 0, \ldots, 0\}$, with the zeroes being added to fill up the rest of the n data values. (This is called "padding" the sequence.) The quantity $\sqrt{n} d_a(\omega)$ is called the *transfer function* and its squared magnitude is the *power transfer function*. Expression (18.46) makes it possible to analyze easily the effects of linear filters. This, coupled with their simplicity and the high speed with which they may be computed makes them a very common method of choice for smoothing a time series and the resulting periodogram.

Example 18.3 (continued) We applied the 5-point linear filter described above to the residuals from the core temperature data following simple harmonic regression, yielding a series of the form (18.45). The top panel of Fig. 18.10 shows the residual series and the middle panel shows the power transfer function. The power transfer function decreases to nearly zero as the frequency increases so that high-frequency components have been essentially eliminated from the filtered series. The resulting series is shown in the bottom panel of Fig. 18.10. The filtered series is smoother than the original series. This 5-point linear filter is predominantly a high frequency filter but, as the middle panel of Fig. 18.10 shows, its effects are not restricted to the highest frequencies: there is a gradual squelching of middle-range frequencies as well. □

We have just found that the 5-point linear filter used in (18.45), and applied above to the data from Example 18.3, acts mostly as a high-frequency filter but also displays some gradual mid-range filtering. This might be considered undesirable and one might consider trying to use an ideal high-frequency (or *low-pass*) filter that has a power transfer function of the form

$$H(\omega) = \begin{cases} 1 & \text{for } 0 \le |\omega| \le \omega_c \\ 0 & \text{for } \omega_c < |\omega| \le \frac{1}{2} \end{cases}$$

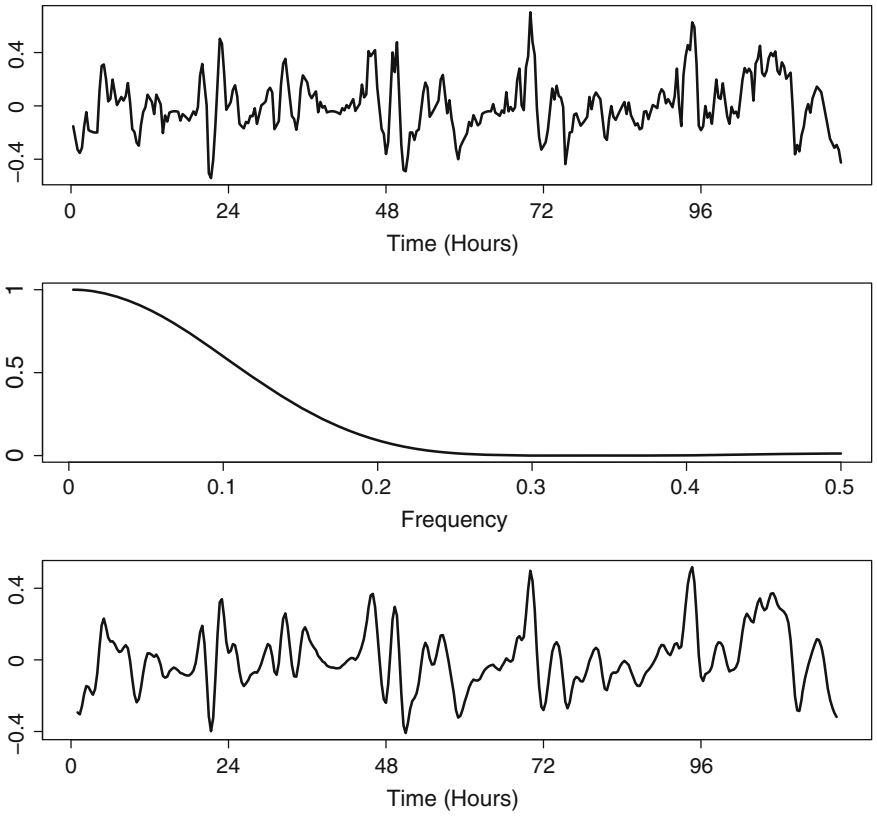

Fig. 18.10 *Top* Core temperature data after removing dominant 24 h effect, i.e., the residuals after simple harmonic regression. *Middle* The power transfer function of the five-point linear filter with coefficients (1, 2, 3, 2, 1)/9, showing a strong diminution of the higher frequency components. *Bottom* Core temperature data after applying the five-point linear filter with coefficients (1, 2, 3, 2, 1)/9.

which would remove all components with frequencies $\omega > \omega_c$ and leave all other components of the series unchanged. One might then, in principle, try to find a filter that corresponds to this power transfer function. This approach turns out to introduce certain technical problems associated with Fourier transforms of discontinuous functions. In practice, time series software typically provides some option for low-pass filtering based on a linear filter, or a combination of linear filters, which aims to approximate the effect of the ideal power transfer function. Similarly, most software provides options for *high-pass* filtering, which approximates an ideal filter that would remove frequencies $\omega < \omega_c$ for some ω_c, and *band-pass* filtering, which approximates an ideal filter that would remove frequencies outside some interval (ω_a, ω_b); the range (ω_a, ω_b) then becomes the frequency band that is retained by the band-pass filter. We illustrated a form of high-pass filtering when we detrended the LFP series

in Example 15.2, with our discussion surrounding Fig. 18.6 (see p. 528), and then again filtered the data in Example 18.3 before fitting the auto-regressive model on p. 532. In the latter case, the detrending method was nonlinear. The advantage of linear filters in practice is the speed with which results may be computed.

All of these remarks about linear filters have theoretical counterparts.

Some details: Suppose $\{X_t; t \in \mathcal{Z}\}$ is a stationary process with spectral density $f_X(\omega)$ and the series $\{a_h; h \in \mathcal{Z}\}$ satisfies

$$\sum_{h=-\infty}^{\infty} |a_h| < \infty.$$

If we let

$$A(\omega) = \sum_{h=-\infty}^{\infty} a_h e^{-2\pi i \omega h},$$

then the filtered process $\{Y_t; t \in \mathcal{Z}\}$ defined by

$$Y_t = \sum_{h=-\infty}^{\infty} a_h X_{t-h}$$

is stationary with spectral density

$$f_Y(\omega) = |A(\omega)|^2 f_X(\omega).$$

Here, the series of coefficients $\{a_h; h \in \mathcal{Z}\}$ is known as the *impulse response function*. □

18.3.5 *Frequency information is limited by the sampling rate.*

While the Fourier frequencies $\omega_k = k/n$ are defined for $k = 1, \ldots, n$, the resulting cosine functions are constrained by the important restriction

$$\cos(2\pi \frac{k}{n} t) = \cos(2\pi \frac{n-k}{n} t) \tag{18.47}$$

for every integer t.

Details: In (18.6) put $u = 2\pi t$ and $v = 2\pi \frac{k}{n} t$ to get

$$\cos(2\pi \frac{n-k}{n} t) = \cos(2\pi t) \cos(2\pi \frac{k}{n} t) + \sin(2\pi t) \sin(2\pi \frac{k}{n} t)$$

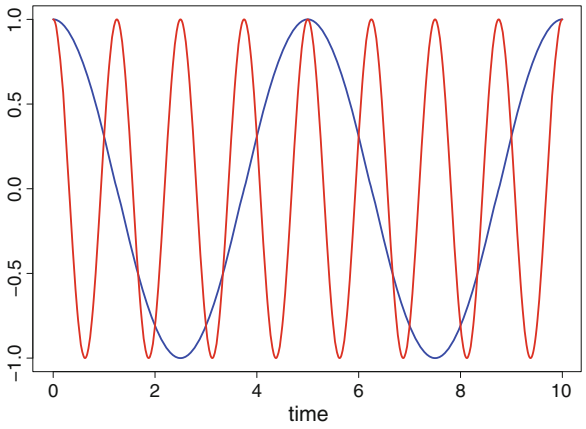

Fig. 18.11 A plot illustrating aliasing of two frequencies for $n = 10$. Two cosine functions are plotted: $\cos(2\pi\omega_1 t)$ and $\cos(2\pi\omega_2 t)$ for $\omega_1 = 2/10$ and $\omega_2 = 8/10$. At all the values $t = 1, \ldots, 10$ these cosine functions agree, so that the frequencies ω_1 and ω_2 are aliased. Note that the time interval between peak and trough corresponding to the second frequency is less than the sampling interval of 1 (equivalently, $\omega_2 > 1/2$) so that, in a sense, the second cosine is oscillating too fast to be determined at this sampling rate. Simple harmonic regression fits for any data sampled at $t = 1, \ldots, 10$ will be the same using ω_2 as using ω_1.

and when t is an integer $\sin(2\pi t) = 0$ while $\cos(2\pi t) = 1$. □

Thus, any cosine with a frequency $\frac{1}{2} < \omega_k < 1$ will have precisely the same values at all integers t as the cosine with frequency $1 - \omega_k$. This is known as *aliasing*: it is not possible to distinguish a cosine function having frequency $\omega^* > \frac{1}{2}$ from another cosine with a frequency in $(0, \frac{1}{2})$. By sampling $x_t = \cos(2\pi\omega t)$ at points $t = 1, \ldots, n$, the fastest visible oscillations occur at the frequency $\omega = \frac{1}{2}$, for which $x_t = \cos(\pi t) = (-1)^t$. (When multiplied by n to get back to the original units of time, this fastest visible frequency of oscillation is called the *Nyquist frequency.*) The situation is illustrated in Fig. 18.11. Corresponding to (18.47) we also have

$$\sin(2\pi\frac{k}{n}t) = -\sin(2\pi(\frac{n-k}{n})t).$$

These aliasing relations have analogues in the DFT. They imply that[11] the second half of the components of the DFT, those for which $\omega_k > \frac{1}{2}$, are redundant with the first. Plots of the periodogram therefore correspond to frequencies only up to $\omega_k = \frac{1}{2}$.

[11] This assumes the data are real numbers. It is occasionally useful, instead, to examine data that consist of complex numbers.

18.3.6 *Tapering reduces the leakage of power from non-Fourier to Fourier frequencies.*

The intuitive description of Fourier analysis in Section 18.2.1 left out an important fact. If we consider the fundamental cosine and sine functions $\cos(2\pi t)$ and $\sin(2\pi t)$, these are functions not only on $[0, 1]$ but on the whole real line. They and all of the resulting cosine and sine functions at harmonic frequencies, i.e., the functions $\cos(2\pi kt)$ and $\sin(2\pi kt)$ for $k = 1, 2, \ldots$, will be periodic on the interval $[0, 1]$. So that all of these functions satisfy

$$f(0) = f(1). \tag{18.48}$$

The rough arguments we gave in Section 18.2.1 make the most sense for functions that satisfy (18.48). When this constraint does not hold, it turns out that the Fourier approximation (18.16) suffers from a failure to adequately represent $f(t)$, which is known as the *Gibbs phenomenon.* The corresponding effect when applying the DFT to data is known as *leakage*.

To describe the problem of leakage, let us consider the periodogram of the cosine function $x_t = \cos(2\pi\omega t)$, for $t = 1, \ldots, n$. Calculation shows that this periodogram (for each Fourier frequency ω_j) is given by

$$I(\omega_j) = n|D_n(\omega - \omega_j)|^2 \tag{18.49}$$

where

$$D_n(\phi) = \frac{\sin(\pi n\phi)}{n\sin(\pi\phi)}$$

is known as the *Dirichlet kernel.* If ω is a Fourier frequency, then $I(\omega_j)$ has a single spike at $\omega_j = \omega$ and is zero at all other Fourier frequencies ω_j. In other words, in this case the periodogram correctly finds the sole cosine component.

> *Details:* Note that as $\phi \to 0$, $D_n(\phi) \to \frac{1}{n}$ (by L'Hopital's rule), so $D_n(\phi)$ at $\phi = 0$ is defined to be $D_n(0) = \frac{1}{n}$. Thus, when $\omega_j = \omega$ we have $I(\omega_j) = \frac{1}{n}$. If ω is a Fourier frequency then $\omega - \omega_j$ has the form $\frac{k}{n}$ for some integer k and $D_n(\omega - \omega_j) = 0$ for all j except when $\omega_j = \omega$. □

On the other hand, when ω is not a Fourier frequency the Dirichlet kernel creates "side lobes," as shown in Fig. 18.12, where $D_n(\omega - \omega_j)$ will be nonzero even for frequencies ω_j that are not immediately non-adjacent to ω. As a consequence, the power at frequency ω will "leak" to other frequencies in the periodogram, so the periodogram indicates misleadingly that those other frequencies are present in the data.

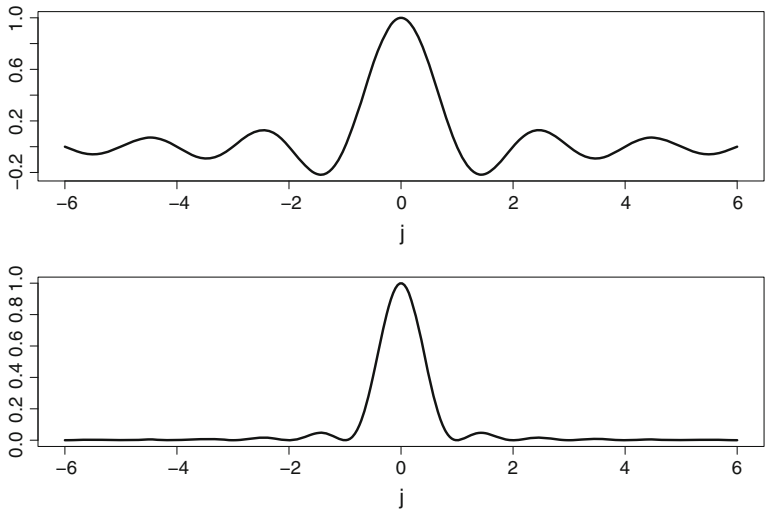

Fig. 18.12 *Top* The Dirichlet kernel $D_{100}(j/100)$, here plotted for values of j ranging from -6 to 6. A continuous *curve* was generated by taking non-integer values of j. *Bottom* The periodogram $I(j/100) = 100|D_{100}(j/100)|^2$, after scaling by dividing by 100.

The problem of leakage is very dramatic when the *dynamic range* of the data is large. Dynamic range refers to the ratio of the largest to smallest positive periodogram values (usually measured on the $\log_{10}$, or decibel, scale).

Illustration: As an illustration, consider

$$x_t = 20\cos(2\pi\omega_1 t) + \cos(2\pi\omega_2 t) \tag{18.50}$$

where $n = 100$, $\omega_1 = .05$ and $\omega_2 = .15$. Its periodogram is shown in the top panel of Fig. 18.13. To see the second frequency it is necessary to use a log scale to plot the periodogram, as shown in the bottom panel of Fig. 18.13. Log periodogram plots are used as defaults in many contexts. Now consider the leakage-prone variant where we take $\omega_1 = 1/22$ rather than $1/20$. Its periodogram is shown in Fig. 18.14. In this case leakage obscures the second peak almost entirely, and if the periodogram were noisy (as it is with real data) it would be extremely difficult to see the second peak at all. □

Leakage is also a problem when there are trends, which cause large low-frequency coefficients in the periodogram.

Example 15.2 (continued from p. 528) We previously showed the log periodogram for the LFP data in Fig. 18.5. The very low frequency trends cause leakage, which obscures the higher frequencies of interest. □

The standard solution to the problem of leakage is to force the data to satisfy (18.48) by applying *tapering*. Tapering decreases bias due to leakage in spectral density

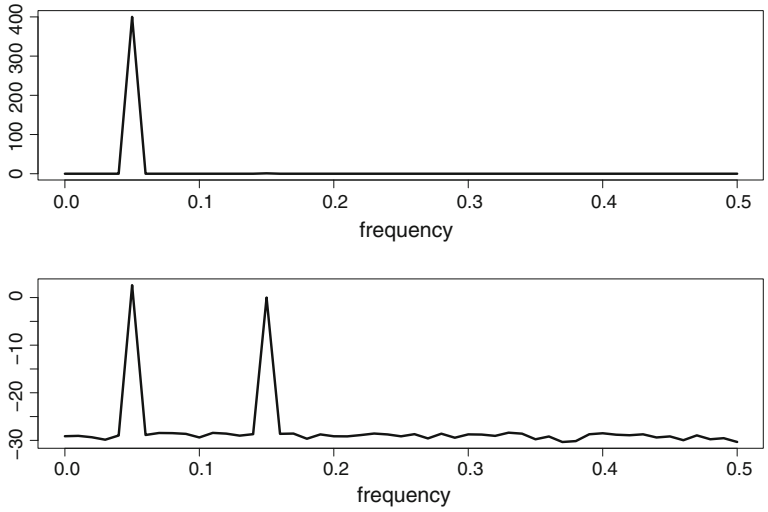

Fig. 18.13 *Top* Periodogram of $x_t = 20\cos(2\pi\omega_1 t) + \cos(2\pi\omega_2 t)$, where $n = 100$, $\omega_1 = .05$ and $\omega_2 = .15$. *Bottom* Log periodogram of x_t. In the log scale the second peak becomes visible.

estimation by damping down the ends of the series toward zero, forcing the series to have period equal to its length (and thus satisfying (18.48)). This is accomplished in standard spectral analysis software. Because the beginning and end of the tapered series have values close to zero, however, this reduces the effective sample size of the series and therefore loses some information. It has been shown that the use of the mean of multiple tapers can recover this information.[12] Multi-taper estimation is used as a default in some software.

18.3.7 Time-frequency analysis describes the evolution of rhythms across time.

Up until this point, Section 18.3 has presented powerful methods for spectral analysis of time series under the assumption of stationarity. We have emphasized that time series should not be considered stationary when there are slowly varying trends, as displayed in Fig. 1.5 of Example 1.6 and Fig. 18.1 of Example 15.2. In many cases, however, a different kind of non-stationarity is present and, in fact, may be of great interest: the frequency content of a signal may change across time.

Example 2.2 (continued from p. 514) The spectrograms in Fig. 2.2 on p. 27 displayed nicely some changes in the frequency content of EEGs across the course of

[12] See Mitra and Pesaran (1999), Percival and Walden (1993), and Thomson (1982).

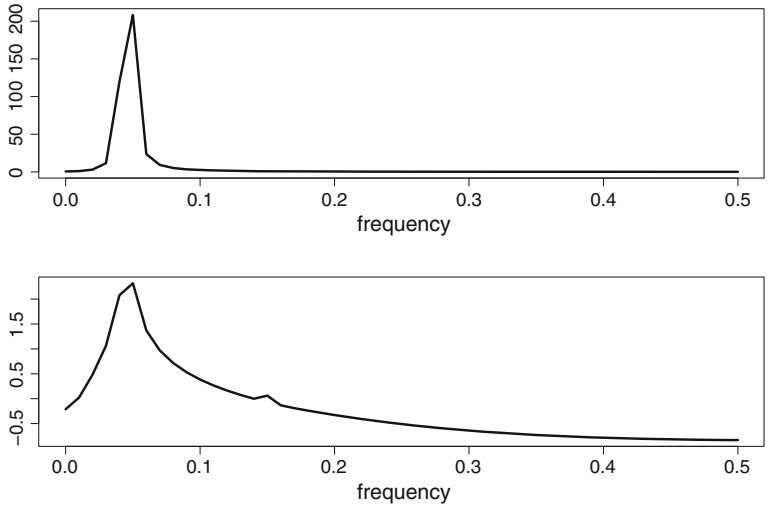

Fig. 18.14 *Top* Periodogram of $x_t = 20\cos(2\pi\omega_1 t) + \cos(2\pi\omega_2 t)$, where $n = 100$, $\omega_1 = 1/22$ and $\omega_2 = .15$. *Bottom* Log periodogram of x_t. Due to leakage, the second peak is obscured.

the experiment. Specifically, the alpha rhythm appeared during an epoch in which the subject's eyes closed, and during induction of anesthesia. □

Spectrograms, such as that in Example 2.2, may be created by segmenting the observation time interval $[0, T]$ into a set of subintervals $[0, T_1], [T_1, T_2], \ldots, [T_k, T]$, and then computing spectral density estimates within each interval. The estimated spectrum is then plotted on the y-axis for every time interval, with time labeled along the x-axis. The intervals must be chosen to be long enough so that there are substantial series from which to estimate the spectrum, yet short enough that the series may be considered stationary within each interval. Some spectrogram software takes as a default 512 observations per interval (with corrections to this to allow for T not being divisible by 512). Some smoothing (and tapering) of the spectral density estimates across time is often incorporated. One way to smooth across time, which is available as an option in most spectrogram software, is to choose the analysis intervals to be overlapping. In some experiments there are repeated trials, in which case the spectrograms may be averaged across trials.

Example 18.2 (continued from p. 518) To display the LFP response to the stimulus Logothetis et al. (2001) used a spectrogram that incorporated tapering and was averaged across trials and across subjects. It showed strong power in the gamma range after onset of the stimulus. □

Time-frequency analysis is often performed using wavelets (Section 15.2.8). Because of the scaling property (the narrowing range) in the definition (15.9), wavelet regression provides a representation that is localized in both time and frequency, with frequency here defined by the scale of the wavelets. See Percival and Walden (2000).

Example 18.1 (continued from p. 518) In their study of MEG oscillatory activity during learning, Chaumon et al. (2009) used Morlet wavelets (see p. 429) to decompose MEG sensor signals across time and frequency. They analyzed the log-transformed power within a 30–48 Hz, band at time 100–400 ms after target onset, from one group of sensors over the occipital lobe and another group of sensors over the frontal lobe. They found that during the learning phase (the first few blocks) of the experiment this gamma band power in the sensors over the occipital lobe was higher for the predictive trials than for the nonpredictive trials ($p < .005$ based on an across-subject paired t-test, using 16 subjects) with the power for the predictive trials being elevated above baseline. On the other hand, during the same learning period, the gamma band power in the sensors over the frontal lobe was depressed for the nonpredictive trials ($p < .0001$), but not for the predictive trials (with the predictive and nonpredictive gamma band power being different, $p < .01$). □

18.4 Propagation of Uncertainty for Functions of the Periodogram

18.4.1 Confidence intervals and significance tests may be carried out by propagating the uncertainty from the periodogram.

The large-sample result described by (18.41) together with the approximate independence of $I(\omega_j)$ and $I(\omega_k)$, for $j \neq k$, provide uncertainty about the estimate of the spectral density and also make it easy to propagate this uncertainty. Importantly, this result holds in the same form for periodograms computed with suitable tapers. (See the brief discussion in Percival and Walden (1993, p. 190), which cites Brillinger (1981, p. 107).)

Now suppose we have computed some feature of the periodogram and we want a 95 % confidence interval associated with that feature. For example, we may have smoothed the periodogram and may want bands to represent our uncertainty. Let $m = (n-1)/2$ if n is odd; $n/2$ if n is even. For a range of ω values, write the smoothed version at frequency ω in the form $g_\omega(I(\omega_1), \ldots, I(\omega_m))$. That is, the operation that produced the smooth value at frequency ω is being written as a function g_ω of the periodogram values. We would say that $g_\omega(I(\omega_1), \ldots, I(\omega_m))$ is an estimator of $f(\omega)$. To apply propagation of error we do the following.

1. For $j = 1$ to J:

   For $i = 1, \ldots, m$:
   generate observations Y_i from an $Exp(1)$ distribution;
   define $U_i^{(j)} = \hat{f}(\omega_i)Y_i$, where $\hat{f}(\omega_i)$ is an estimate of $f(\omega_i)$ (based on a smoothed periodogram).
   Compute $W^{(j)} = g_\omega(U_1^{(j)}, U_2^{(j)}, \ldots, U_m^{(j)})$.

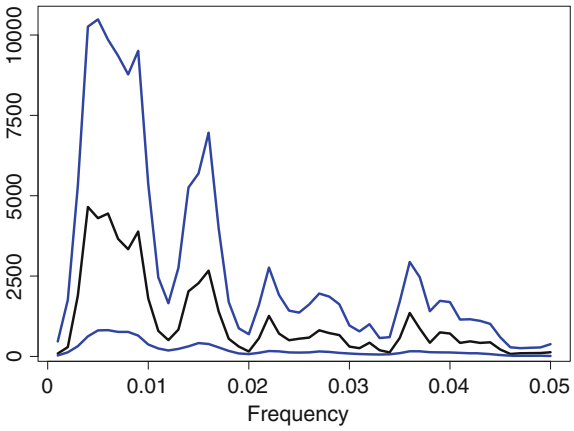

Fig. 18.15 Smoothed periodogram and approximate, pointwise 95 % confidence bands, from the beginning-period LFP detrended series.

2a. Set $\overline{W} = \frac{1}{J}\sum W^{(j)}$ and then $SE^2 = \frac{1}{J-1}\sum(W^{(j)} - \overline{W})^2$ is the squared standard error of $g_\omega(I(\omega_1), \ldots, I(\omega_m))$.

2b. Let $W_{.025}$ and $W_{.975}$ be .025 and .975 quantiles in the sample $W^{(1)}, \ldots, W^{(J)}$. Then $(W_{.025}, W_{.975})$ is an approximate 95 % confidence interval (for $f(\omega)$) associated with $g_\omega(I(\omega_1), \ldots, I(\omega_m))$.

In practice, we would compute a whole set of $W^{(j)}$ values for different g_ω functions, corresponding to different values of ω. This would give us approximate pointwise[13] confidence bands on the smoothed periodogram.

In step 1 of the algorithm above an estimate $\hat{f}(\omega_i)$ (based on the smoothed periodogram) is used in place of $f(\omega_i)$, because the latter is unknown and so can't be computed. This is usually called a bootstrap, analogously to the bootstrap procedures in Chapter 9.

Example 15.2 (continued from p. 528) Returning to the pair of 1 s average LFP recordings, we noted previously, in Figs. 18.1 and 18.5, the need to detrend the time series before looking for periodicities under the assumption of stationarity. Figure 18.6 displayed the smoothed periodograms of the detrended series. Pointwise 95 % confidence bands together with the smoothed periodogram for the first period, obtained by propagation of uncertainty, are shown in Fig. 18.15.

We next consider whether the first and last periods have the same spectral density (an indication of stationarity). Figure 18.16 shows the two smoothed periodograms overlaid. A significance test may be based on the integrated squared difference between the two smooth curves. Specifically, if $\hat{f}_1(\omega)$ and $\hat{f}_2(\omega)$ are the two spectral

[13] By pointwise we mean that at any given frequency ω the bands would provide an approximate 95 % confidence interval. An alternative is to compute approximate *simultaneous* confidence bands, meaning bands that provide approximate 95 % confidence simultaneously for all ω. This may be accomplished with a suitable adaptation of the algorithm.

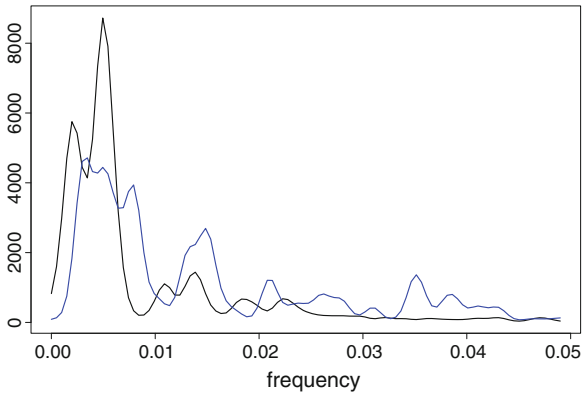

Fig. 18.16 Smoothed periodograms from beginning and end periods, overlaid.

density estimates, then we use

$$t_{obs} = \sum_k (\hat{f}_1(\omega_k) - \hat{f}_2(\omega_k))^2$$

as the test statistic. To compute a p-value under $H_0 = f_1(\omega) = f_2(\omega)$ for all ω, we take as a "pooled" estimate

$$\hat{f}(\omega_k) = \frac{1}{2}(\hat{f}_1(\omega_k) + \hat{f}_2(\omega_k))$$

for $k = 1, \ldots, m$. We then generate a pseudo-sample of pairs of periodograms using $\hat{f}(\omega)$ as the spectral density, and for each generated pair of periodograms, apply smoothing and compute t. We then see what fraction of the generated t values is greater than t_{obs}. This is our approximate p-value. In this case, we obtained $p = 0.53$, indicating no evidence that the spectra from the two recording intervals are different. □

18.4.2 Uncertainty about functions of time series may be obtained from time series pseudo-data.

The method above propagates the uncertainty from the asymptotic distribution of the periodogram to anything computed from it. If, however, an analytical technique bypasses the periodogram a different method must be used to propagate uncertainty. A more general idea is to use the approximate normal distributions on the coefficients, in order to propagate the uncertainty from the DFT itself. In other words, one may begin with the uncertainty in the DFT obtained from the data, and then apply an

inverse DFT to generate time series that behave the same as the original series in the sense of having (approximately) the same spectrum. The resulting time series pseudo-data are sometimes called *surrogate data*.

An efficient method of carrying out such simulations (based on "circulant embedding") is described in Percival and Constantine (2006). Code by these authors is available in the CRAN library of R packages, within the package `fractal`. See below. As described in the Percival and Constantine paper, the method is closely related to *surrogate time series*, e.g., Schreiber and Schmitz (2000). Additional "bootstrap" resampling methods for spectral analysis, with an emphasis on theoretical results, are discussed in Chapter 9 of Lahiri (2003b). We omit detailed discussion of this topic and note only that the pseudo data generated by this approach are normal (Gaussian), and so do not reflect any sources of uncertainty arising from substantial non-normal variation in the data.

18.5 Bivariate Time Series

Suppose $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ are sequences of observations made across time, and the problem is to describe their sequential relationship. For example, an increase in y_t may tend to occur following some increase or decrease in a linear combination of some of the preceding x_t values. This is the sort of possibility that bivariate time series analysis aims to describe.

Example 18.4 Beta oscillations during a sensorimotor task. Brovelli et al. (2004) recorded local field potentials from multiple sites simultaneously while a subject (a rhesus monkey) performed a Go/No-Go visuomotor task. Results were reported for two monkeys. The task required the subject hold down a lever during an interval having a randomly determined length while a stimulus appeared. On Go trials, a reward was given if the monkey released the lever within 500 ms. The purpose of the study was to look for coordinated rhythmic activity across the recording sites during a task that required focused attention. Of particular interest was the range of frequencies identified as *beta oscillations*, which the authors took to be 14–30 Hz. The specific question was whether local field potentials in sensory and motor regions exhibit co-ordinated patterns within the beta range of frequencies. □

The theoretical framework of such efforts begins, again, with stationarity. A joint process $\{(X_t, Y_t), t \in \mathcal{Z}\}$ is said to be *strictly stationary* if the joint distribution of $\{(X_t, Y_t), \ldots, (X_{t+h}, Y_{t+h})\}$ is the same as that of $\{(X_s, Y_s), \ldots, (X_{s+h}, Y_{s+h})\}$ for all integers s, t, h. The process is *weakly stationary* if each of X_t and Y_t is weakly stationary with means and covariance functions μ_X, $\gamma_X(h)$ and μ_Y, $\gamma_Y(h)$, and, in addition, the cross-covariance function

$$\gamma_{XY}(s,t) = E((X_s - \mu_X)(Y_t - \mu_Y))$$

depends on s and t only through their difference $h = t - s$, in which case we write it in the form

$$\gamma_{XY}(h) = E((X_{t-h} - \mu_X)(Y_t - \mu_Y)).$$

Note that $\gamma_{XY}(h) = \gamma_{YX}(-h)$. The *cross-correlation* function of $\{(X_t, Y_t)\}$ is

$$\rho_{XY}(h) = \frac{\gamma_{XY}(h)}{\sigma_X \sigma_Y}$$

where $\sigma_X = \sqrt{\gamma_X(0)}$ and similarly for Y_t. The cross-correlation $\rho_{XY}(h)$ is the ordinary correlation between the random variable X_{t-h} and Y_t. Just as the ordinary correlation ρ may be interpreted as a measure of linear association between two random variables, the cross-correlation $\rho(h)$ may be interpreted as a measure of linear association between two stationary processes at lag h. The cross-covariance and cross-correlation functions are estimated by their sample counterparts:

$$\hat{\gamma}_{XY}(h) = \frac{1}{n} \sum_{t=1}^{n-h} (x_t - \bar{x})(y_{t+h} - \bar{y})$$

with $\hat{\gamma}_{XY}(-h) = \hat{\gamma}_{YX}(h)$, and

$$\hat{\rho}(h) = \frac{\hat{\gamma}_{XY}(h)}{\hat{\sigma}_X \hat{\sigma}_Y}.$$

The univariate Eqs. (18.29)–(18.31) have immediate extensions to the bivariate case: if

$$\sum_{h=-\infty}^{\infty} |\gamma_{XY}(h)| < \infty$$

then there is a *cross-spectral density function* $f_{XY}(\omega)$ for which

$$\gamma_{XY}(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega h} f_{XY}(\omega) d\omega \tag{18.51}$$

and

$$f_{XY}(\omega) = \sum_{h=-\infty}^{\infty} \gamma_{XY}(h) e^{-2\pi i \omega h}.$$

The cross-spectral density is, in general, complex valued. Because $\gamma_{YX}(h) = \gamma_{XY}(-h)$ we have

$$f_{YX}(\omega) = \overline{f_{XY}(\omega)} \tag{18.52}$$

i.e., $f_{YX}(\omega)$ is the complex conjugate of $f_{XY}(\omega)$. In Section 18.3.1 we said that a smoothed periodogram could be considered an estimator of the theoretical spectral

density, and we based that interpretation on a finite-sample expression (18.33), which gave the periodogram as a scaled DFT of the sample covariance function. Similarly, an estimate $\hat{f}_{XY}(\omega)$ of $f_{XY}(\omega)$ may be obtained by smoothing a scaled DFT of the sample cross-covariance function $\hat{\gamma}_{XY}(h)$. In Section 18.5.1 we discuss the important concept of *coherence*, which is defined in terms of the cross-spectral density.

18.5.1 The coherence $\rho_{XY}(\omega)$ between two series X and Y may be considered the correlation of their ω-frequency components.

There is a very nice way to decompose into frequencies the linear dependence between a pair of stationary time series. This frequency-based measure of linear dependence forms an analogy with ordinary correlation which, as we noted in Section 4.2.1, may be interpreted as a measure of linear association. To substantiate this interpretation for the ordinary correlation ρ between two random variables Y and X we provided on p. 81 a theorem concerning the linear prediction of Y from $\alpha + \beta X$, giving the formula for α and β that minimized the mean squared error of prediction, $E\left((Y - \alpha - \beta X)^2\right)$ and showing that when these optimal values of α and β are plugged in, the minimum mean squared error became

$$E\left((Y - \alpha - \beta X)^2\right) = \sigma_Y^2(1 - \rho^2), \tag{18.53}$$

which was Eq. (4.11).

In Eq. (18.53) we considered the linear prediction of Y based on X, meaning the prediction of Y based on a linear function of X. The analogous problem for $\{(X_t, Y_t), t \in \mathcal{Z}\}$ is to assume

$$Y_t = \sum_{h=-\infty}^{\infty} \beta_h X_{t-h} + W_t, \tag{18.54}$$

where W_t is a stationary process independent of $\{X_t\}$, with $E(W_t) = 0$ and $V(W_t) = \sigma_W^2$, and to minimize the mean squared error

$$MSE = E\left(Y_t - \sum_{h=-\infty}^{\infty} \beta_h X_{t-h}\right)^2. \tag{18.55}$$

Some manipulations show that the solution satisfies

$$\min MSE = \int_{-\frac{1}{2}}^{\frac{1}{2}} f_Y(\omega)(1 - \rho_{XY}(\omega)^2)d\omega \tag{18.56}$$

where

$$\rho_{XY}(\omega)^2 = \frac{|f_{XY}(\omega)|^2}{f_X(\omega) f_Y(\omega)} \tag{18.57}$$

is the *squared coherence*. Thus, in analogy with (18.53), $f_Y(\omega)(1 - \rho_{XY}(\omega)^2)$ is the ω-component of the minimum-*MSE* fit of (18.54). In (18.56) we have $MSE \geq 0$ and $f_Y(\omega) \geq 0$, which together imply that $0 \leq \rho_{XY}(\omega)^2 \leq 1$ for all ω, and when

$$Y_t = \sum_{h=-\infty}^{\infty} \beta_h X_{t-h}$$

we have $\rho_{XY}(\omega)^2 = 1$ for all ω. These facts, together with (18.56), give the interpretation that the squared coherence is a frequency-based analogue to squared correlation between two theoretical time series.

Additional details: The interpretation of coherence in terms of correlation may be pushed further, but is somewhat subtle. In defining the cross-spectral spectral density we mentioned that it is complex valued. Let $\theta(f_{XY}(\omega))$ be the phase of $f_{XY}(\omega)$, which we may write in terms of the real and imaginary parts of $f_{XY}(\omega)$,

$$\theta(f_{XY}(\omega)) = \arctan \frac{Im(f_{XY}(\omega))}{Re(f_{XY}(\omega))}$$

so that

$$f_{XY}(\omega) = |f_{XY}| \exp\left(i\theta(f_{XY}(\omega))\right).$$

The function $\theta(f_{XY}(\omega))$ is often called the *phase coherence*. The *coherence* is then the complex-valued function defined by

$$\rho_{XY}(\omega) = \frac{f_{XY}(\omega)}{\sqrt{f_X(\omega) f_Y(\omega)}}.$$

This complex-valued coherence contains phase information, which is necessary when considering the tendency of two signal components at frequency ω to vary together. The magnitude of the coherence is often considered to be a measure of phase-locking of the two signals, but it also depends on the relationship of their amplitudes.
A more complete explanation of coherence is beyond the scope of our presentation here.[14] □

From a pair of observed time series the squared coherence may be estimated by

[14] One helpful fact is that an average coherence across a given frequency band may be shown to be equal to the complex-valued correlation between band-pass filtered versions of the two series; see Ombao and Vanbellegem (2008).

$$\hat{\rho}^2_{XY}(\omega) = \frac{|\hat{f}_{XY}(\omega)|^2}{\hat{f}_X(\omega)\hat{f}_Y(\omega)} \tag{18.58}$$

where, again, $\hat{f}_{XY}(\omega)$ is a smoothed version of the DFT of $\hat{\gamma}_{XY}(h)$. However, the smoothing in this estimation process is crucial. The raw cross-periodogram $I_{XY}(\omega)$ satisfies the relationship

$$|I_{XY}(\omega)|^2 = I_X(\omega)I_Y(\omega)$$

so that plugging the raw periodograms into (18.58) will always yield the value 1. Thus, again, it is imperative to smooth periodograms before interpreting them.

Example 18.4 (continued from p. 553) Brovelli et al. collected approximately 900 successful Go trials, using data from 90 ms prior to stimulus onset to 500 ms after onset. They subtracted out the trial-averaged signals to produce approximately stationary multiple time series. To look for the presence of beta oscillations in sensorimotor cortex they recorded from six sites in one animal and four in another. The sites are shown in Fig. 18.17. The sites shown in part A of the figure appear to be in (1) the arm area of primary motor cortex (M1), (2) the arm area of sensory cortex (S1), (3) anterior intraparietal cortex (AIP, object and hand shape representation), (4) lateral intraparietal cortex (used in guiding saccades and identifying visual locations), (5) ventral premotor cortex, (6) dorsal premotor cortex. In part B of the figure the sites appear to be in (1) the wrist area of M1 or ventral premotor cortex, (2) the wrist area of S1, (3) AIP, (4) medial intraparietal cortex (related to goals or targets of intended reach).

The authors computed squared coherence for each pair of sites, as in (18.57), with ω in the beta range, then found the maximum squared coherence across all values of ω, and performed a permutation significance test (see Section 11.2.1) to see whether that maximum was sufficiently large to form clear evidence of underlying coherence in LFP across brain regions. Their results are depicted on the left side of Fig. 18.17. The authors found that primary motor cortex (M1, site 1 in both monkeys), primary sensory cortex (S1, site 2), and anterior intraparietal cortex (AIP, site 3) were all engaged in coherent oscillatory activity during the task. □

18.5.2 In examining cross-correlation or coherence of two time series it is advisable first to pre-whiten the series.

In Section 12.2.3 we highlighted the importance of the assumption of independent errors in linear regression: we showed that the squared correlation between two *independent* $AR(1)$ time series is likely to be statistically significant, erroneously indicating association. A similar phenomenon occurs for the cross-correlation, and for coherence. To avoid it, the serial dependence should be removed from the two series before the cross-correlation or coherence is computed. For example, if we have two series $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ we could fit appropriate AR models to each

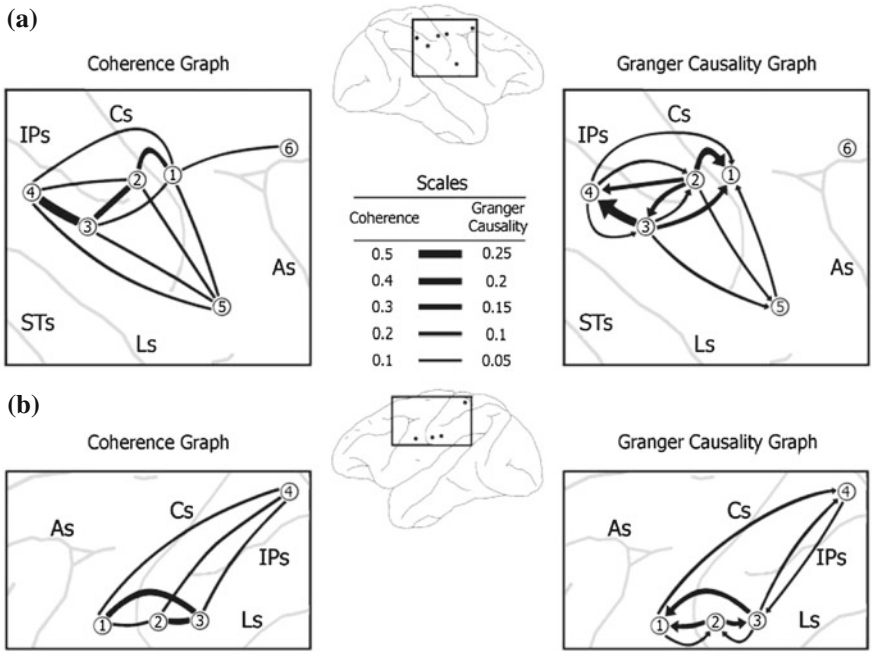

Fig. 18.17 Figure adapted from Brovelli et al. showing coherence and Granger causality among six recording sites in one monkey (part A) and four in another (part B). On the *left* are lines representing statistically significant coherence between a pair of sites ($p < .005$ based on a permutation test with a correction for multiple comparisons), with thickness indicating the magnitude of coherence as shown on the scale graphic in the middle of the figure. On the *right* are lines, some of which have *arrows*, representing statistically significant Granger causality, with magnitudes again indicated by line thickness as shown on the scale graphic in the middle of the figure. Recording sites are shown above and below the scale graphic.

series and then work instead with the residuals obtained from subtracting the *AR* fits. An alternative procedure involves fitting an *AR* (or *ARMA*) model then applying a suitable filter that removes the serial dependence. See Box et al. (2008) for discussion of this approach.

Example 18.2 (continued from p. 518) In their study Logothetis et al. (2001) reported the distribution of R^2 values between[15] LFP and BOLD signals across trials, which were generally substantial, with a mean of .52. Before computing these correlations, however, they pre-whitened the series using $AR(10)$ models. □

[15] Actually, they reported R^2 between stimulus-based impulse response functions (see p. 544) found from the LFP and BOLD signals.

18.5.3 *Granger causality measures the linear predictability of one time series by another.*

The squared coherence provides a frequency-based measure of linear association between two time series. Just as the correlation $Cor(X, Y)$ is symmetrical in its arguments X and Y, so too is the squared coherence. In contrast, regression is directional. We now develop a simple directional assessment of linear predictability of one time series from another.

The idea is very simple. In ordinary regression we assess the influence of a variable (or set of variables) X_2 on Y in the presence of another variable (or set of variables) X_1 by examining the reduction in variance when we compare the regression of Y on (X_1, X_2) with the regression of Y on X_1 alone. If the variance is reduced sufficiently, then we conclude that X_2 helps explain (predict) Y. Here, we replace Y with Y_t, replace X_1 with $\{Y_s, s < t\}$ and X_2 with $\{X_s, s < t\}$. In other words, we examine the additional contribution to predicting Y_t made by the past observations of X_s after accounting for the autocorrelation in $\{Y_t\}$. The "causality" part comes when the past of X_s helps predict Y_t but the past of Y_s does *not* help predict X_t.

Let us begin by defining what it means for $\{(X_t, Y_t), t \in \mathcal{Z}\}$ to follow a joint $AR(p)$ process. Working by analogy with the definition (18.27), we write

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \sum_{i=1}^{p} \begin{pmatrix} \phi_i^{XX} & \phi_i^{XY} \\ \phi_i^{YX} & \phi_i^{YY} \end{pmatrix} \begin{pmatrix} X_{t-i} \\ Y_{t-i} \end{pmatrix} + \begin{pmatrix} W_t^{X|XY} \\ W_t^{Y|XY} \end{pmatrix} \tag{18.59}$$

where $W_t^{X|XY}$ and $W_t^{Y|XY}$ are independently $N(0, \sigma^2_{X|XY})$ and $N(0, \sigma^2_{Y|XY})$. The notational superscripts and subscripts $X|XY$ and $Y|XY$ are used to indicate variables or variances for the joint $AR(p)$ model (18.59), in which both $X_1, \ldots, X_{t-p}$ and $Y_1, \ldots, Y_{t-p}$ appear on the right-hand side. This is in contrast to the usual univariate $AR(p)$ models for $\{Y_t, t \in \mathcal{Z}\}$,

$$Y_t = \sum_{i=1}^{p} \phi_i^Y Y_{t-i} + W_t^Y, \tag{18.60}$$

where W_t^Y are independently[16] $N(0, \sigma^2_{Y|Y})$, and for $\{X_t, t \in \mathcal{Z}\}$,

$$X_t = \sum_{i=1}^{p} \phi_i^X X_{t-i} + W_t^X, \tag{18.61}$$

where W_t^X are independently $N(0, \sigma^2_{X|X})$. We may now say that $\{X_t, t \in \mathcal{Z}\}$ is predictive of $\{Y_t, t \in \mathcal{Z}\}$ if $\sigma_{Y|XY} < \sigma_{Y|Y}$. In this situation, $\{X_t, t \in \mathcal{Z}\}$ is also said to be

[16] Here $\sigma^2_{Y|Y}$ is a constant; the notation is intended only to indicate that it is the error variance when Y appears on both the left-hand side and the right-hand side of the model.

Granger causal of $\{Y_t, t \in \mathcal{Z}\}$. Similarly, we say $\{Y_t, t \in \mathcal{Z}\}$ is predictive (Granger causal) of $\{X_t, t \in \mathcal{Z}\}$ if $\sigma_{X|XY} < \sigma_{X|X}$. This kind of predictability is often quantified by the *Granger causality measure*

$$F_{X\to Y} = 2\log\frac{\sigma_{Y|Y}}{\sigma_{Y|XY}}.$$

Theoretical analysis of this approach was given by Geweke (1982), based on earlier work by Granger (1969).[17]

In applications, to evaluate whether a time series $x_t, t = 1, \ldots, n$ is predictive of $y_t, t = 1, \ldots, n$, the basic procedure is to (1) fit a bivariate $AR(p)$ model, then (2) test the hypothesis $H_0 : \phi_i^{YX} = 0$ for all i, which is equivalent to testing $H_0 : F_{X\to Y} = 0$.

Illustration As an illustration, we simulated a bivariate time series of length 1,000 using the model

$$\begin{aligned} X_t &= .5X_{t-1} + U_t \\ Y_t &= .2Y_{t-1} + .5X_{t-1} + V_t \end{aligned}$$

where $U_t \sim N(0, (.2)^2)$ and $V_t \sim N(0, (.2)^2)$, independently. We then fit a linear regression model of the form

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 X_{t-1} + \epsilon_t$$

and, similarly, fit another model of the same form but with the roles of X and Y reversed. The results for the two regressions are shown in the following table.

Variable	Coefficient	Std. Err.	t-ratio	p-value
Intercept	−.001	.006	−.211	.83
x_{t-1}	.496	.012	42.7	$<10^{-15}$
y_{t-1}	.192	.018	10.7	$<10^{-15}$
Intercept	.008	.016	.536	.59
x_{t-1}	.508	.029	17.1	$< 10^{-15}$
y_{t-1}	−.055	.045	−1.3	.228

As expected, the first fit indicates that X_{t-1} provides additional information beyond Y_{t-1} in predicting Y_t, while the second fit shows that Y_{t-1} does *not* provide additional information beyond X_{t-1} in predicting X_t. This is sometimes summarized by saying

[17] In addition, Geweke (1982) defined a spectral measure $f_{X\to Y}(\omega)$ representing the ω-component of Granger causality in the sense that

$$F_{X\to Y} = \int_{-\frac{1}{2}}^{\frac{1}{2}} f_{X\to Y}(\omega)d\omega.$$

X_t is *causally* related to Y_t, but we must keep in mind that "causal" is used in a predictive, time-directed sense. □

This illustration sweeps under the rug the selection of auto-regressive order p in part of the problem, in step (1) above. In applications this is non-trivial, and care should be taken to make sure interpretations do not depend on choices of p that involve substantial uncertainty.

Example 18.4 (continued from p. 556) Results of Brovelli et al. based on coherence analysis were discussed on p. 556 and were displayed on the left-hand side of Fig. 18.17. Those authors went on to fit an $AR(10)$ model to the data from both monkeys, noting that $AR(5)$ and $AR(15)$ gave consistent results, and that AIC (see Section 11.1.6) would select $AR(15)$ (they considered $AR(p)$ models up through order $p = 15$). They then applied Granger causality[18] analysis, which allowed them to produce the additional directional interpretations shown on the right-hand side of Fig. 18.17. In particular, beta rhythms in primary sensory cortex (site 2 in both monkeys) were predictive of the rhythms in other locations, while primary motor cortex (site 1) tended to be predicted by both sensory and AIP signals and was itself only weakly predictive of signals at other sites. □

[18] They used the spectral decomposition mentioned in the footnote on p. 559 to plot the frequency representation of Granger causality, found its peak, and performed a permutation test analogously to what they had done in analyzing coherence.

Chapter 19
Point Processes

At the beginning of this book, in Example 1.1 (p. 3), we described the activity of a neuron recorded from the supplementary eye field. Interpreting Fig. 1.1 we said that, toward the end of each trial, the neuron fired more rapidly under one experimental condition than under the other. In that discussion we took for granted one of the foundational teachings[1] of neurophysiology, that neurons respond to a stimulus or contribute to an action by increasing their firing rate. But what, precisely, do we mean by "firing rate?" The definition of firing rate turns out to be both subtle and important for statistical analysis of neural data.

Perhaps the simplest conception is that firing rate (FR) is number of spikes (action potentials) per unit time. To compute it we would then count spikes over a time interval of length Δt and write

$$FR = \frac{\text{number of spikes}}{\Delta t}. \tag{19.1}$$

While useful in many contexts, Eq. (19.1) suffers from a fundamental difficulty: it depends strongly on the interval used in the calculation. As an extreme case, suppose we were to examine an interval of length $\Delta t = 100\,\text{ms}$ containing a single spike. Rewriting in terms of seconds, we get $\Delta t = .1\,\text{s}$ (seconds) and this would give us $FR = 10$ spikes per second (10 Hz). But now suppose we shrink the interval down to $\Delta t = 5\,\text{ms}$. Then we would have $\Delta t = .005\,\text{s}$ and we would get $FR = 200\,\text{Hz}$, which is drastically different. How would we know what interval to choose?

To avoid this conundrum, and to begin the process of formulating a statistical model, we do two things. First, we replace the spike count by its theoretical counterpart, the expected spike count, and then we pass to the limit as $\Delta t \to 0$ so that we obtain a firing rate at time t that no longer involves an interval. In other words, we define a theoretical *instantaneous firing rate*. Note that for small Δt the

[1] Description of this phenomenon began with work of Edgar Adrian and Keffer Hartline and their colleagues (e.g., Adrian and Zotterman 1926; Hartline and Graham 1932).

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_19,

spike count in (19.1) is either 0 or 1, which is a Bernoulli event with expected value $P(\text{spike in } (t, t+\Delta t))$. The theoretical instantaneous firing rate at time t then becomes

$$FR(t) = \lim_{\Delta t \to 0} \frac{P(\text{spike in } (t, t+\Delta t))}{\Delta t}. \tag{19.2}$$

However, the definition in (19.2) omits any mention of the experimental context of the observed firing rate. A more inclusive way to write firing rate as a function of time is to allow it to depend on variables we write, collectively, as a vector x_t. The vector x_t might refer to an experimental condition or it could involve such things as refractory effects due to a previous spike shortly before time t (see Section 19.1.3), or a local field potential that represents a substantial component of synaptic input to the cell. We therefore have a more complete conceptualization of firing rate by putting it in the form

$$FR(t|x_t) = \lim_{\Delta t \to 0} \frac{P(\text{spike in } (t, t+\Delta t)|x_t)}{\Delta t}. \tag{19.3}$$

To flesh this out we must say how we calculate the probability in the numerator of (19.3), which will take us through Section 19.3.2. Granting that we will get there, we may state the central idea in statistical modeling of spike train data: neurophysiological phenomena may be represented through variables x_t that are thought to influence spiking activity. A statistical model for spike trains involves two things: (1) a simple, universal formula for the probability density of the spike train in terms of the instantaneous firing rate function, and (2) a specification of the way the firing rate function depends on variables x_t.

A major theme of this book is the use of probability to describe variation. In Chapter 3 we considered events, which led to our description of variation using probability distributions, and in Chapter 18 we examined sequences of temporally-dependent observations, which were modeled as time series. Spike trains, however, don't quite fit into any of the molds we have constructed in the foregoing chapters. They are sequences of varying *event times*, times at which action potentials (spikes) occur—in repeated trials the spike times typically vary, as may be seen in Fig. 1.1 of Example 1.1. To handle such sequences of event times we invoke a special class of models called *point processes*. As we discuss in Section 19.3.4, the tools needed for fitting point processes to spike train data are generalized linear models (Chapter 14) and nonparametric regression (Chapter 15). Indeed, the models we discuss that involve instantaneous firing rate, conceptualized by (19.3), are called *point process regression models*. The purposes of this chapter are, first, to review the way point process representations of spike trains are defined in terms of instantaneous firing-rate functions and, second, to show how point process regression models help in understanding neural behavior.

The name "point process" reflects the localization of the events as points in time together with the notion that the probability distributions evolve across time according to a *stochastic process*. Point processes can be more general, so that the points

can lie in a higher-dimensional physical or abstract space. In PET imaging, for example, a radioisotope that has been incorporated into a metabolically active molecule is introduced into the subject's bloodstream and after these molecules become concentrated in specific tissues the radioisotopes decay, emitting positrons which may be detected. These emissions represent a four-dimensional *spatiotemporal* point process because they are localized occurrences both spatially, throughout the tissue, and in time. Here, however, we focus on point processes in time and their application to modeling spike trains.

The simplest point processes are *Poisson processes*, which are *memoryless* in the sense that the probability of an event occurring at a particular time does not depend on the occurrence or timing of past events. In Section 19.2.1 we discuss *homogeneous* Poisson processes, which can describe highly irregular sequences of event times that have no discernible temporal structure. When an experimental stimulus or behavior is introduced, however, time-varying characteristics of the process become important. In Section 19.2.2 we discuss Poisson processes that are *inhomogeneous* across time. In Section 19.3 we describe ways that more general processes can retain some of the elegance of Poisson processes while gaining the ability to describe a rich variety of phenomena.

Spike trains are fundamental to information processing in the brain, and point processes form the statistical foundation for distinguishing signal from noise in spike trains. We have already seen in Chapters 14 and 15 examples of spike train analysis using Poisson regression with spike counts. For this purpose, the Poisson regression model may be conceptualized as involving counts observed over time bins of width Δt based on a neural firing rate $FR(t)$. In Poisson regression, each Poisson distribution has mean equal to $FR(t) \cdot \Delta t$ and then $FR(t)$ is related to the stimulus (or the behavior) by a formula we may write in short-hand as

$$\log FR(t) = \text{stimulus effects}, \tag{19.4}$$

meaning that $\log FR(t)$ is some function that is determined by the stimulus or behavior. In Example 14.5, for instance, the right-hand side of (19.4) involved a quadratic function that represented the effective distance of a rat from the preferred location of a particular hippocampal place cell, and the result was a Poisson regression model of the place cell's activity. This sort of model may be considered a kind of simplified prototype. When we pass to the limit as in (19.2) and use instantaneous firing rate, the Poisson regression model becomes a Poisson process regression model.

Poisson processes are important, and they are especially useful for analyzing the trial-averaged firing rate. When, in Example 15.1, we displayed the smoothed PSTH under two experimental conditions, we were comparing two trial-averaged firing-rate functions. We spell this out in Section 19.3.3. On the other hand, many phenomena can only be studied *within trials*. For instance, oscillatory behavior, bursting, and some kinds of influences of one neuron on another show substantial variation across trials and may be difficult or impossible to detect from across-trial summaries like the PSTH. Careful examination of spike trains within trials usually reveals non-Poisson behavior: neurons tend not to be memoryless, but instead exhibit effects

of their past *history* of spiking (e.g., of refractory effects or recent burst activity). Non-Poisson models that incorporate history effects are described in Section 19.3, and methods developed in that section produce within-trial analyses of spike trains. In such cases, the instantaneous firing rate takes the form (19.3) and Eq. (19.4) must be modified by including additional terms (as components of the variable x_t) on the right-hand side to incorporate effects that occur differently on each trial. For instance, a firing-rate model might have the form

$$\log FR(t|x_t) = \text{stimulus effects} + \text{history effects} + \text{coupling effects}. \tag{19.5}$$

In Section 19.3.4 we indicate how spike train data may be analyzed by fitting models suggested by conceptualizations like (19.5), again using the methods developed in Chapters 14 and 15.

19.1 Point Process Representations

19.1.1 A point process may be specified in terms of event times, inter-event intervals, or event counts.

If $s_1, s_2, \ldots, s_n$ are times at which events occur within some time interval we may take $x_i = s_i - s_{i-1}$, i.e., x_i is the elapsed time between s_{i-1} and s_i, and define $x_1 = s_1$. This gives the inter-event waiting times x_i from the event times and we could reverse the arithmetic to find the event times from a set of inter-event waiting times $x_1, \ldots, x_n$ using $s_j = \sum_{i=1}^{j} x_i$. In discussing point processes, both of these representations are useful. In the context of spike trains, $s_1, s_2, \ldots, s_n$ are the spike times, while $x_1, \ldots, x_n$ are the inter-spike intervals (ISIs). Nearly all of our discussion of event-time sequences will involve modeling of spike train behavior.

To represent the variability among the event times we let $X_1, X_2, \ldots$ be a sequence of positive random variables. Then the sequence of random variables $S_1, S_2, \ldots$ defined by $S_j = \sum_{i=1}^{j} X_i$ is a *point process* on $(0, \infty)$. In fitting point processes to data, we instead consider finite intervals of time over which the process is observed, and these are usually taken to have the form $(0, T]$, but for many theoretical purposes it is more convenient to assume the point process ranges across $(0, \infty)$.

Another useful way to describe a set of event times is in terms of the counts of events observed over time intervals. The event count in a particular time interval may be considered a random variable. For theoretical purposes it is helpful to introduce a function $N(t)$ that counts the total number of events that have occurred up to and including time t. $N(t)$ is called the *counting process* representation of the point process. See Fig. 19.1. If we let $\Delta N_{(t_1,t_2]}$ denote the number of events observed in the interval $(t_1, t_2]$, then we have $\Delta N_{(t_1,t_2]} = N(t_2) - N(t_1)$. The count $\Delta N_{(t_1,t_2]}$ is often called the *increment* of the point process between t_1 and t_2. In the case of a

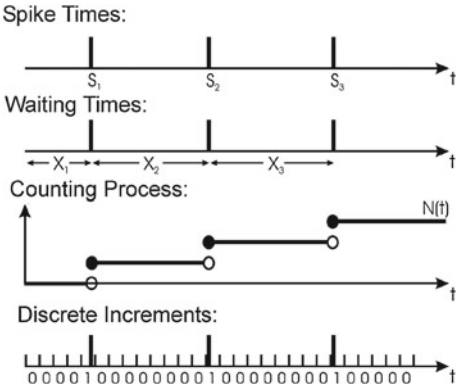

Fig. 19.1 Multiple specifications for point process data: the process may be specified in terms of spike times, waiting times, counts, or discrete binary indicators.

neural spike train, S_i would represent the time of the ith spike, X_i would represent the ith inter-spike interval (ISI), and $\Delta N_{(t_1,t_2]}$ would represent the spike count in the interval $(t_1, t_2]$. For event times S_i and inter-event waiting times X_i we are dealing with mathematical objects that are already familiar, namely sequences of random variables, with the index i being a positive integer. The counting process, $N(t)$, on the other hand, is a *continuous-time stochastic process*, which determines count increments that are random variables.

Keeping track of the times at which the count increases is equivalent to keeping track of increments. Furthermore, for successive spike times s_i and s_{i+1},if we set $t_1 = s_i$ and consider $t_2 < s_{i+1}$ then $\Delta N_{(t_1,t_2]} = 0$ but when $t_2 = s_{i+1}$ then $\Delta N_{(t_1,t_2]} = 1$. Thus, keeping track of the times at which the count increases is equivalent to keeping track of events themselves and, therefore, the counts provide a third way to characterize a point process.

As an example of the way we may identify the event times with the counting process, the set of times for which the counting process is less than some value j, $\{t : N(t) < j\}$, is equivalent to the set of times for which the jth spike has not yet occurred, $\{t : S_j > t\}$. Both of these representations express the set of all times that precede the jth spike, but they do so differently. We can describe a point process using spike times, interspike intervals, or counting processes and specifying any one of these fully specifies the other two. It is often possible to simplify theoretical calculations by taking advantage of these multiple equivalent representations.

19.1.2 A point process may be considered, approximately, to be a binary time series.

At the beginning of the chapter we said that point process data are analyzed using the framework of generalized linear models. This requires the discrete representation

given at the bottom of Fig. 19.1. The event times, inter-event intervals, and counting process all specify the point process in *continuous* time. Suppose we take an observation interval $(0, T]$ and break it up into n small, evenly-spaced time bins. Let $\Delta t = T/n$, and $t_i = i \cdot \Delta t$, for $i = 1, \ldots, n$. We can now consider the discrete increments $\Delta N_i = N(t_i) - N(t_{i-1})$, which count the number of events in a single bin. If we make Δt small enough, it becomes extremely unlikely for there to be more than one event in a single bin. The set of increments $\{\Delta N_i; i = 1, \ldots, n\}$ then becomes a sequence of 0s and 1s, with the 1s indicating the bins in which the events are observed (see Fig. 19.1). In the case of spike trains, data are often recorded in this form, with $\Delta t = 1$ ms. To emphasize the point, we define $Y_i = \Delta N_i$, and put $p_i = P(Y_i = 1)$, so that $Y_i \sim Bernoulli(p_i)$. The Y_is form a binary time series, that is, a sequence of Bernoulli random variables that may be inhomogeneous (the p_i may be different) and/or dependent. Such a discrete-time process is yet another way to represent a point process, at least approximately. It loses some information about the precise timing of events within each bin, but for sufficiently small Δt this loss of information becomes irrelevant for practical purposes. Also, for small Δt we have small p_i and the Bernoulli distributions may be approximated by Poisson distributions, according to the result in Section 5.2.2. In other words, for small Δt we may consider the point process to be essentially a sequence of Poisson random variables. This will allow us to use Poisson regression methods (which are part of generalized linear model methodology) in analyzing data modeled as point processes. The rest of this chapter is largely devoted to filling in the details and fleshing out the consequences, thereby supplying the substance behind the informal statements (19.4) and (19.5).

19.1.3 Point processes can display a wide variety of history-dependent behaviors.

In many stochastic systems, past behavior influences the future. The biophysical properties of ion channels, for example, make it impossible for a neuron to fire again immediately following a spike, creating a short interval known as the absolute refractory period. In addition, after the absolute refractory period there is a relative refractory period during which the neuron can fire again, but requires stronger input in order to do so. These refractory effects are important cases of *history dependence* in neural spike trains. To describe spike train variability accurately (at least for moderate to high firing rates where the refractory period is important), the probability of a spike occurring at a given time must depend on how recently the neuron has fired in the past. A more complicated history-dependent neural behavior is bursting, which is characterized by short sequences of spikes with small interspike intervals. In addition, spike trains are sometimes oscillatory. For example, neurons in the CA1 region of rodent hippocampus tend to fire at particular phases of the EEG theta rhythm. Thus,

in a variety of settings, probability models for spike trains make dependence on spiking history explicit.

Example 19.1 Retinal ganglion cell under constant conditions Neurons in the retina typically respond to patterns of light displayed over small sections of the visual field. When retinal neurons are grown in culture and held under constant light and environmental conditions, however, they will still spontaneously fire action potentials. In a fully functioning retina, this spontaneous activity is sometimes described as background firing activity, which is modulated as a function of visual stimuli. A short segment of the spiking activity from one neuron appeared in Fig. 16.1. A histogram of the ISIs appears in the left panel of Fig. 19.10. Even though this neuron is not responding to any explicit stimuli, we can still see structure in its firing activity. Although most of the ISIs are shorter than 20 ms, some are much longer: there is a small second mode in the histogram around 60–120 ms. This suggests that the neuron may experience two distinct states, one in which there are bursts of spikes (with short ISIs) and another, more quiescent state (with longer ISIs). From Fig. 16.1 we may also get an impression that there may be bursts of activity, with multiple spikes arriving in quick succession of one another. □

Example 19.2 Beta oscillations in Parkinson's disease Parkinson's disease, a chronic progressive neurological disorder, causes motor deficits leading to difficulty in movement. Clinical studies have shown that providing explicit visual cues, as guides, can improve movement in many patients, a possible explanation being that cortical drive associated with cues may lead to dampening of pathological beta oscillations (10–30 Hz) in the basal ganglia. To investigate this phenomenon, Sarma et al. (2012) recorded from neurons in the basal ganglia (specifically, the substantia nigra) while patients carried out a hand movement task. Because the period associated with a 20 Hz oscillation is 50 ms, if a neuron's activity is related to a beta oscillation it will tend to fire roughly every 50 ms. Therefore, its probability of firing at time t will be elevated if it fired previously 50 ms prior to time t. This is a form of history effect, which the authors built into their neural models in order to examine whether it was dampened due to visual cues. □

Example 19.3 Spatiotemporal correlations in visual signaling To better understand the role of correlation among retinal ganglion cells, Pillow et al. (2008) examined 27 simultaneously-recorded neurons from an isolated monkey retina during stimulation by binary white noise. The authors used a model having the form of (19.5). They concluded, first, that spike times appear more precise when the spiking behavior of coupled neighboring neurons is taken into account and, second, that in predicting (decoding) the stimulus from the spike trains, inclusion of the coupling term improved prediction by 20 % compared with a method that ignored coupling and instead assumed independence among the neurons. □

19.2 Poisson Processes

19.2.1 Poisson processes are point processes for which event probabilities do not depend on occurrence or timing of past events.

The discussion in Section 19.1.3 indicated the importance of history dependence in spike trains. On the other hand, a great simplification is achieved by ignoring history dependence and, instead, assuming the probability of spiking at a given time has no relationship with previous spiking behavior. This assumption leads to the class of *Poisson processes,* which are very appealing from a mathematical point of view: although they rarely furnish realistic models for data from individual spike trains, they are a pedagogical—and often practical—starting point for point processes in much the way that the normal distribution is for continuous random variables. As we shall see below, it is not hard to modify Poisson process models to make them more realistic.

Two kinds of Poisson processes must be distinguished. When event probabilities are invariant in time Poisson processes are called *homogeneous*; otherwise they are called *inhomogeneous*. We begin with the homogeneous case.

Definition: A homogeneous Poisson process with intensity λ is a point process satisfying the following conditions:

1. For any interval, $(t, t + \Delta t]$, $\Delta N_{(t,t+\Delta t]} \sim P(\mu)$ with $\mu = \lambda \Delta t$.
2. For any non-overlapping intervals, $(t_1, t_2]$ and $(t_3, t_4]$, $\Delta N_{(t_1,t_2]}$ and $\Delta N_{(t_3,t_4]}$ are independent.

For spike trains, the first condition states that for any time interval of length Δt, the spike count is a Poisson random variable with mean $\mu = \lambda \cdot \Delta t$. In particular, the mean, which is the expected number of spikes in the interval, increases in proportion to the length of the interval. Furthermore, the distribution of the spike count depends on the length of the interval, but not on its starting time: $\Delta N_{(t,t+h]}$ has the same distribution as $\Delta N_{(s,s+h]}$ for all positive values of s, t, h. This homogeneous process is *time-invariant*, and is said to have *stationary increments*. The second condition states that the spike counts (the counting process increments) from non-overlapping intervals are independent. In other words, the distribution of the number of spikes in an interval does not depend on the spiking activity outside that interval. Another way to state this definition is to say that a homogeneous Poisson process is a point process with stationary, independent increments.

A detail: There is one technical point to check: we need to be sure that the distributions of overlapping intervals, given in the definition

above, are consistent. For example, if we consider intervals (t_1, t_2) and (t_2, t_3) we must be sure that the Poisson distributions for the counts in each of these are consistent with the Poisson distribution for the count in the interval (t_1, t_3). Specifically, in this case, we must know that the sum of two independent Poisson random variables with means $\mu = \lambda(t_2 - t_1)$ and $\mu = \lambda(t_3 - t_2)$ is a Poisson random variable with mean $\mu = \lambda(t_3 - t_1)$. But this follows from the fact that if $W_1 \sim P(\mu_1)$ and $W_2 \sim P(\mu_2)$ independently, and we let $W = W_1 + W_2$, then $W \sim P(\mu_1 + \mu_2)$. We omit the details. □

We now come to an important characterization of homogeneous Poisson processes.

Theorem: A point process is a homogeneous Poisson process with intensity λ if and only if its inter-event waiting times are i.i.d. $Exp(\lambda)$.

Proof: We derive the waiting-time distribution for a homogeneous Poisson process. Recalling that X_i is the length of the inter-event interval between the $(i-1)^{\text{st}}$ and ith event times, we have $X_i > t$ precisely when $\Delta N_{(S_{i-1}, S_{i-1}+t]} = 0$. From the definition of a homogeneous Poisson process, $P\left(\Delta N_{(S_{i-1}, S_{i-1}+t]} = 0\right) = e^{-\lambda t}$. Therefore, the CDF of X_i is $F_{X_i}(t) = P(X_i \leq t) = 1 - e^{-\lambda t}$, which is the CDF of an $Exp(\lambda)$ random variable.
The converse of this theorem involves additional calculations and is omitted. □

Recall from Section 5.4.2 that the exponential distribution is memoryless. According to this theorem, for a homogeneous Poisson process, at any given moment the time at which the next event will occur does not depend on past events. Thus, the homogeneous Poisson process "has no memory" of past events.

Another way to think about homogeneous Poisson processes is that the event times are scattered "as irregularly as possible." One characterization of the "irregularity" notion is that, as noted on p. 120, the exponential distribution $Exp(\lambda)$ maximizes the entropy among all distributions on $(0, \infty)$ having mean $\mu = 1/\lambda$. Here is another.

Result: Suppose we observe $N(T) = n$ events from a homogeneous Poisson process on an interval $(0, T]$. Then the distribution of the event times is the same as that of a sample of size n from a uniform distribution on $(0, T]$.

Proof: This appears as a corollary to the theorem on p. 577, where it is also stated more precisely. □

Example 19.4 Miniature excitatory post-synaptic currents Figure 19.2 displays event times of miniature excitatory postsynaptic currents (MEPSCs) recorded from neurons in neonatal mice at multiple days of development. To record these events, the neurons are patch clamped at the cell body and treated so that they cannot propagate action potentials. These MEPSCs are thought to represent random activations of the dendritic arbors of the neuron at distinct spatial locations, so that the two assumptions of a Poisson process are plausible. The sequence of events in Fig. 19.2 looks highly

Fig. 19.2 A sequence of MEPSC event times. The inter-event intervals are highly irregular.

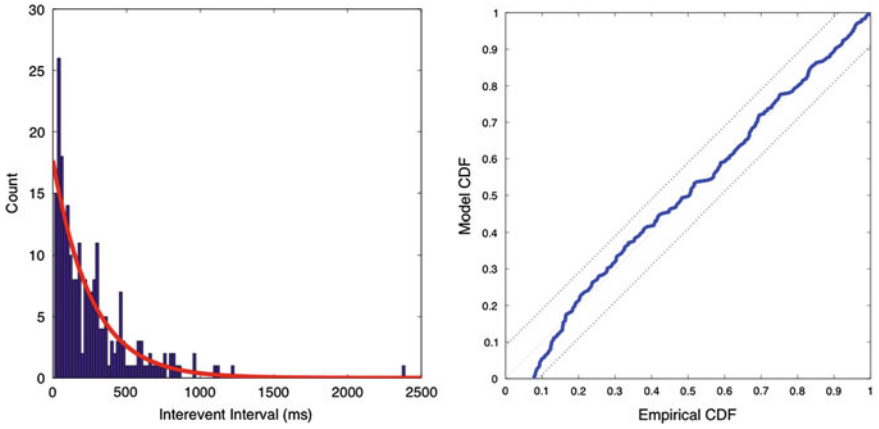

Fig. 19.3 Histogram and P–P plot of MEPSC inter-event intervals. *Left* Overlaid (in *red*) on the histogram is an exponential pdf. *Right* P–P plot falls within diagonal bands, indicating no lack of fit according to the Kolmogorov-Smirnov test (discussed in Section 19.3.5).

irregular, with no temporal structure. Figure 19.3 displays a histogram of the intervals between MEPSC events. The distribution of waiting times is captured well by an exponential fit, as shown both in left panel of Fig. 19.3 and in the P–P plot, in the right panel, which compares[2] the empirical CDF to that of an exponential. □

Important intuition may be gained by considering a discrete time representation of a sequence of event times, as discussed in Section 19.1.2. Suppose we have an observation interval $(0, T]$ and we consider partitioning $(0, T]$ into successive time bins of width Δt. If we make Δt sufficiently small we can force to nearly zero the probability of getting more than 1 event in any time bin. We then ignore the possibility of getting more than 1 event in any bin and, as in Section 19.1.2, we then let Y_i be the binary random variable that indicates whether an event has occurred in the ith time bin with $P(Y_i = 1) = p_i$, for $i = 1, \ldots, n$ (so that there are n time bins and $T = n\Delta t$). Each Y_i is a *Bernoulli* (p_i) random variable. If these Bernoulli random variables are homogeneous ($p_1 = p_2 = \cdots = p_n = p$ for some p) and independent, so that they form Bernoulli trials, then we have

1. For the ith time bin $(i\Delta t, (i+1)\Delta t]$, $\Delta N_{(i\Delta t,(i+1)\Delta t)} \sim$ *Bernoulli* (p).

[2] The small deviation of the curve from the diagonal in the lower left-hand corner of the P–P plot is probably due to inaccuracy of measurement for very short inter-event intervals.

2. For any two distinct time bins, $(i\Delta t, (i+1)\Delta t]$ and $(j\Delta t, (j+1)\Delta t]$, $\Delta N_{(i\Delta t,(i+1)\Delta t)}$ and $\Delta N_{(j\Delta t,(j+1)\Delta t)}$ are independent.

Let us now put $\lambda = p/\Delta t$ and use the Poisson approximation to the binomial distribution (see Section 5.2.2) as $\Delta t \to 0$. The two properties above then become essentially (for sufficiently small Δt) the same as the two properties in the definition of a Poisson process, given on p. 570. Therefore, leaving aside some mathematical details (see (19.11)), we may say that the sequence of Bernoulli trials converges to a Poisson process as $\Delta t \to 0$. That is, a homogeneous Poisson process is essentially a sequence of Bernoulli trials. We used this idea repeatedly in interpreting the Poisson distribution in Section 5.2. Rewriting $\mu = p/\Delta t$ as $p = \lambda \Delta t$ and replacing Δt with the infinitesimal dt we obtain the shorthand summary

$$P(\text{event in } (t, t+dt]) = \lambda dt. \tag{19.6}$$

We extend the fundamental connection between Bernoulli random variables and Poisson processes (and therefore also Poisson distributions) to the inhomogeneous case in Section 19.2.2.

19.2.2 Inhomogeneous Poisson processes have time-varying intensities.

We made two assumptions in defining a simple Poisson process: that the increments were (i) stationary, and (ii) independent for non-overlapping intervals. The first step in modeling a larger class of point processes is to eliminate the stationarity assumption. For spike trains, we would like to construct a class of models where the spike count distributions vary across time. In terms of the Bernoulli-trial approximation, we wish to allow the event probabilities p_i to differ.

Definition: An inhomogeneous Poisson process with intensity function $\lambda(t)$ is a point process satisfying the following conditions:

1. For any interval, $(t, t+\Delta t]$, $\Delta N_{(t,t+\Delta t]} \sim P(\mu)$ with $\mu = \int_{t_1}^{t_2} \lambda(t)dt$.
2. For any non-overlapping intervals, $(t_1, t_2]$ and $(t_3, t_4]$, $\Delta N_{(t_1,t_2]}$ and $\Delta N_{(t_3,t_4]}$ are independent.

This process is called an inhomogeneous Poisson process because it still has Poisson increments but each increment has its own mean, determined by the value of the rate function over the interval in question. The inhomogeneous Poisson process is no longer stationary, but its increments remain independent and, as a result, it retains the memoryless property, according to which the probability of spiking at any instant

does not depend on occurrences or timing of past spikes. In shorthand notation we modify (19.6) by writing

$$P(\text{event in } (t, t+dt]) = \lambda(t)dt. \tag{19.7}$$

At the beginning of the chapter we said that point process data are analyzed using the framework of generalized linear models, and in Section 19.1.2 we identified as a key step the representation of a point process as a binary time series, at least approximately. To take this step we need to equate, at least approximately, the point process likelihood function and the likelihood function for a suitable binary time series. In general, a likelihood function is proportional to the joint pdf of the data. Suppose we have observed event times $s_1, \ldots, s_n$. We assume these arise as observed values of random variables $S_1, \ldots, S_{N(T)}$, where $N(T)$ is the number of event times in $(0, T]$ and is itself a random variable. We write the joint pdf of $s_1, \ldots, s_n$ as $f_{S_1,\ldots,S_{N(T)}}(s_1, \ldots, s_n)$, where we acknowledge in our subscript notation[3] that $N(T)$ is also a random variable (taking the value $N(T) = n$ in data consisting of n events). Now suppose this joint pdf depends on some parameter vector θ. The likelihood function becomes

$$L(\theta) = f_{S_1,\ldots,S_{N(T)}}(s_1, \ldots, s_n|\theta). \tag{19.8}$$

In Example 14.5, for instance, we could consider the spike times to follow an inhomogeneous Poisson process and the parameter vector in (19.8) would consist of the parameters characterizing the spatial place cell distribution, $\theta = (\mu_x, \mu_y, \sigma_x, \sigma_y, \sigma_{xy})$. To get a formula for the likelihood function, the mathematical result we need is the formula for the joint pdf of the spike times. To be sure we get essentially the same likelihood function when we instead treat the spike train as a binary time series we also need a statement that the joint pdf of the spike times is approximately equal to the joint pdf for the binary time series. We provide both of these results below. We then also present an additional fact about inhomogeneous Poisson processes that aids intuition.

We begin with the joint pdf.

Theorem The event time sequence $S_1, S_2, \ldots, S_{N(T)}$ from a Poisson process with intensity function $\lambda(t)$ on an interval $(0, T]$ has joint pdf

$$f_{S_1,\ldots,S_{N(T)}}(s_1, \ldots, s_n) = \exp\left\{-\int_0^T \lambda(t)dt\right\} \prod_{i=1}^{n} \lambda(s_i). \tag{19.9}$$

Proof: See Section 19.4. □

[3] A more explicit notation would be $f_{S_1,\ldots,S_{N(T)},N(T)}(S_1 = s_1, \ldots, S_{N(T)} = s_n, N(T) = n)$, see p. 577, where we make explicit the randomness due to $N(T)$.

We now turn to our ability to treat an inhomogeneous Poisson process as if it were approximately the same as a binary time series described in Section 19.1.2, with

$$P(\text{event in } (t, t+\Delta t]) \approx \lambda(t)\Delta t. \tag{19.10}$$

We give a rigorous statement that the joint pdf of the spike times is approximately equal to the joint pdf for the corresponding binary time series. More specifically, we show that the joint pdf in Eq. (19.9) is the limit of relevant binary pdfs as $\Delta t \to 0$.

Let us consider a set of points $s_1, \ldots, s_n$ in the interval $(0, T]$ that, while conceptually representing event times, are for the purposes of the analysis below, taken to be fixed. They represent the observed data. We will call them "atoms" because they are points where probability mass will be placed. Suppose $(0, T]$ is decomposed into N subintervals of length Δt, so that $\Delta t = T/N$. For $i = 1, \ldots, N$ let $x_i = 1$ if the ith subinterval contains one of the atoms and 0 otherwise.

Theorem Let $\lambda(t)$ be a continuous function on $[0, T]$, set $\lambda_i = \lambda(t_i)$ for subinterval midpoints t_i, and let $p_i = (\Delta t)\lambda_i$. Then as $\Delta t \to 0$ we have

$$\frac{1}{(\Delta t)^n} \prod_{i=1}^{n} p_i^{x_i}(1-p_i)^{1-x_i} \to e^{-\int_0^T \lambda(t)dt} \prod_{i=1}^{n} \lambda(s_i). \tag{19.11}$$

To prove this result we need two lemmas. Let $S = S_n$ be the set of i indices for which $x_i = 1$ and S^c the set of indices for which $x_i = 0$.

Lemma 1 As $\Delta t \to 0$ we have

$$\prod_{S} \lambda(t_i) \to \prod_{i=1}^{n} \lambda(s_i).$$

Proof: The lemma follows immediately from continuity of $\lambda(t)$. □

Lemma 2 As $\Delta t \to 0$ we have

$$\sum_{S^c} \log(1 - (\Delta t)\lambda_i) \to -\int_0^T \lambda(t)dt.$$

Proof: This follows immediately from a first-order Taylor series expansion of the log (Equation (A.5)), together with the definition of the integral as a limit[4] of sums. □

Proof of the theorem: Putting the two lemmas together we easily prove the theorem. We have

[4] The limit of the sum over S^c is the same as the limit of the sum over $S \cup S^c$ because S has n elements for all sufficiently small values of Δt, so that $\lim \sum_S \Delta t\lambda_i = 0$.

$$\begin{aligned}\frac{1}{(\Delta t)^n}\prod_{i=1}^{N} p_i^{x_i}(1-p_i)^{1-x_i} &= \frac{1}{(\Delta t)^n}\Big(\prod_{S}(\Delta t)\lambda_i\Big)\Big(\prod_{S^c} 1-(\Delta t)\lambda_i\Big)\\ &= \Big(\prod_{S}\lambda_i\Big)e^{\sum_{S^c}\log(1-(\Delta t)\lambda_i)}\\ &\to e^{-\int_0^T \lambda(t)dt}\prod_{i=1}^{n}\lambda(s_i).\end{aligned}$$ □

To recap: taken together, the two theorems above show that the inhomogeneous Poisson process spike time joint pdf is approximately equal to a binary time series joint pdf, which allows us to use the binary random variables Y_i (with $p_i = P(Y_i = 1)$) defined in Section 19.1.2 in place of the Poisson process. The memorylessness of the Poisson process translates into independence among the Y_is. However, the values of p_i may vary across time, corresponding to the inhomogeneity of the process. Importantly, we may estimate $\lambda(t)$ by likelihood methods, applying Poisson regression with suitably small time bins (e.g., having width 1 ms).

Example 1.1 (continued) In Chapter 1 we introduced the SEF neuron example, the problem being to characterize the neural response under two different experimental conditions. In Chapter 8 we returned to the example to describe the benefit of smoothing the PSTH, and in Chapter 15, p. 422, we showed how smoothing may be accomplished using Poisson regression splines. The smoothing model was

$$Y_i \sim P(\lambda_i) \tag{19.12}$$
$$\log \lambda_i = f(t_i) \tag{19.13}$$

where t_i was the time at the midpoint of the ith time bin (of the PSTH), Y_i was the corresponding spike count in that bin, and $f(t)$ was taken to be a natural cubic spline with two knots at specified locations.

An inhomogeneous Poisson process model may be constructed that is very similar to the PSTH-based regression model. To get a Poisson process model we must take the time bins to be smaller—small enough that on any trial there is at most one spike in any bin. For instance, we may take the bins to have width 1 ms. Then, we must define the resulting binary counts: for trial r let Y_{ri} be 1 if a spike occurs in the ith bin and 0 otherwise. We write the model

$$Y_{ri} \sim P(\lambda_i) \tag{19.14}$$
$$\log \lambda_i = f(t_i) \tag{19.15}$$

where, again, $f(t)$ is a natural cubic spline with two knots at the locations specified previously. Comparing (19.14) and (19.15) with (19.12) and (19.13) we have a model of almost the same form. Aside from the width of the time bins, the distinction is that (19.14) and (19.15) is a within-trial model, in terms of Y_{ri}, while (19.12) and (19.13) is a model that pools events across trials by using the PSTH spike counts Y_i. It turns

out that the intensity that results from fitting (19.14) and (19.15) is nearly identical to the fit of $f(t)$ resulting from (19.12) and (19.13). The closeness of results holds quite generally because the smoothing of the PSTH is not very sensitive to the choice of bin widths as long as the firing rate varies slowly enough to be nearly constant within bins. Smoothing the PSTH amounts to fitting a Poisson process after jittering all the spike times within a bin so that they are equal to the midpoint of that bin. □

The final theorem of this section gives another interesting way to think about inhomogeneous Poisson processes. Let us begin by considering the PSTH, as used in Examples 1.1 and 15.1. The PSTH is the peristimulus time *histogram*. But in what sense is it a histogram? A histogram is a plot that displays counts, as does the PSTH, but the counts are presumed to be repeated observations from a random variable, and the histogram is supposed to be a rough estimate of the random variable's pdf. What are the repeated observations that generate the PSTH? And what pdf is it estimating? The data are the event times. But, as we have already taken pains to point out, these event times are not i.i.d. observations from a fixed distribution: they follow a point process, which is different. How are they transformed into i.i.d. observations that are suitable for making a histogram and estimating a pdf? While these questions are puzzling at first, the answer turns out to be simple. According to the next theorem, given some number n of events in an interval $(0, T]$, the event times will be scattered across $(0, T]$ as if they were i.i.d. observations from a distribution having as its pdf the normalized intensity $\lambda(t)$. In other words, the positions of the event times are just like i.i.d. observations; therefore, the PSTH is just like a histogram, and could be treated as if it were an estimator of the normalized intensity function.

To state the result, let us first recall that the length of the sequence of event times $S_1, S_2, \ldots, S_{N(T)}$ depends on the random quantity $N(T)$. Thus, to be more thorough we might write the joint pdf above in the form

$$f_{S_1,\ldots,S_{N(T)}}(s_1, \ldots, s_n) = f_{S_1,\ldots,S_{N(T)},N(T)}(S_1 = s_1, \ldots, S_{N(T)} = s_n, N(T) = n).$$

That is, the pdf on the left-hand side is really a short-hand notation for the pdf on the right-hand side. This observation is used in the proof of the following theorem. We will write $f_N(n)$ for the pdf of $N(T)$ and note that, for a Poisson process with intensity $\lambda(t)$, $N(T) \sim P(\mu)$ with $\mu = \int_0^T \lambda(t)dt$.

Theorem Let $S_1, S_2, \ldots, S_{N(T)}$ be an event sequence from a Poisson process with intensity function $\lambda(t)$ on an interval $(0, T]$. Conditionally on $N(T) = n$, the sequence $S_1, S_2, \ldots, S_n$, has the same joint distribution as an ordered set of i.i.d. observations from a univariate distribution having pdf

$$g(t) = \frac{\lambda(t)}{\int_0^T \lambda(u)du}.$$

Proof: We write the conditional pdf as

$$
\begin{aligned}
f_{S_1,\ldots,S_{N(T)}}(s_1,\ldots,s_n|N(T)=n) &= \frac{f_{S_1,\ldots,S_{N(T)}}(s_1,\ldots,s_n)}{f_N(n)} \\
&= \frac{e^{-\int_0^T \lambda(t)dt}\prod_{i=1}^n \lambda(s_i)}{e^{-\int_0^T \lambda(t)dt}\frac{\left(\int_0^T \lambda(t)dt\right)^n}{n!}} \\
&= n!\prod_{i=1}^n \frac{\lambda(s_i)}{\int_0^T \lambda(t)dt} \\
&= n!\prod_{i=1}^n g(s_i).
\end{aligned}
$$

Noting that there are $n!$ ways to order the observations $s_1,\ldots,s_n$, this completes the proof. □

The theorem says that we may consider an inhomogeneous Poisson process with intensity $\lambda(t)$ to be equivalent to a two-stage process in which we (1) generate an observation $N=n$ from a Poisson distribution with mean $\mu=\int_0^T \lambda(t)dt$; this tells us how many events are in $(0,T]$; we then (2) generate n i.i.d. observations from a distribution having $g(t)=\lambda(t)/\int_0^T \lambda(u)du$ as its pdf. We motivated the theorem by suggesting that it shows how the PSTH acts like a histogram: the intensity function $\lambda(t)$ describes the event times that come from pooling together all the spike times across all of the trials; the PSTH then estimates $\lambda(t)/\int_0^T \lambda(u)du$. Not only does this explain the sense in which the PSTH is actually a histogram, it also motivates application of a density estimator (e.g., a normal kernel density estimator or Gaussian filter), as in Section 15.4, to smooth the PSTH.

When we specialize the theorem above to homogeneous Poisson processes we get, as a corollary, the result stated as a theorem on p. 571.

Corollary Let $S_1,S_2,\ldots,S_{N(T)}$ be an event sequence from a homogeneous Poisson process with intensity λ on an interval $(0,T]$. Conditionally on $N(T)=n$, the sequence $S_1,S_2,\ldots,S_n$, has the same joint distribution as an ordered set of i.i.d. observations from a uniform distribution on $[0,T]$.

Proof: This is a special case of the theorem in which $\lambda(t)=\lambda$ so that $g(t)=1/T$, i.e., $g(t)$ is the pdf of the uniform distribution on $(0,T]$. □

19.3 Non-Poisson Point Processes

19.3.1 Renewal processes have i.i.d. inter-event waiting times.

The homogeneous Poisson process developed in Section 19.2.1 assumed that the point process increments were both stationary and independent of past event history.

To accommodate event probabilities that change across time, we generalized from homogeneous to inhomogeneous Poisson processes. This eliminated the assumption of stationary increments but it preserved the independence assumption, which entailed history independence. Systems that produce point process data, however, typically have physical mechanisms that lead to history-dependent variation among the events, which cannot be explained with Poisson models. Therefore, it is necessary to further generalize by removing the independence assumption.

The simplest kind of history-dependent behavior occurs when the probability of the ith event depends on the occurrence time of the previous event s_{i-1}, but not on any events prior to that. If the ith waiting time X_i is no longer memoryless, then $P(X_i > t + h | X_i > t)$ may not be equal to $P(X_i > u + h | X_i > u)$ when $u \neq t$, but X_i is independent of event times prior to S_{i-1}, and is therefore independent of all waiting times X_j for $j < i$. Thus, the waiting time random variables are all mutually independent. In the time-homogeneous case, they also all have the same distribution. A point process with i.i.d waiting times is called a *renewal process*. We already saw that homogeneous Poisson processes have i.i.d. exponential waiting times. Therefore, renewal processes may be considered generalizations of homogeneous Poisson processes.

A renewal model is specified by the distribution of the inter-event waiting times. Typically, this takes the form of a probability density function, $f_{X_i}(x_i)$, where x_i can take values in $[0, \infty)$. In principle we can define a renewal process using any probability distribution that takes on positive values, but there are some classes of probability models that are more commonly used either because of their distributional properties, or because of some physical or physiological features of the underlying process.

For example, the gamma distribution, which generalizes the exponential, may be used when one wants to describe interspike interval distributions using two parameters: the gamma shape parameter gives it flexibility to capture a number of characteristics that are often observed in point process data. If this shape parameter is equal to one, then the gamma distribution simplifies to an exponential, which as we have shown, is the ISI distribution of a simple Poisson process. Therefore, renewal models based on the gamma distribution generalize simple Poisson processes, and can be used to address questions about whether data are actually Poisson. If the shape parameter is less than one, then the density drops off faster than an exponential. This can provide a rough description of ISIs when a neuron fires in rapid bursts. If the shape parameter is greater than one, then the gamma density function takes on the value zero at $x_i = 0$, rises to a maximum value at some positive value of x_i, and then falls back to zero. This can describe the ISIs for a relatively regular spike train, such as those from a neuron having oscillatory input. Thus, this very simple class of distributions with only two parameters is capable of capturing, at least roughly, some interesting types of history dependent structure.

While the gamma distribution is simple and flexible, it doesn't have any direct connection with the physiology of neurons. For neural spiking data, a renewal model with a stronger theoretical foundation is the inverse Gaussian. As described in Section 5.4.6, the inverse Gaussian also has two parameters and is motivated by

the integrate-and-fire conception of neural spiking behavior. Thus, a renewal process with inverse Gaussian ISIs would be a simple yet natural model for neural activity in a steady state.

One way to quantify the regularity of a renewal process is through the ISI coefficient of variation. We noted in (3.14) that exponentially-distributed random variables have CV $= 1$, so this corresponds to a Poisson process. When CV < 1 the process is more regular than Poisson (as would be a spike train from an oscillatory neuron), while when CV > 1 the process is more irregular than Poisson (as would be a spike train from a bursty neuron). This regularity or irregularity of a renewal process will also be apparent in the distribution of counts and is often measured by the Fano factor,

$$F(t, t + \Delta t) = \frac{V(\Delta N_{(t,t+\Delta t]})}{E(\Delta N_{(t,t+\Delta t]})}.$$

For a Poisson process we have $F(t, t + \Delta t) = 1$. The counts will be relatively less dispersed for regular renewal processes, so that $F(t, t+\Delta t) < 1$, and more dispersed for irregular processes, so that $F(t, t + \Delta t) > 1$.

A general result that has implications for spike train analysis is the *renewal theorem*, which[5] examines the expected number of events in an interval $(t, t + h]$ as $t \to \infty$. For a Poisson process with intensity λ we have $E(\Delta N_{(t,t+h]}) = \lambda h$, and the waiting time distribution is exponential with mean $\mu = 1/\lambda$. In other words, the expected number of events in $(t, t + h]$ is $\lambda h = h/\mu$, so that the expected number of events is just the length of the interval divided by the average waiting time for an event. For a renewal process the same statement is approximately true for large t.

Renewal Theorem Suppose a renewal process has waiting times with a continuous pdf and a mean μ. Defining $\lambda = 1/\mu$ we have

$$\lim_{t\to\infty} E(\Delta N_{(t,t+h]}) = \lambda h.$$

Proof: Omitted. □

Notice that if we take h sufficiently small in the renewal theorem, the count $\Delta N_{(t,t+h]}$ will, with high probability, be either 0 or 1 and then its expectation is $E(\Delta N_{(t,t+h]}) = P(\Delta N_{(t,t+h]} = 1)$. Thus, if we pick a large t and ask for the probability of an event in the infinitesimal interval $(t, t + dt]$ by ignoring the time of the most recent event and instead letting the renewal process start at time 0 and run until we get to time t, we find that (19.6) continues hold.

A related result arises when we consider what happens when we combine multiple renewal processes by pooling together all their event times. This sort of pooling occurs, for example, in a PSTH when multiple spike trains are collected across multiple trials: in making the PSTH every spike time is used but the trial on which it occurred is ignored. Such combination of point processes is called *superposition*. Specifically, if we have counting processes $N^i(t)$, for $i = 1, \ldots, n$ then

[5] A more general version of this result is often called *Blackwell's Theorem*.

$N(t) = \sum_{i=1}^{n} N^i(t)$ is the process resulting from superposition. First, we consider the Poisson case.

Theorem For $i = 1, \ldots, n$, let $N^i(t)$ be the counting process representation of a homogeneous Poisson process having intensity λ_i. Then the point process specified by $N(t) = \sum_{i=1}^{n} N^i(t)$ is a homogeneous Poisson process having intensity $\lambda = \sum_{i=1}^{n} \lambda_i$.

Sketch of Proof: Because the sum of independent Poisson random variables is Poisson, condition 1 of the definition of a homogenous Poisson process is satisfied for the superposition process. Because condition 2 is satisfied for all n independent processes, it is also satisfied for the superposition process. □

Result The superposition of a large number of independent renewal processes having waiting times with continuous pdfs and finite means is, approximately, a Poisson process.

Proof: The mathematics involved in stating this result precisely are rather intricate. We omit the proof, but offer the following heuristics to make the result plausible.

Suppose that the n independent renewal processes have mean waiting times $\mu_i = 1/\lambda_i$, for $i = 1, \ldots, n$. Let us consider intervals $(t, t + h]$, with h so small that, with large probability, across all n processes at most 1 event occurs. Then the superposition increments $\Delta N_{(t,t+h]}$ are essentially binary variables. For the superposition to be Poisson, these binary variables must be homogeneous and independent. By the renewal theorem, for large t,

$$P(\Delta N^i_{(t,t+h]} = 1) \approx \lambda_i h,$$

where $\lambda_i = 1/\mu_i$ and

$$P(\Delta N^i_{(t,t+h]} = 0) \approx 1 - \lambda_i h.$$

When we pool all the processes together, the event $\Delta N_{(t,t+h]} = 1$ will occur if at least one process has an event, and otherwise $\Delta N_{(t,t+h]} = 0$, which has probability

$$P(\Delta N_{(t,t+h]}{=}0) \approx (1-\lambda_1 h)(1-\lambda_2 h)\cdots(1-\lambda_n h) \approx e^{-\lambda t} \approx 1-\lambda h$$

and this, in turn, shows that

$$P(\Delta N_{(t,t+h]} = 1) \approx \lambda h,$$

as for a Poisson process, so that homogeneity holds, approximately. As far as independence is concerned, the key point is that the renewal processes are independent of one another, so that the only dependence in the superposition is due to events from the same process, which are very rare among the large numbers of events in the superposition process. That is, if we assume n is so large that, for all k, $P(\Delta N_{(t,t+h]} = 1) >> P(\Delta N^k_{(t,t+h]} = 1)$, then when we consider two non-overlapping intervals $(t_1, t_1 + h]$ and $(t_2, t_2 + h]$, relative to the superposition process, the probability that the kth process has events in both intervals is negligible. This is another way of saying that the identity of events in the superposition gets washed out as the number of processes increases. □

By combining this superposition result and the renewal theorem we obtain a practical implication: the superposition of multiple renewal processes will be approximately a Poisson process, but we can expect the approximation to be better for large t, after initial conditions die out. If, for example, we take multiple spike trains, and if time $t = 0$ has a physiological meaning related to the conditions of the experiment, then we may expect the initial conditions to affect the spike trains in a reproducible way from trial to trial so that even after pooling we might see non-Poisson behavior near the beginning of the trial; as such effects dissipate across time we would expect the pooled spike trains to exhibit Poisson-process-like variation.

19.3.2 The conditional intensity function specifies the joint probability density of spike times for a general point process.

In Section 19.2.2 we described the structure of an inhomogeneous Poisson process in terms of an intensity function that characterized the instantaneous probability of firing a spike at each instant in time, as in (19.6). In an analogous way, a general point process may be characterized by its *conditional intensity function*. Poisson processes are memoryless but, in general, if we want to find the probability of an event in a time interval $(t, t + \Delta t]$ we must consider the timing of the events preceding time t. Let us denote the number of events prior to t by $N(t-)$,

$$N(t-) = max_{u<t} N(u).$$

We call the sequence of event times prior to time t the *history* up to time t and write it as $H_t = (S_1, S_2, \ldots, S_{N(t-)})$. For a set of observed data we would write $H_t = (s_1, s_2, \ldots, s_n)$ with the understanding that $N(t-) = n$. The conditional intensity function is then given by

$$\lambda(t|H_t) = \lim_{\Delta t \to 0} \frac{P(\Delta N_{(t,t+\Delta t]} = 1|H_t)}{\Delta t}, \tag{19.16}$$

where $P(\Delta N_{(t,t+\Delta t]} = 1|H_t)$ is the conditional probability of an event in $(t, t + \Delta t]$ given the history H_t. Taking Δt to be small we may rewrite Eq. (19.16) in the form

$$P(\Delta N_{(t,t+\Delta t]} = 1|H_t) \approx \lambda(t|H_t)\Delta t. \tag{19.17}$$

Or, in shorthand,

$$P(\text{event in } (t, t + dt]|H_t) = \lambda(t|H_t)dt, \tag{19.18}$$

which generalizes (19.6). According to (19.18) the conditional intensity function expresses the instantaneous probability of an event. It serves as the fundamental building block for constructing the probability distributions needed for general point processes.[6] A mathematical assumption needed for theoretical constructions is that the point process is *orderly*, which means that for a sufficiently small interval, the probability of more than one event occurring is negligible. Mathematically, this is stated as

$$\lim_{\Delta t \to 0} \frac{P(\Delta N_{(t,t+\Delta t]} > 1|H_t)}{\Delta t} = 0. \tag{19.19}$$

This assumption is biophysically plausible for a point process model of a neuron because neurons have an absolute refractory period. In most situations, the probability of a neuron firing more than one spike is negligibly small for $\Delta t < 1$ ms.

Once we specify the conditional intensity for a point process, it is not hard to write down the pdf for the sequence of event times in an observation interval $(0, T]$. In fact, the argument is essentially the same as in the case of the inhomogeneous Poisson process, with the conditional intensity $\lambda(t|H_t)$ substituted for the intensity $\lambda(t)$. The key observation is that the conditional intensity behaves essentially like a hazard function, the only distinction being the appearance of the stochastic history H_t.

Theorem The event time sequence $S_1, S_2, \ldots, S_{N(T)}$ of an orderly point process on an interval $(0, T]$ has joint pdf

$$f_{S_1,\ldots,S_{N(T)}}(s_1, \ldots, s_n) = \exp\left\{-\int_0^T \lambda(t|H_t)dt\right\} \prod_{i=1}^{n} \lambda(s_i|H_{s_i}) \tag{19.20}$$

where $\lambda(t|H_t)$ is the conditional intensity function of the process.

---

[6] Because the history $H_t = (S_1, S_2, \ldots, S_{N(t-)})$ is itself a point process, it is stochastic and, therefore, the conditional intensity is stochastic. The definition (19.18) includes two separable steps: first, we define the conditional intensity

$$\lambda(t|s_1, \ldots, s_n) = \lim_{\Delta t \to 0} \frac{P(\Delta N_{(t,t+\Delta t]} = 1|N(t-) = n, S_1 = s_1, \ldots, S_n = s_n)}{\Delta t}$$

for every possible vector $(s_1, \ldots, s_n)$ making up the history H_t, and then we replace the specific values $N(t-) = n$ and $(S_1 = s_1, \ldots, S_n = s_n)$ with their stochastic counterparts written as $H_t = (S_1, S_2, \ldots, S_{N(t-)})$.

Proof: See Section 19.4. □

Equation (19.20) has the same form as (19.9), the only distinction being the replacement of the Poisson intensity $\lambda(t)$ in (19.9) with the conditional intensity $\lambda(t|H_t)$ in (19.20).

We may also approximate a general point process by a binary process. For small Δt, the probability of an event in an interval $(t, t + \Delta t]$

$$P(\text{event in } (t, t + \Delta t]|H_t) \approx \lambda(t|H_t)\Delta t \tag{19.21}$$

and the probability of no event is

$$P(\text{no event in } (t, t + \Delta t]|H_t) \approx 1 - \lambda(t|H_t)\Delta t. \tag{19.22}$$

Equation (19.21) generalizes (19.10). If we consider the discrete approximation, analogous to the Poisson process case, we may define $p_i = \int \lambda(t|H_t)dt$ where the integral is over the ith time bin. We again get Bernoulli random variables Y_i with $P(Y_i = 1) = p_i$ but now these Y_i random variables are *dependent*, e.g., we may have $P(Y_i = 1|Y_{i-1} = 1) \neq p_i$. The theorem giving (19.11) holds again when we replace $\lambda(t)$ with $\lambda(t|H_t)$. In practice, spike train analyses using dependent binary variables are a little more complicated than those using independent binary variables, but it remains relatively easy to formulate history-dependent models for these dependent variables by following a regression strategy that is very similar to that used previously, on p. 576. We give examples in Section 19.3.4.

19.3.3 The marginal intensity is the expectation of the conditional intensity.

Equation (19.16) gave the definition of the conditional intensity function. We now define the unconditional or *marginal intensity function* as

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(\Delta N_{(t,t+\Delta t]} = 1)}{\Delta t}. \tag{19.23}$$

Definition (19.23) may be rewritten in some informative ways. First, note that if X is a binary random variable its expectation is $E(X) = P(X = 1)$, as in (15.2). For Δt sufficiently small, $\Delta N_{(t,t+\Delta t]}$ is a binary random variable so that (19.23) may be written

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{E(\Delta N_{(t,t+\Delta t]})}{\Delta t}. \tag{19.24}$$

That is, the marginal intensity is the expected spike count density.

Next, according to the law of total probability (p. 86), for a pair of random variables Y and X and an event A we have $P(X \in A) = E_Y(P(X \in A|Y))$. Letting H_t play the role of Y and $\Delta N_{(t,t+\Delta t]} = 1$ the role of $X \in A$, we get, similarly,

$$P(\Delta N_{(t,t+\Delta t]} = 1) = E_{H_t}\left(P(\Delta N_{(t,t+\Delta t]} = 1|H_t)\right)$$

and

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{E_{H_t}\left(P(\Delta N_{(t,t+\Delta t]} = 1|H_t)\right)}{\Delta t}.$$

By interchanging[7] the expectation and limiting operation we may then write

$$\lambda(t) = E_{H_t}(\lambda(t|H_t)). \tag{19.25}$$

Equation (19.25) explains the name "marginal" intensity. The intensity $\lambda(t)$ is marginal in much the same sense as when we have a pair of random variables (X, Y) and speak of the distribution of X as a marginal distribution because it is derived by averaging over all possible values of Y. Here, $\lambda(t)$ results from averaging the conditional intensity over all possible histories H_t. In the case of spike trains, the conditional intensity would apply to individual trials, while the marginal intensity would be the theoretical time-varying firing rate after averaging across trials. Importantly, we may consider $\lambda(t)$ to be the function being estimated by the PSTH. This does not require us to assume the trials are in any sense all the same. There could be some source of trial-to-trial variation, or even systematic variation (such as effects associated with learning across trials). Consideration of $\lambda(t)$ takes place whenever the average across trials seems meaningful and interesting.

As in Eq. (19.17) we may also write

$$P(\Delta N_{(t,t+\Delta t]} = 1) \approx \lambda(t)\Delta t \tag{19.26}$$

and we have the shorthand

$$P(\text{event in } (t, t + dt]) = \lambda(t)dt, \tag{19.27}$$

keeping in mind that we also take the left-hand side to mean

$$P(\text{event in } (t, t + dt]) = E_{H_t}P(\text{event in } (t, t + dt]|H_t).$$

Equation (19.27) must be compared with (19.18) and, of course, it has the same form as (19.6). We may therefore think of the average across histories (for spike trains, the average across trials) as defining a theoretical inhomogeneous Poisson process intensity. This is the intensity that is estimated by the PSTH.

[7] General theory justifying the interchange of limit and expectation applies here.

The distinction between conditional and marginal intensities is so important for spike train analysis that we emphasize it, as follows.

If we consider spike trains to be point processes, within trials the instantaneous firing rate is $\lambda(t|H_t)$ and we have

$$P(\text{spike in } (t, t+dt]|H_t) = \lambda(t|H_t)dt,$$

while the across-trial average firing rate is $\lambda(t)$ and we have

$$P(\text{spike in } (t, t+dt]) = \lambda(t)dt.$$

19.3.4 Conditional intensity functions may be fitted using Poisson regression.

On p. 576 we discussed the way Poisson regression may be used to fit inhomogeneous Poisson process models. The key theoretical result that made this possible was Eq. (19.11) in conjunction with (19.10). As we said on p. 584, that theorem holds again for conditional intensity functions using Eq. (19.21). This means that Poisson regression can again be used for non-Poisson point processes.

We now give some examples in which conditional intensity functions have been fitted to spike train data.

Example 19.1 (continued from p. 569) Let us take time bins to have width $\Delta t = 1$ ms and write $\lambda_k = \lambda(t_k|H_{t_k})$, where t_k is the midpoint of the kth time bin. Defining

$$\log \lambda_k = \alpha_0 + \sum_{j=1}^{120} \alpha_j \Delta N_{(k-j-1,k-j]}, \tag{19.28}$$

we get a model with 120 history-related explanatory variables, each indicating whether or not a spike was fired in a 1 ms interval at a different time lag. The parameter α_0 provides the log background firing rate in the absence of prior spiking activity within the past 121 ms. Using Poisson regression with ML estimation (as in Section 14.1) we obtained $\hat{\alpha}_0 = 3.8$ so that, if there were no spikes in the previous 121 ms, the conditional intensity would become $\lambda_k = \exp(\hat{\alpha}_0) = 45$ spikes per second, corresponding to an average ISI of 22 ms. The MLEs $\hat{\alpha}_i$ obtained from the data are plotted in Fig. 19.4, in the form $\exp\{\hat{\alpha}_i\}$. The $\hat{\alpha}_i$ values related to 0–2 ms after a spike are large negative numbers, so that $\exp\{\hat{\alpha}_i\}$ is close to zero, leading to a refractory period when the neuron is much less likely to fire immediately after

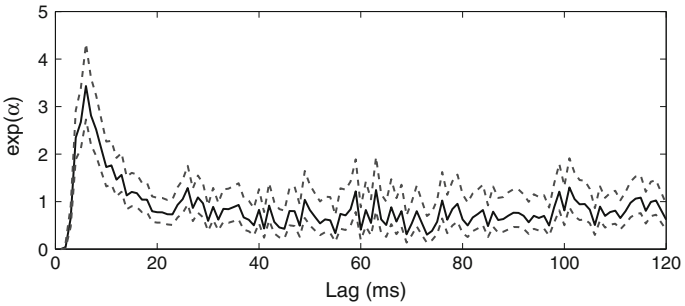

Fig. 19.4 Parameter estimates for history-dependent retinal conditional intensity model (*bold line*) together with confidence intervals (*dotted line*), which indicate uncertainty in the estimates (based on maximum likelihood). The x-axis indicates the lag time in milliseconds.

another spike. However, the estimates related to 4–13 ms after a spike are substantially positive, leading to an increase in the firing probability. For example, if the only spike in the 120 ms history occurred 6 ms in the past, then the background conditional intensity of 45 spikes per seconds is multiplied by a factor of about 3.1, leading to a conditional intensity of 140 spikes per second. This phenomenon accounts for the rapid bursts of spikes observed in the data. (The same data were discussed in the context of burst detection in Example 16.3 on p. 458.) Many of the remaining parameters are close to zero, and hence $\exp\{\hat{\alpha}_i\}$ is close to one, indicating that the corresponding history term has no effect on the spiking probability. Figure 19.5 displays the ISI histogram with exponential and Inverse Gaussian renewal model pdfs overlaid, and also the pdf for the model of Eq. (19.28). The exponential model overestimates the number of very short ISIs (0–4 ms), and both renewal models underestimate the number of ISIs between 5–10 ms and overestimate the number of ISIs between 10–60 ms. In contrast, the conditional intensity model in Eq. (19.28) accurately predicts the number of ISIs across all ISI lengths. □

Example 19.2 (continued) On p. 569 we said that a beta oscillation at 20 Hz could be represented in the history effects as an elevated probability of firing at time t when the neuron fired previously 50 ms prior to time t. Using Eq. (19.28) this would be represented by positive α_j coefficients around $j = 50$. Sarma et al. reduced the number of parameters, replacing (19.28) with

$$\log \lambda_k = \alpha_0 + \sum_{j=1}^{10} \alpha_j \Delta N_{k-j} + \sum_{i=1}^{14} \gamma_i \Delta N_{(k-(10i+9),k-10i]}. \tag{19.29}$$

In this version of the model, when γ_i is positive there is an increase in the log probability of firing when the neuron previously fired in the interval from $10i$ to $10i + 9$ ms in the past. Thus, the presence of a beta oscillation would produce a positive coefficient γ_5 (corresponding to 50–59 ms in the past, or 17–20 Hz). An example of a neuron having a positive γ_5 coefficient was given by the authors, reproduced here

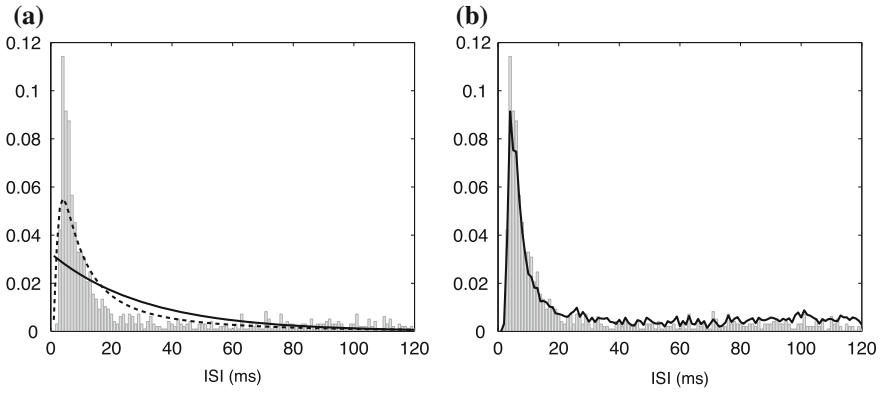

Fig. 19.5 ISI histogram and fitted pdfs. Panel **a:** ISI histogram overlaid with pdfs from exponential (*solid line*) and inverse Gaussian (*dashed line*) renewal models. Panel **b:** ISI histogram overlaid with pdf (*solid line*) from model defined by Eq. (19.28).

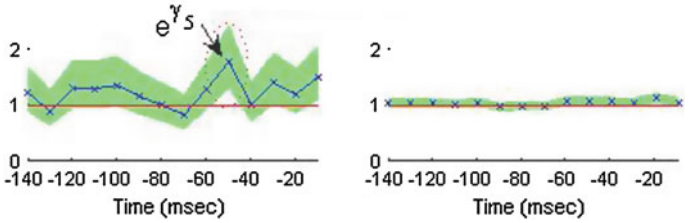

Fig. 19.6 Plots of γ coefficients using model (19.29) for a neuron recorded from the substantia nigra for a cued hand movement. *Left* coefficients before initiation of movement. *Right* coefficients after initiation of movement. Adapted from Sarma et al.

in Fig. 19.6. Results before and after movement initiation are shown in Fig. 19.6, when an explicit visual cue showed the subject where to move. In this case there was a dampening of beta oscillations during movement. The authors decomposed the timing of beta oscillations further and found that, among many substantia nigra cells, there was evidence of decreased beta oscillation beginning immediately following illumination of the visual cue. Based on additional results they suggested that execution of a motor plan following a cue may be suppressing pathological activity in the substantia nigra, which may explain improved task performance. □

A second way to introduce history dependence is to begin with the hazard function of a renewal process and then modify the conditional intensity so that it can vary across time. This extends to renewal processes the method used for allowing Poisson processes to become inhomogeneous. In a homogeneous Poisson process, the waiting times are not only i.i.d., they are also memoryless: the probability of an event does not depend on when the last event occurred. To get an inhomogeneous Poisson process, we retain the memorylessness but introduce a time-varying conditional intensity. A simple idea is to take a renewal process and, similarly, introduce a time-varying

factor. For a renewal process, the probability of an event at time t depends on the timing of the most recent previous event $s_*(t)$, but not on any events prior to $s_*(t)$. If we allow the conditional intensity to depend on both time t and the time of the previous event $s_*(t)$ we obtain a form

$$\lambda(t|H_t) = g(t, s_*(t)) \tag{19.30}$$

where $g(x, y)$ is a function to be specified. Models of this type are sometimes called *Markovian* or *Inhomogeneous Markov Interval* (IMI) models.[8] In an inhomogeneous Poisson process the conditional intensity takes the form

$$\lambda(t|H_t) = g_0(t)$$

where $g_0(t)$ becomes the intensity $\lambda(t)$. In a renewal process the conditional intensity takes the form

$$\lambda(t|H_t) = g_1(t - s_*(t))$$

where $g_1(t - s_*(t))$ becomes the hazard function for the waiting time distribution. The IMI model generalizes both of these, creating an inhomogeneous version of a renewal model.[9] The simplest IMI model takes the conditional intensity to be of the multiplicative form[10]

$$\lambda(t|H_t) = g_0(t)g_1(t - s_*(t)). \tag{19.31}$$

A point process having conditional intensity of the form (19.30) or (19.31) may be fitted using binary Poisson regression, as in Example 1.1 on p. 576, except now with the additional terms representing the function $g_1(u)$ (where $u = t - s_*(t)$). A simple method is to fit the functions $g_0(t)$ and $g_1(u)$ using Poisson regression splines, in much the same way as discussed previously on p. 422 and 576 for Example 1.1.

Example 1.1 (continued from p. 576) Kass and Ventura (2001) fitted a model of the form (19.31) to data from an SEF neuron recorded for the study of Olson et al (2000). To do this they wrote

$$\log \lambda(t|H_t) = \log g_0(t) + \log g_1(t - s_*(t))$$

---

[8] The terminology is intended to signify that the history dependence is limited to the previous spike time. A discrete-time stochastic process is a Markov process if the probability that the process will be in a particular state at time t depends only on the state of the process at time $t - 1$.

[9] Because integrate-and-fire neurons reset to a baseline subthreshold voltage after firing, they necessarily follow Eq. (19.30). Further discussion of IMI models and their relationship to integrate-and-fire neurons is given in Koyama and Kass (2008).

[10] The functions $g_0(t)$ and $g_1(u)$ are defined only up to a multiplicative constant. That is, for any nonzero number c if we multiply $g_0(t)$ by c and divide $g_1(u)$ by c we do not change the result. Some arbitrary choice of scaling must therefore be introduced. In Fig. 19.7 the constant was chosen so that $g_0(t)$ was equal to the Poisson process intensity at time $t = 50$ ms after the appearance of the visual cue.

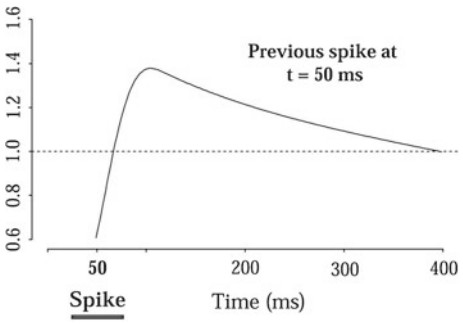

Fig. 19.7 Plot of the function $g_1(t - s_*(t))$ defined in (19.31) for the SEF data. The function is scaled so that a value of 1 makes the conditional intensity equal to the Poisson process intensity at time $t = 50$ ms after the appearance of the visual cue. Adapted from Kass and Ventura (2001).

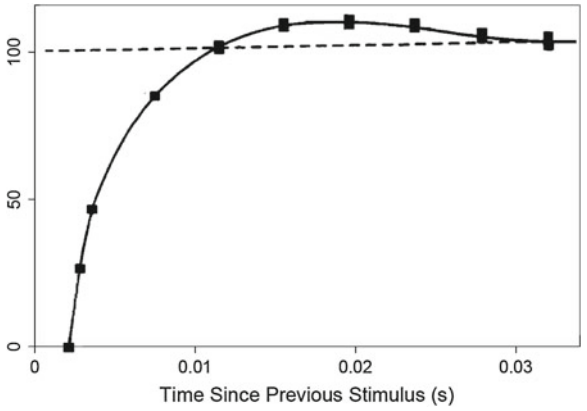

Fig. 19.8 Refractory effects in sciatic nerve of a frog. The y-axis is the reciprocal of the voltage threshold required to induce a second spike following a previous spike. The value 100 on the y-axis indicates the required reciprocal voltage when there was a long gap between the two successive action potentials. Adapted from Adrian and Lucas (1912).

which is an instance of (19.5) without coupling terms. Kass and Ventura took both $\log g_0(t)$ and $\log g_1(u)$ to be splines with a small number of knots and applied Poisson regression (see p. 422) using standard software. They showed that the model fitted the data better than an inhomogeneous Poisson model (using the graphical method in Section 19.3.5), and that inclusion of cross-product terms did not improve the fit (the likelihood ratio test for the additional terms was not significant).

A plot of the resulting non-Poisson recovery function $g_1(u)$ is shown in Fig. 19.7. For a Poisson process this function would be constant and equal to 1. The plot shows neural firing to be inhibited, compared with Poisson, for about 10 milliseconds and then it becomes *more* likely to fire, with the increase declining gradually until it returns to a baseline value. □

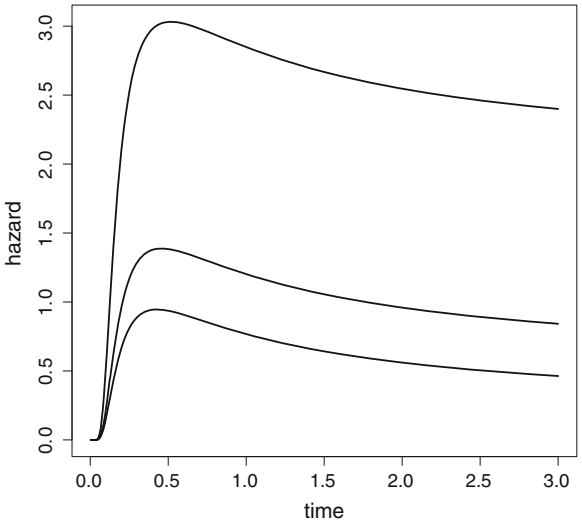

Fig. 19.9 Plots of inverse Gaussian hazard function for three different values of the coefficient of variation, .7 (*top curve*), 1 (*middle curve*), and 1.3 (*bottom curve*). These values correspond to the rough range of those commonly observed in cortical interspike interval data. The theoretical coefficient of variation is given by Eq. (5.16).

The non-monotonic behavior of the recovery function $g_1(t - s_*(t))$ in the foregoing analysis of Example 1.1 may seem somewhat surprising, but anecdotal evidence suggests it may be common. Interestingly, Adrian and Lucas (1912) found a qualitatively similar result by a very different method. They stimulated a frog's sciatic nerve through a second electrode and examined the time course of "excitability," which they defined as the reciprocal of the voltage threshold required to induce an action potential. Figure 19.8 plots this excitability as a function of time since the previous stimulus. There is again a relative refractory period of approximately 10 ms followed by an overshoot and a gradual return to the baseline value. Furthermore, the theoretical inter-spike interval distribution for an integrate-and-fire neuron (following a random walk generated by excitatory and inhibitory post-synaptic potentials) is inverse Gaussian (see Section 5.4.6), and the hazard function for an inverse Gaussian has a non-monotonic shape, shown in Fig. 19.9, that closely resembles the typical recovery function. The qualitative shape of the recovery function shown in Fig. 19.7 is thus consistent with what we would expect from the point of view of theoretical neurobiology.

In many experimental settings spike trains are collected to see how they differ under varying experimental conditions. The conditions may be summarized by a variable or vector, often called a *covariate* (as in regression, see p. 332). Furthermore, there may be other variables that may be related to spiking activity, which could be time-varying, such as a local field potential. Let us collect any such covariates into a vector denoted by u_t if we regard them as fixed by the experimenter, and V_t if

they should be considered stochastic. We then write $X_t = (H_t, u_t, V_t)$ and let the conditional intensity become a function not only of time and history, but also of the covariate vector X_t. Thus, for an observation $X_t = x_t$ we write the conditional intensity in the form $\lambda(t|x_t)$. With this in hand we may generalize the statement on p. 586, allowing it to cover the interesting cases implied by our discussion surrounding Eq. (19.5), as follows:

If we consider spike trains to be point processes, within trials the instantaneous firing rate is $\lambda(t|x_t)$ and we have

$$P(\text{spike in } (t, t+dt]|H_t) = \lambda(t|x_t)dt. \tag{19.32}$$

We may also generalize formula (19.20).

Theorem If the conditional intensity of an orderly point process on an interval $(0, T]$ depends on the random process X_t, so that when $X_t = x_t$ it may be written in the form $\lambda(t|x_t)$, then, conditionally on $X_t = x_t$, the event time sequence $S_1, S_2, \ldots, S_{N(T)}$ has joint pdf

$$f_{S_1,\ldots,S_{N(T)}|X_t}(s_1, \ldots, s_n|X_t = x_t) = \exp\left\{-\int_0^T \lambda(t|x_t)dt\right\} \prod_{i=1}^{n} \lambda(s_i|x_t). \tag{19.33}$$

Proof: The proof is the same as that given for (19.20) in Section 19.4 with x_t replacing H_t. □

A detail: If we are interested in the variation of the conditional intensity with the random vector X_t we can emphasize this by writing it in the form $\lambda(t|X_t)$. For example, in a multi-trial experiment, the firing rate may vary across trials, and the conditional intensity could include a component that changes across trials (see Ventura et al. 2005b). In such situations, the model includes two distinct sources of variability: one due to variability described by the point process pdf in (19.33) and the second due to the way the conditional intensity varies with X_t. The resulting point process is often called *doubly stochastic*. □

Example 16.6 (continued from p. 472) We now give some additional details about the model used by Frank et al (2002). They applied a multiplicative IMI model to characterize spatial receptive fields of neurons from both the CA1 region of the hippocampus and the deep layers of the entorhinal cortex (EC) in awake, behaving rats. In their model, each neuronal spike train was described in terms of a conditional intensity function of the form (19.31), where the temporal factor $g_0(t)$ became

$$g_0(t) = g^S(t, u_t)$$

where u_t is the animal's two-dimensional spatial location at time t. In other words, $g^S(t, u_t)$ is a time-dependent place field. As we said on p. 472 the authors adopted a state-space model (see Section 16.2.4), where the state variables involved features of the place field. By modeling the resulting conditional intensity in the form

$$\lambda(t|x_t) = g^S(t, u_t)g_1(t - s_*(t))$$

the authors found consistent patterns of plasticity in both CA1 hippocampal neurons and deep entorhinal cortex (EC) neurons, which were distinct: the spatial intensity functions of CA1 neurons showed a consistent increase over time, whereas those of deep EC neurons tended to decrease. They also found that the ISI-modulating factor $g_1(t - s_*(t))$ of CA1 neurons increased only in the "theta" region (75–150 ms), whereas those of deep EC neurons decreased in the region between 20 and 75 ms. In addition, the minority of deep EC neurons whose spatial intensity functions increased in area over time fired in a more spatially specific manner than non-increasing deep EC neurons. This led them to suggest that this subset of deep EC neurons may receive more direct input from CA1 and may be part of a neural circuit that transmits information about the animal's location to the neocortex. □

It is easy to supplement (19.31) with terms that consider not only the spike $s_*(t)$ immediately preceding time t, but also the spike $s_{2*}(t)$ preceding $s_*(t)$, $s_{3*}(t)$ preceding $s_{2*}(t)$, etc. One way to do this is to write

$$\lambda(t|H_t) = g_0(t)g_1(t - s_*(t))g_2(t - s_{2*}(t))g_3(t - s_{3*}(t)) \tag{19.34}$$

or, equivalently,

$$\begin{aligned}\log \lambda(t|H_t) = &\log g_0(t) + \log g_1(t - s_*(t)) \\ &+ \log g_2(t - s_{2*}(t)) + \log g_3(t - s_{3*}(t))\end{aligned}$$

and then use additional spline-based terms to represent $\log g_2(t - s_{2*}(t))$ and $\log g_3(t - s_{3*}(t))$ in a Poisson regression.

Example 1.1 (continued) In their study of the model (19.31) for SEF neurons, described on p. 589, Kass and Ventura also used a model that included several spikes preceding time t, as in (19.34). The implementation again used splines with a small number of knots to represent each of the additional functions $g_2(t - s_{2*})$, $g_3(t - s_{3*})$, etc. The authors found the extra terms did not improve the fit (the likelihood ratio test was not significant). □

> *A detail:* In applying (19.34) using regression splines, Kass and Ventura allowed the functions $g_1(t - s_*)$, $g_2(t - s_{2*})$, $g_3(t - s_{3*})$, to be distinct. A plausible alternative is to assume they have the same functional form, which would mean that they have the same knots and the

same coefficients. This would say that the way a spike at time s prior to time t alters the probability of neural firing at time t depends only on $t - s$ and not on how many spikes occur between time s and time t. In this case (19.34) is replaced by

$$\lambda(t|H_t) = g_0(t)g_1(t - s_*(t))g_1(t - s_{2*}(t))g_1(t - s_{3*}(t)).$$

This simplification reduces the number of parameters in the model. Models of this type were used by Pillow et al (2008). □

Another way model (19.31) may be extended is to include terms corresponding to coupling between neurons, as indicated by (19.5). To illustrate, we may consider the effect of neuron B on a given neuron A by letting $u_*(t)$ be the time of the neuron B spike that precedes time t and, similarly, letting $u_{2*}(t)$ and be the time of the spike preceding $u_*(t)$ and $u_{3*}(t)$ the time of the spike preceding $u_{2*}(t)$. Then we may append to (19.34) a series of factors that represent the coupling effects. In logarithmic form, considering 3 spikes back in time, this becomes

$$\begin{aligned}\log \lambda(t|H_t) = &\log g_0(t) + \log g_1(t - s_*(t)) \\ &+ \log g_2(t - s_{2*}(t)) + \log g_3(t - s_{3*}(t)) \\ &+ \log h_1(t - u_*(t)) + \log h_2(t - u_{2*}(t)) \\ &+ \log h_3(t - u_{3*}(t)).\end{aligned} \tag{19.35}$$

Once again (19.35) takes the form of (19.5), and some version of Poisson regression may be applied.

Example 19.3 (continued) In introducing this example on p. 569 we said that the authors used a model having the form of (19.5). Let us be somewhat more specific. In terms of (19.35), Pillow et al. took the receptive-field stimulus effects ($g_0(t)$, here spatio-temporal as in Example 16.6) to be linear, i.e., a linear combination of 5×5 stimulus pixel intensities across 30 time bins. For the history effects and the coupling effects they did not use splines but rather used an alternative set of basis functions such that $\log \lambda(t|H_t)$ remained linear, as it does with regression splines in (19.35). They then applied Poisson regression. However, because their model involved a large number of free parameters they had to use a modified fitting criterion (a form of penalized fitting similar to that used with smoothing splines) which is beyond the scope our presentation here. □

19.3.5 Graphical checks for departures from a point process model may be obtained by time rescaling.

As described in Section 3.3.1, Q–Q and P–P plots may be used to check the fit of a probability distribution to data. These plots indicate the discrepancy between the

empirical cdf $\hat{F}(x)$ and the theoretical cdf $F(x)$, the idea being that when $\hat{F}(x)$ is based on i.i.d. random variables we have $\hat{F}(x) \to F(x)$ for all x (if the distribution is continuous) as the sample size grows indefinitely large. In the case of point processes we may examine the inter-event waiting times $X_1, \ldots, X_n$. For a homogeneous Poisson process these are i.i.d. $Exp(\lambda)$. Thus, to assess the fit of a homogeneous Poisson process to a sequence of event times we may simply compute the inter-event waiting times and examine a Q–Q or P–P plot under the assumption that the true waiting-time distribution is exponential. For an inhomogeneous Poisson process, or a more general point process, the waiting times are no longer i.i.d. Thus, this method can not be applied in the same form. However, a a version of the probability integral transform (p. 122) may be used to create a homogeneous Poisson process from *any* point process. We begin with a conditional intensity function in the general form of Eq. (19.32).

Time Rescaling Theorem. Suppose we have a point process with conditional intensity function $\lambda(t|x_t)$ on $(0, T]$ and with occurrence times $0 < S_1 < S_2, \ldots, < S_{N(T)} \leq T$. Suppose further that the waiting time distributions are continuous with $f_{X_j|S_{j-1}}(x) > 0$ on $(s_{j-1}, T]$, for all $j \geq 1$. If we define

$$Z_1 = \int_0^{S_1} \lambda(t|x_t)dt \tag{19.36}$$

and

$$Z_j = \int_{S_{j-1}}^{S_j} \lambda(t|x_t)dt \tag{19.37}$$

for $j = 2, \ldots, N(T)$, then $Z_1, \ldots, Z_{N(T)}$ are i.i.d. $Exp(1)$ random variables.[11]

Proof: See Section 19.4. □

This result is called the time rescaling theorem because we can think of the transformation as stretching and shrinking the time axis based on the value of the conditional intensity function. If $\lambda(t|x_t)$ were constant and equal to one everywhere, then the process would be a homogeneous Poisson process with independent, exponential ISIs, and time does not need to be rescaled. When $\lambda(t|x_t)$ is less than one, the transformed event times z_j accumulate slowly and represent a shrinking of time, so that distant event times are brought closer together. Likewise, when $\lambda(t|x_t)$ is greater than one, the event times z_j accumulate more rapidly and represent a stretching of time, so that neighboring event times are drawn further apart.

With time rescaling in hand, we may now apply Q–Q or P–P plots to detect departures from a point process model: using the conditional intensity function we transform the time axis and judge the extent to which the resulting waiting times deviate from those predicted by an $Exp(1)$ distribution. Furthermore, in conjunction with a P–P plot, the Kolmogorov-Smirnov test (Section 10.3.7) may be applied to test

[11] Extending the argument slightly to include the interval (s_N, T) it may also be shown that $Z_1, \ldots, Z_{N(T)}$ follow a homogeneous Poisson process with intensity $\lambda = 1$.

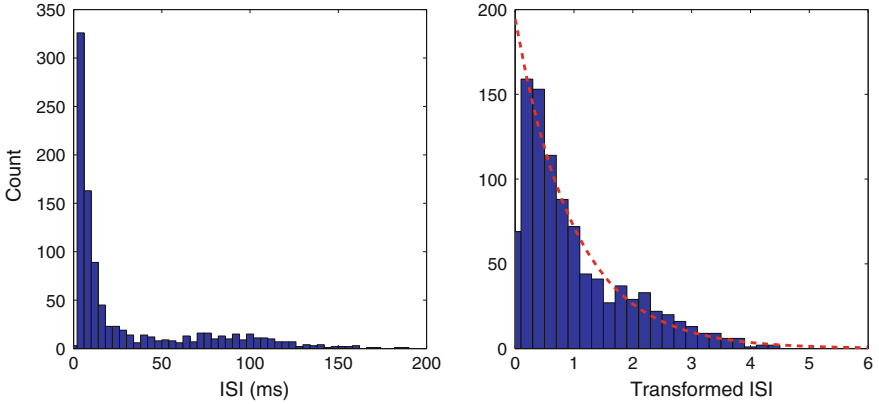

Fig. 19.10 *Left* Histogram of ISIs for the retinal ganglion cell spike train. *Right* Histogram of time-rescaled ISIs. *Dashed red line* is the $Exp(1)$ pdf.

the null hypothesis that the transformed waiting times follow an $Exp(1)$ distribution, which becomes an assessment of fit of the conditional intensity function. If the P–P plot consists of pairs (x_r, y_r), for $r = 1, \ldots, n$, the usual approach is to use the points $(x_r, y_r + 1.36/\sqrt{n})$ and $(x_r, y_r - 1.36/\sqrt{n})$ to define upper and lower bands for visual indication of fit, as illustrated in Fig. 19.11. Specifically, to make a P–P plot for a conditional intensity function $\lambda(t|x_t)$ used to model spike times $s_1, s_2, \ldots, s_n$ we do the following:

1. From (19.36) and (19.37) find transformed spike times $z_1, \ldots, z_n$;
2. for $j = 1, \ldots, n$ define $u_j = 1 - \exp(-z_j)$;
3. put the values $u_1, \ldots, u_n$ in ascending order to get $u_{(1)}, \ldots, u_{(n)}$;
4. for $r = 1, \ldots, n$ (see p. 67) plot the (x, y) pair $\left(\frac{r-.5}{n}, u_{(r)}\right)$;
5. produce upper and lower bands: for $r = 1, \ldots, n$ plot the (x, y) pair $\left(\frac{r-.5}{n}, u_{(r)} + 1.36/\sqrt{n}\right)$ and $\left(\frac{r-.5}{n}, u_{(r)} - 1.36/\sqrt{n}\right)$.

Example 19.1 (continued from p. 586) Using the conditional intensity of Eq. (19.28) we may apply time rescaling. Figure 19.10 displays a histogram of the original ISIs for this data. The smallest bin (0–2 ms) is empty due to the refractory period of the neuron. We can also observe two distinct peaks at around 10 and 100 ms respectively. It is clear that this pattern of ISIs is not described well by an exponential distribution, and therefore the original process cannot be accurately modeled as a simple Poisson process. However the histogram in the right panel of the figure shows the result of transforming the observed ISIs according to the conditional intensity model. Figure 19.11 displays a P–P plot for the intervals in the right panel of Fig. 19.10. Together, these figures show that the model in Eq. (19.28) does a good job of describing the variability in the retinal neuron spike train. □

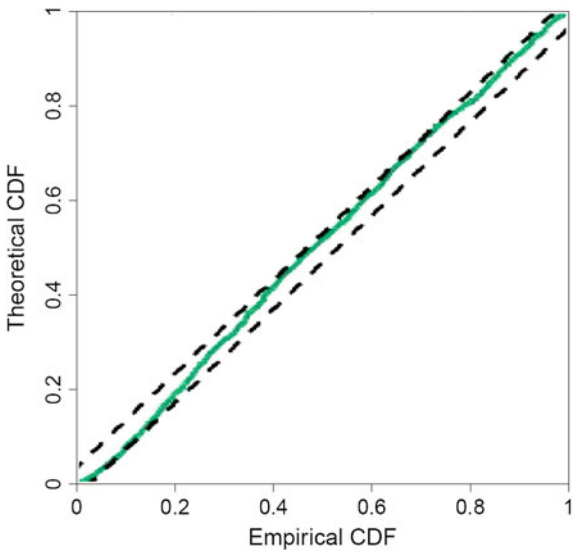

Fig. 19.11 P–P plot for the distribution of rescaled intervals shown in Fig. 19.10.

Example 19.5 Spike trains from a locust olfactory bulb. Substantial insight about sensory coding has been gained by studying olfaction among insects. An insect may come across thousands of alternative odors in its environment, among millions of potential possibilities, but only particular odors are important for the animal's behavior. A challenge has been to describe the mechanisms by which salient odors are learned. A series of experiments carried out by Dr. Mark Stopfer and colleagues (e.g., Stopfer et al. 2003) has examined the way neural responses to odors may evolve over repeated exposure. To capture subtle changes it is desirable to have good point process models for olfactory spike trains. Figure 19.12 displays P–P plots for the fit of an inhomogeneous Poisson model and a multiplicative IMI model to a set of spike trains from a locust olfactory bulb. The spike trains clearly deviate from the Poisson model; the fit of the multiplicative IMI model to the data is much better. □

19.3.6 There are efficient methods for generating point process pseudo-data.

It is easy to devise a computer algorithm to generate observations from a homogeneous Poisson processes, or some other renewal process: we simply generate a random sample from the appropriate waiting-time distribution; the *i*th event time will then be the sum of the first *i* waiting times. In particular, to generate a homogeneous

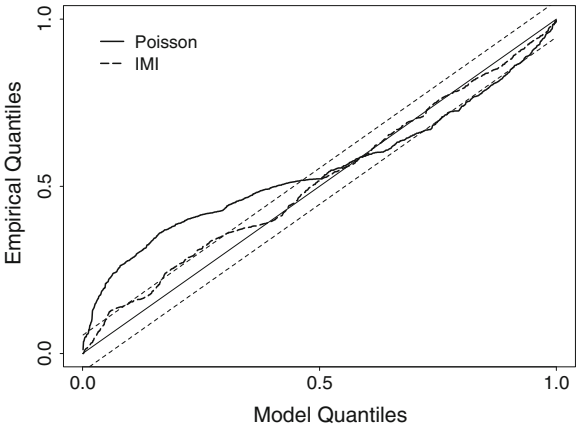

Fig. 19.12 P–P plots of inhomogeneous Poisson and multiplicative IMI models for spike train data from a locust olfactory bulb. For a perfect fit the curve would fall on the diagonal line $y = x$. The data-based (empirical) probabilities deviate substantially from the Poisson model but much less so from the IMI model. When the curve ranges outside the diagonal bands above and below the $y = x$ line, some lack of fit is indicated according to the Kolmogorov-Smirnov test (discussed in Section 10.3.7).

Poisson process with rate λ, we can draw a random sample from an $Exp(\lambda)$ distribution and take the ith event time to be $s_i = \sum_{j=1}^{i} x_j$.

Generating event times from a general point process is more complicated. One simple approach, based on the Bernoulli approximation, involves partitioning the total time interval into small bins of size Δt: in the kth interval, centered at t_k, we generate an event with probability $p_k = \lambda(t_k|x_{t_k})\Delta t$, where x_{t_k} depends on the history of previously generated events. This works well for small simulation intervals. However, as the total time interval becomes large and as Δt becomes small, the number of Bernoulli samples that needs to be generated becomes very large, and most of those samples will be zero, since $\lambda(t|x_t)\Delta t$ is small. In such cases the method becomes very inefficient and thus may take excessive computing time. Alternative approaches generate a relatively small number of i.i.d. observations, and then manipulate them so that the resulting distributions match those of the desired point process.

Thinning To apply this algorithm, the conditional intensity function $\lambda(t|x_t)$ must be bounded by some constant, $\lambda_{\max}$. The algorithm follows a two-stage process. In the first stage, a set of candidate event times is generated as a simple Poisson process with a rate $\lambda_{\max}$. Because $\lambda_{\max} \geq \lambda(t|x_t)$, these candidate event times occur more frequently than they would for the point process we want to simulate. In the second stage they are "thinned" by removing some of them according to a stochastic scheme. We omit the details. In practice, thinning is typically only used when simulating inhomogeneous Poisson processes with bounded intensity functions.

Time rescaling Another approach to simulating general point processes is based on the time-rescaling theorem. According to the statement of the theorem in Section 19.3.5, the transformed Z_i random variables follow an $Exp(1)$ distribution, with the transformation being based on the integral of the conditional intensity function. This suggests generating a sequence of $Exp(1)$ random variables and then back-transforming to get the desired point process. That idea turns out to work rather well in practice. Here is the algorithm for generating a process on the interval $(0, T]$ with conditional intensity $\lambda(t|x_t)$:

1. Initialize $s_0 = 0$ and $i = 1$.
2. Sample z_i from an $Exp(1)$ distribution.
3. Find s_i as the solution to
$$z_i = \int_{s_{i-1}}^{s_i} \lambda(t|x_t)dt.$$
4. If $s_i > T$ stop.
5. Set $i = i + 1$ and go to 2.

19.3.7 Spectral analysis of point processes requires care.

Because point processes may be considered, approximately, to be binary time series (see Section 19.1.2) it is tempting to treat them as a time series and use spectral methods to find frequency-based components, as in Section 18.3. This is possible, but requires attention to the nature of point processes.

In the first place, spectral analysis applies to stationary time series. To define stationarity (see on p. 515) for a point process we require that the counts $\Delta N_{(t_1,t_2]}$, $\Delta N_{(t_2,t_3]}, \ldots, \Delta N_{(t_{k-1},t_k]}$ have the same joint distribution as $\Delta N_{(t_1+h,t_2+h]}$, $\Delta N_{(t_2+h,t_3+h]}, \ldots, \Delta N_{(t_{k-1}+h,t_k+h]}$ for all h and all $t_1 < t_2 < \cdots < t_k$. However, we previously defined point processes only on the positive real line $(0, \infty)$ and for stationarity to make sense the process must be defined on the whole real line $(-\infty, \infty)$. One way to extend a point process to the negative half of the real line is to define the counts to be negative when $t < 0$. For example, suppose we have a homogeneous Poisson process on $(0, \infty)$ with rate λ. Let its counting process representation be $M_1(t)$. Now take another homogeneous Poisson process with rate λ and counting process $M_2(t)$ and define $N(t) = M_1(t)$ for $t > 0$ and $N(t) = -M_2(-t)$ for $t < 0$, and set $N(0) = 0$. Then $N(t)$ becomes the counting process representation of a stationary Poisson process with rate λ.

We now assume that we have counts $\Delta N_{(t_1,t_2]}$ defined for all t and that the resulting point process is stationary. In Section 18.3 the spectral density was defined as the Fourier transform of the autocovariance function. The expectation of a count was given in terms of the marginal intensity in (19.24). In the stationary case the marginal intensity must be time-invariant and therefore equal to a constant λ. We may define a covariance intensity function analogously as

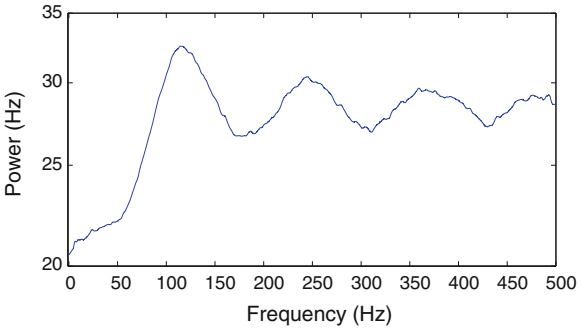

Fig. 19.13 Estimated spectral density from a simulated spike train. The simulated spike train had an average firing rate of roughly 28 Hz, a 5 ms refractory period, and an increased probability of spiking after a previous spike roughly 8 ms in the past. The estimated spectral density does not appear to reflect these properties and is easily misinterpreted.

$$\begin{aligned}\kappa(s,t) &= \lim_{\Delta t\to 0} \frac{E(\Delta N_{s,s+\Delta t]}\Delta N_{t,t+\Delta t]}) - E(\Delta N_{s,s+\Delta t]})E(\Delta N_{t,t+\Delta t]})}{(\Delta t)^2} \\ &= \lim_{\Delta t\to 0} \frac{E(\Delta N_{s,s+\Delta t]}\Delta N_{t,t+\Delta t]})}{(\Delta t)^2} - \lambda^2. \end{aligned} \tag{19.38}$$

This holds for $s \neq t$. In the stationary case $\kappa(s,t)$ is a function only of the difference $h = t - s$ so we write $\kappa(h)$ and use (19.38) for $h \neq 0$. For $s = t$ we have, for small Δt (because $\Delta N_{t,t+\Delta t]}$ is binary),

$$E(\Delta N_{t,t+\Delta t]}\Delta N_{t,t+\Delta t]}) = E(\Delta N_{t,t+\Delta t]})$$

which implies that the limit in (19.38) vanishes. Instead, we define

$$\kappa(0) = \lim_{\Delta t\to 0} \frac{V(\Delta N_{t,t+\Delta t]})}{\Delta t} = \lambda. \tag{19.39}$$

We therefore must analyze separately[12] the cases $\kappa(0)$ and $\kappa(h)$ when $h \neq 0$. Keeping this in mind, we may now state that the point process spectrum is the Fourier transform of the covariance function. We omit details (see Brillinger 1972).

These technicalities are an indication that point process spectra are likely to behave somewhat differently than continuous spectra. It is possible to apply the discrete Fourier transform to spike train data and then try to interpret the result. Figure 19.13 displays an example of the estimated spectrum of a simulated spike train. Visual inspection of the estimated spectrum shows a dip at low frequencies, a large peak around 120 Hz, and maintained power out to 500 Hz. A näive interpretation from

[12] These may be combined by writing the covariance function, often called the *complete covariance function* as $\kappa(0)\delta(h) + \kappa(h)$ where $\delta(h)$ is the *Dirac delta function*, which is infinite at 0 and 0 for all other values of h.

this spectrum might presume that this spiking process has no very low frequency firing, tends to fire around 120 Hz, but also has considerable high frequency activity, suggesting no refractoriness. However, this interpretation is incorrect. The point process generating this spike train actually has an average firing rate around 28 Hz and reflects realistic spiking features including a 5 ms refractory period and an increased probability of firing 8 ms after a previous spike. The error here does not come from the computation of the estimated spectrum, but rather from the näive interpretation.

We do not pursue further the estimation of point process spectra. Our discussion of Fig. 19.13 is intended to show that point process spectra must be interpreted carefully.

19.4 Additional Derivations

Derivation of Equation (19.9) We start with a lemma.

Lemma The pdf of the ith waiting-time distribution is

$$f_{S_i}(s_i|S_{i-1} = s_{i-1}) = \lambda(s_i) \exp\left\{-\int_{s_{i-1}}^{s_i} \lambda(t)dt\right\}. \tag{19.40}$$

Proof of the lemma: Note that $\{S_i > s_i|S_{i-1} = s_{i-1}\}$, is equivalent to there being no events in the interval $(s_{i-1}, s_i]$. Therefore, from the definition of a Poisson process on p. 574 together with the Poisson pdf in Eq. (5.3), we have $P(S_i > s_i|S_{i-1} = s_{i-1}) = P\left(\Delta N_{(s_{i-1},s_i]} = 0\right) = \exp\left\{-\int_{s_{i-1}}^{s_i} \lambda(t)dt\right\}$, and the ith waiting time CDF is therefore $P(S_i \le s_i|S_{i-1} = s_{i-1}) = 1 - \exp\left\{-\int_{s_{i-1}}^{s_i} \lambda(t)dt\right\}$. The derivative of the CDF

$$f_{S_i}(s_i|S_{i-1} = s_{i-1}) = \frac{d}{ds_i}\left(1 - \exp\left\{-\int_{s_{i-1}}^{s_i} \lambda(t)dt\right\}\right)$$

gives the desired pdf. □

Proof of the theorem: We have

$$\begin{aligned} &f_{S_1,\ldots,S_{N(T)}}(s_1, \ldots, s_n) \\ &= f_{S_1}(s_1) f_{S_2}(s_2|S_1 = s_2) \cdots f_{S_{N(T)}}(s_n|S_{n-1} = s_{n-1}) \cdot P(\Delta N_{(s_n,T]} = 0). \end{aligned}$$

The factors involving waiting-time densities are given by the lemma. The last factor is

$$P(\Delta N_{(s_n,T]} = 0) = \exp\left(-\int_{s_n}^{T} \lambda(t)dt\right).$$

Combining these gives the result. □

Derivation of Equation (19.20) We need a lemma, which is analogous to the lemma used in deriving (19.9).

Lemma For an orderly point process with conditional intensity $\lambda(t|H_t)$ on $[0, T]$, the pdf of the ith waiting-time distribution, conditionally on $S_1 = s_1, \ldots, S_{i-1} = s_{i-1}$, for $t \in (s_{i-1}, T]$ is

$$f_{S_i|S_1,\ldots,S_{i-1}}(s_i|S_1 = s_1, \ldots, S_{i-1} = s_{i-1}) = \lambda(s_i|H_t) \exp\left\{-\int_{s_{i-1}}^{s_i} \lambda(t|H_t)dt\right\}. \tag{19.41}$$

Proof of the lemma: Let X_i be the waiting time for the ith event, conditionally on $S_1 = s_1, \ldots, S_{i-1} = s_{i-1}$. For $t > s_{i-1}$ we have $X_i \in (t, t + \Delta t)$ if and only if $\Delta N_{(t,t+\Delta t)} > 0$. Furthermore, if the ith event has not yet occurred at time t we have $H_t = (s_1, \ldots, s_{i-1})$. We then have

$$\begin{aligned} &\lim_{\Delta t \to 0} \frac{P(X_i \in (t, t + \Delta t)|X_i > t, S_1 = s_1, \ldots, S_{i-1} = s_{i-1})}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(\Delta N_{(t,t+\Delta t)} > 0|H_t))}{\Delta t} \end{aligned}$$

and, because the point process is regular, the right-hand side is $\lambda(t|H_t)$. Just as we argued in the case of hazard functions, in Section 3.2.4, the numerator of the left-hand side may be written

$$P(X_i \in (t, t + \Delta t)|X_i > t, H_t) = \frac{F(t + \Delta t|H_t) - F(t|H_t)}{1 - F(t|H_t)}$$

where F is the CDF of the waiting time distribution, conditionally on H_t. Passing to the limit again gives

$$\lim_{\Delta t \to 0} \frac{P(X_i \in (t, t + \Delta t)|X_i > t, H_t)}{\Delta t} = \frac{f(t|H_t)}{1 - F(t|H_t)}.$$

In other words, just as in the case of a hazard function, the conditional intensity function satisfies

$$\lambda(t|H_t) = \frac{f(t|H_t)}{1 - F(t|H_t)}.$$

Proceeding as in the case of the hazard function we then get the conditional pdf

$$f(t|H_t) = \lambda(t|H_t)e^{-\int_{s_{i-1}}^{t} \lambda(u|x_u)du}$$

as required. □

Proof of the theorem: The argument follows from the lemma by the same steps as the theorem for inhomogeneous Poisson processes. □

Proof of the time rescaling theorem Note that the transformed waiting times are

$$Z_j = \int_{s_{j-1}}^{s_j} \lambda(u|x_u)du$$

where $s_0 = 0$. Applying the theorem on producing exponential random variables from the probability integral transform (p. 122) to $X_1 = S_1$ with $Z_1 = G(X_1)$ and $G(t) = G_1(t)$ where

$$G_1(t) = \int_0^t \lambda(u|x_u)du,$$

we get $Z_1 \sim Exp(1)$. Continuing to the next event time and defining $X_2 = S_2 - S_1$ with $Z_2 = G(X_2)$ and $G(t) = G_2(t)$ where

$$G_2(t) = \int_{s_1}^t \lambda(u|x_u)du,$$

we get $Z_2 \sim Exp(1)$ and, furthermore, this same distribution results regardless of the value of $Z_1 = z_1$. Thus, the conditional density $f_{Z_2|Z_1}(z_2|Z_1 = z_1)$ does not depend on z_1; therefore Z_2 is independent of Z_1. Continuing on, we get $Z_j \sim Exp(1)$ independently of all Z_i for $i < j$, for all $j = 1, \ldots, n$ and for all possible values $n = N(T)$ of the random variable $N(T)$. □

Appendix
Mathematical Background

A.1 Introduction

The data we discuss in this book consist of numbers we conceptualize, abstractly, as values of variables in the sense of elementary algebra: a variable x can take on many possible numerical values. We talk about relationships between measured variables, such as x and y, in terms of functions, writing expressions like $y = f(x)$. Strings of numbers form vectors, while arrays of numbers form matrices, and matrix algebra extends many concepts and manipulations involving one or two variables to those involving many variables. The purpose of this appendix is to review the essential properties of numbers, vectors, matrices, and functions that are used repeatedly in the analysis of neural data. Our goal is not to teach the concepts, but rather to offer convenient reminders.

A.2 Numbers and Vectors

Rational numbers have the form $\frac{m}{n}$ where m and n are integers. Real numbers include not only rational numbers but also algebraic numbers like $\sqrt{2}$ and transcendental numbers like π. Real numbers are those that correspond to points on the number line. They are used to represent measurements. When we say that a variable x (representing a measurement) may take on a range of values in an interval (a, b) we mean that x may be any real number such that $a < x < b$. However, every measurement is limited to a certain accuracy, and thus to a pre-specifiable finite number of possible values. Thus, data that are somehow recorded by a physical device and are represented in the output of software are rational numbers and it is, therefore, not literally true that a measurement can take on any real value in (a, b); for example, most of the values in (a, b) are irrational. Instead, the use of intervals of real numbers to represent measurements is an abstraction, but it is the starting point in applying modern mathematics to the real world. When we speak of a number we mean a real

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1,

number unless we specifically say otherwise. Complex numbers are discussed in Section A.10.

Throughout the book we identify multiple unspecified values of a particular variable by using subscripts. Thus, x_1, x_2, x_3 might represent three values of x. We then also use the summation notation,

$$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3$$

and, more generally,

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n.$$

Similarly, we use the product notation

$$\prod_{i=1}^{3} x_i = x_1 x_2 x_3 = x_1 \times x_2 \times x_3$$

and, more generally,

$$\prod_{i=1}^{n} x_i = x_1 x_2 \cdots x_n.$$

We also use subscripts in another way when we work with vectors. A 2-dimensional vector is an ordered pair (x, y) and a 3-dimensional vector is an ordered triple (x, y, z). More generally, n-tuples have the form $(x_1, x_2, \ldots, x_n)$. We say that $(x_1, x_2, \ldots, x_n)$ is an n-dimensional vector having ith component x_i, for $i = 1, \ldots, n$. The set of all such n-dimensional vectors is labelled R^n (which we read as "r n"), for reasons we discuss in Section. A.9. Vectors and vector manipulations are a convenient way to consider, together, all the components. When we consider matrix manipulations we need to distinguish column vectors

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

from row vectors $(x_1, \ldots, x_n)$, but for other purposes we may ignore this distinction. The sum of two n-dimensional vectors $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ is

$$x + y = (x_1 + y_1, \ldots, x_n + y_n)$$

and their dot product is

$$x \cdot y = \sum_{i=1}^{n} x_i y_i. \tag{A.1}$$

A 1-dimensional vector is a number, and in the context of vector and matrix manipulations is often referred to as a *scalar*. The product of multiplying a vector $x = (x_1, \ldots, x_n)$ by a scalar c is

$$cx = (cx_1, \ldots, cx_n).$$

A.3 Functions and Linear Approximation

A function is a mapping from one set to another such that each element of the first set is taken to a particular element of the second set. We will be interested mainly in functions of real numbers or vectors that map into real numbers. If x is a real number or vector, we often write $y = f(x)$ to indicate that the function f maps x to y.

Suppose f is a function on a real interval. For many, many calculations it is useful to approximate f linearly, i.e., to write $y = f(x) \approx a + bx$ for suitable coefficients a and b. This is accomplished using the *derivative* of f, which is given by

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h},$$

assuming this limit is well-defined. We may also write

$$\frac{df}{dx} = f'(x)$$

and if we wish to specify that the derivative is evaluated at $x = x_0$ we write

$$\left.\frac{df}{dx}\right|_{x=x_0} = f'(x_0).$$

The linear approximation of f at a value x_0 is given by $b = f'(x_0)$. If $y_0 = f(x_0)$ we may then plug (x_0, y_0) into $y = a + bx$ to get $a = y_0 - f'(x_0)x_0$ and then we have $y \approx a + bx$ as the linear approximation to f for values of x close to x_0. By rearranging terms we can also write this in the form

$$y \approx f(x_0) + f'(x_0)(x - x_0) \tag{A.2}$$

or

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0). \tag{A.3}$$

When this kind of linear approximation is put in a form that explicitly recognizes the approximation error it is called a *first-order Taylor series*. Thus, a first-order Taylor series of the function $f(x)$ is the linear approximation having the form

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + R$$

where the remainder R satisfies $R \to 0$ as $x \to x_0$. Taylor series may be carried out to higher terms, involving higher derivatives.

Functions of several variables also have linear approximations based on derivatives, but the derivatives must be taken with respect to each of the function arguments and are then called *partial derivatives*. If $y = f(x_1, x_2)$ we write the partial derivatives as

$$\frac{\partial f}{\partial x_1} = \lim_{h \to 0} \frac{f(x_1 + h, x_2) - f(x_1, x_2)}{h}$$
$$\frac{\partial f}{\partial x_2} = \lim_{h \to 0} \frac{f(x_1, x_2 + h) - f(x_1, x_2)}{h},$$

if these limits exist, and then the linear approximation of $y = f(x_1, x_2)$ near $(x_1, x_2) = (a, b)$, which generalizes (A.2), is

$$y \approx f(a, b) + \left.\frac{\partial f}{\partial x_1}\right|_{(x_1,x_2)=(a,b)} (x_1 - a) + \left.\frac{\partial f}{\partial x_2}\right|_{(x_1,x_2)=(a,b)} (x_2 - b).$$

Linear approximations of functions $y = f(x_1, x_2, \ldots, x_n)$ are analogous.

A.4 The Exponential Function and Logarithms

For a number A and positive integer k, A^k is the k-fold product of A with itself. Exponentiation begins with this process, and extends to cases A^z where z is any complex number. For now let us assume x is real, and write $f(x) = A^x$, but let us leave the value of A arbitrary. The inverse function is the logarithm: $\log_A(y) = f^{-1}(y)$, in other words, $\log_A(f(x)) = x$.

The defining property of exponentiation is that it converts addition into multiplication, i.e.,

$$f(a + b) = f(a)f(b). \tag{A.4}$$

Logarithms convert multiplication into addition:

$$f^{-1}(ab) = f^{-1}(a) + f^{-1}(b).$$

Although mathematics books usually define exponentiation via convergent Taylor series (which is quick), Eq. (A.4) may, literally, be used to define exponentiation: if

a function satisfies (A.4) it must have the form $f(x) = A^x$ for some A. The derivative of $f(x)$ has the form $f'(x) = cf(x)$ for some proportionality constant c. If we choose the proportionality constant to be 1, i.e., $f'(x) = f(x)$, we obtain the "natural" base for exponentiation, which is the number $A = e$. We sometimes write $e^x = \exp(x)$. We will always mean the natural logarithm (base e) when we write $\log(x)$, unless we say otherwise. It may be shown that the only solutions to the differential equation

$$f'(x) = cf(x)$$

for a constant c are functions of the form $f(x) = ae^{bx}$.

Using (A.3) with $x_0 = 0$ we get

$$\exp(x) \approx 1 + x$$

when x is near zero. More formally, we say that as $t \to 0$ we have $\exp(x)/(1+x) \to 1$. Similarly, the derivative of $\log(x)$ is $1/x$, and with $f(t) = \log(1+t)$ we have $f(0) = 0$ and $f'(0) = 1$. Equation (A.3) then gives

$$\log(1 + t) \approx t \tag{A.5}$$

for small t. Formally, we say that as $t \to 0$ we have $(1/t)\log(1 + t) \to 1$.

Now consider $\log(1 + x/n)$. For n large use (A.5) to get

$$\log(1 + \frac{x}{n}) \approx \frac{x}{n}$$

so that

$$n \log(1 + \frac{x}{n}) \approx x.$$

From the logarithm property $\log(a^b) = b\log(a)$ we have

$$\log\left((1 + \frac{x}{n})^n\right) \approx x$$

and exponentiating both sides we obtain, for large n,

$$e^x \approx (1 + \frac{x}{n})^n,$$

or, more formally, we say that as $n \to \infty$ we have

$$(1 + \frac{x}{n})^n \to e^x. \tag{A.6}$$

A.5 Trigonometry, Inner Products, and Orthogonal Projections

In any right triangle, if θ is one of the acute angles, its cosine, written as $\cos\theta$, is the ratio of the length of the adjacent side to the length of the hypotenuse and its sine, written as $\sin\theta$, is the ratio of the length of the opposite side to the length of the hypotenuse. More generally, if we let the two-dimensional vector (x, y) lie on the unit circle defined by $x^2+y^2 = 1$, and if the angle of this vector with the horizontal vector (1,0) is θ, then the cosine and sine functions are given by $x = \cos\theta$ and $y = \sin\theta$. From this definition of sine and cosine the vector $(\cos\theta, \sin\theta)$ is the rotation of the vector (1,0) counter-clockwise through an angle θ. Because $(\cos\theta, \sin\theta)$ is on the unit circle we also obtain $(\cos\theta)^2 + (\sin\theta)^2 = 1$ for all θ, which is usually written $\cos^2\theta + \sin^2\theta = 1$ for all θ. The tangent function is $\tan\theta = \sin\theta/\cos\theta$. Angles are measured either in radians or degrees. We will almost always use radians: 2π radians $= 360$ degrees.

Because (0, 1) results from rotating (1, 0) by an angle $\frac{\pi}{2}$, the y-component of a point on the unit circle at angle θ is the same as the x-component of a point at angle $\theta - \frac{\pi}{2}$, so the sine and cosine functions are simply phase translations of each other:

$$\sin\theta = \cos(\theta - \frac{\pi}{2}). \tag{A.7}$$

The cosine and sine functions are periodic, with period 2π, that is, $\cos(\theta + 2k\pi) = \cos\theta$ for any integer k. The sine is an odd function, $\sin(-\theta) = -\sin\theta$, and the cosine is an even function, $\cos(-\theta) = \cos\theta$. The inverse functions of sine, cosine, and tangent are the arcsine, arccosine, and arctangent, and they are written $\arccos(x)$, $\arcsin(x)$, and $\arctan(x)$.

Consider a triangle with angles A, B, C having opposite sides of length a, b, c. The value of A (in radians) may be determined from B and C using $A = \pi - B - C$. The value of a may be determined from b, c, and A as follows (see Fig. A.1). Let h be the height of the perpendicular dropped from the vertex having angle C onto the side of length c. We have $h = a\sin A$. This perpendicular, together with the side of length a, form a right triangle. Call the length of its third side d. We have $d = c - b\cos A$. Because it is a right triangle, $a^2 = h^2 + d^2$. Plugging in the expressions for h and d we get the *law of cosines*,

$$a^2 = b^2 + c^2 - 2bc\cos A. \tag{A.8}$$

Next, consider two unit vectors v_1 and v_2 at angles θ_1 and θ_2 with the x-axis. They have coordinates $v_1 = (\cos\theta_1, \sin\theta_1)$ and $v_2 = (\cos\theta_2, \sin\theta_2)$. Let $v = v_1 - v_2$. The length $||v||$ may be found by the ordinary (Euclidean) distance formula

$$||v||^2 = (\cos\theta_1 - \cos\theta_2)^2 + (\sin\theta_1 - \sin\theta_2)^2$$

and by the law of cosines (see the bottom panel of Fig. A.1)

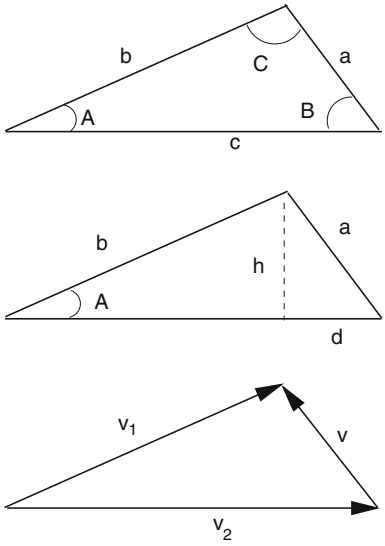

Fig. A.1 *Top two panels* Illustration of law of cosines. The *top panel* displays a triangle with sides of lengths *a, b, c*, and opposite angles *A, B, C*. The *second panel* displays the same triangle, but with the addition of the perpendicular of length *h* dropped from the top vertex onto its opposite side. *Bottom panel* The vector version of the law of cosines. The vectors v_1, v_2 and $v = v_1 - v_2$ form a triangle. If we take $a = ||v||$, $b = ||v_1||$ and $c = ||v_2||$ the law of cosines may be applied to produce the formula for $||v||^2$ given in the text.

$$||v||^2 = 2 - 2\cos(\theta_1 - \theta_2). \tag{A.9}$$

Equating these gives the important cosine addition (or subtraction) formula

$$\cos(\theta_1 - \theta_2) = \cos\theta_1 \cos\theta_2 + \sin\theta_1 \sin\theta_2. \tag{A.10}$$

The corresponding formula for sine addition, obtained from (A.10) by rewriting cosines as sines according to (A.7), is

$$\sin(\theta_1 - \theta_2) = \sin\theta_1 \cos\theta_2 - \sin\theta_2 \cos\theta_1. \tag{A.11}$$

A general sinusoidal function of period T is given by

$$f(t) = R\cos(2\pi\omega t - \phi),$$

where $\omega = 1/T$ is the frequency in cycles per unit t and ϕ is the phase. Using the addition formula (A.10), this function may instead be written

$$f(t) = A\cos(2\pi\omega t) + B\sin(2\pi\omega t)$$

where $A = \cos\phi$, $B = \sin\phi$, $R = \sqrt{A^2 + B^2}$, and $\phi = \arctan(-B/A)$. This representation is very important in regression analysis of periodic data.

The derivatives of the cosine and sine functions are

$$\frac{d}{d\theta}\sin(\theta) = \cos\theta$$

and

$$\frac{d}{d\theta}\cos(\theta) = -\sin\theta.$$

For θ near zero we have

$$\sin\theta \approx \theta \tag{A.12}$$

and

$$\cos\theta \approx 1. \tag{A.13}$$

Now consider a pair of two-dimensional vectors $v_1 = (x_1, y_1)$ and $v_2 = (x_2, y_2)$ (which need not be unit vectors), let θ be the angle between them and let $v = v_1 - v_2$. We may, as above, obtain the length $||v||$ from both the ordinary distance formula and the law of cosines. The distance formula gives

$$||v||^2 = (x_1 - y_1)^2 + (x_2 - y_2)^2 = x_1^2 - 2x_1y_1 + y_1^2 + x_2^2 - 2x_2y_2 + y_2^2$$

and the law of cosines gives

$$||v||^2 = ||v_1||^2 + ||v_2||^2 - 2||v_1||||v_2||\cos\theta = x_1^2 + y_1^2 + x_2^2 + y_2^2 - 2||v_1||||v_2||\cos\theta.$$

Equating these gives

$$x_1y_1 + x_2y_2 = ||v_1||||v_2||\cos\theta$$

the left-hand side of which is the dot product $v_1 \cdot v_2$, as in (A.1). This is also the Euclidean inner product:

$$\langle v_1, v_2\rangle = x_1y_1 + x_2y_2.$$

The Euclidean inner product formula extends immediately to n-dimensional vectors $v_1 = (x_1, \ldots, x_n)$ and $v_2 = (y_1, \ldots, y_n)$. The vectors v_1 and v_2 lie in a plane (which is the set of all vectors formed as linear combinations of v_1 and v_2), and when we speak of the angle between v_1 and v_2 we mean the angle between them within that plane. We have

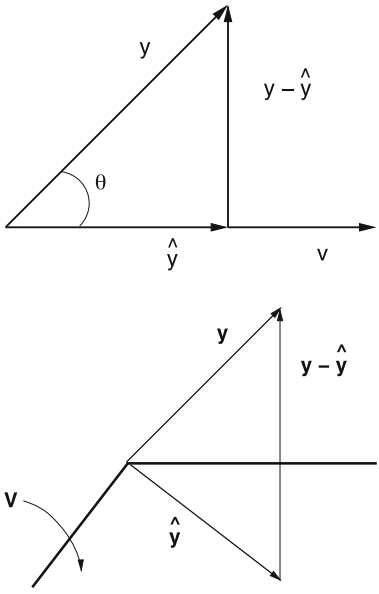

Fig. A.2 *Top panel* Orthogonal projection of the vector y onto the vector v, resulting in the vector $\hat{y}$ in the direction of v. *Bottom panel* Orthogonal projection of the vector y onto the vector subspace V resulting in the vector $\hat{y}$ in V.

$$\begin{aligned}\langle v_1, v_2\rangle &= \sum_{i=1}^{n} x_i y_i \\ &= x_1 y_1 + x_2 y_2 + \cdots + x_n y_n \\ &= ||v_1||||v_2|| \cos\theta\end{aligned} \tag{A.14}$$

where θ is the angle between v_1 and v_2. The squared length of a vector $v = (x_1, \ldots, x_n)$ is

$$||v||^2 = \langle v, v\rangle = v \cdot v = \sum_{i=1}^{n} x_i^2.$$

If $||v|| = 1$ the vector v is called a unit vector. For any vectors v and w and constants a and b, $\langle av, bw\rangle = ab\langle v, w\rangle$.

Two n-dimensional vectors v_1 and v_2 are said to be orthogonal if $\langle v_1, v_2\rangle = 0$. They are orthonormal if, in addition, they are unit vectors. The *orthogonal projection* of a vector y onto a vector v is the vector $\hat{y}$ that has the form $\hat{y} = cv$, for some nonzero constant c, and satisfies

$$\langle \hat{y}, y - \hat{y}\rangle = 0 \tag{A.15}$$

(see Fig. A.2). From (A.15) the vector $\hat{y}$ satisfies the Pythagorean relationship

$$||y||^2 = ||y - \hat{y}||^2 + ||\hat{y}||^2 \tag{A.16}$$

and $\hat{y}$ is the closest vector to y having the form cv in the sense that it minimizes the Euclidean distance

$$||y - \hat{y}|| = \min ||y - cv|| \tag{A.17}$$

where the minimum is taken over nonzero constants c.

We may solve for c by substituting cv for $\hat{y}$ in (A.15) to get

$$\langle cv, y\rangle = \langle cv, cv\rangle$$

so that

$$c = \frac{\langle v, y\rangle}{\langle v, v\rangle} \tag{A.18}$$

and, therefore,

$$\hat{y} = \frac{\langle v, y\rangle}{\langle v, v\rangle} v. \tag{A.19}$$

Let $u_1 = v/||v||$, which is the normalized version of v, meaning the unit vector in the same direction as v. Another expression for $\hat{y}$ is

$$\hat{y} = \langle u_1, y\rangle u_1 = (||y|| \cos\theta) u_1$$

where θ is the angle between y and v. The vector $y - \hat{y}$ is in the same plane as y and v. Let $r = y - \hat{y}$ and define $u_2 = r/||r||$. Then u_1 and u_2 are an orthonormal pair of vectors that lie in the same plane as y and v. We return to orthogonal projections in Section A.9.

A.6 Matrices

An $m \times k$ rectangular array of numbers, with m rows and k columns, is called an $m \times k$ *matrix*. The numbers m and k are are the dimensions of the matrix. We refer to the elements of a matrix by using subscripts of the form ij where i is the row and j is the column. For example, the 2×3 matrix A having rows (A_{11}, A_{12}, A_{13}) and (A_{21}, A_{22}, A_{23}) is

$$A = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \end{pmatrix}.$$

The value A_{ij} is the (i, j) element of A. To distinguish matrices from numbers, in several places we will instead use lower case a_{ij} (a number) to denote the (i, j) element of A (a matrix). An $n \times 1$ dimensional matrix is an n-dimensional vector. If A is an $m \times k$ matrix then its ith row, written $\text{row}_i(A)$, is a $1 \times k$ vector and its jth

column, written $\text{col}_j(B)$ is an $m \times 1$ vector. We say that a vector or matrix is non-zero if at least one of its elements is non-zero. The n-dimensional zero vector is the vector consisting of n zeroes and the $m \times k$ zero matrix is the $m \times k$ matrix all of whose elements are zero.

If A is an $m \times k$ matrix having elements a_{ij} for $i = 1, 2, \ldots, m, j = 1, 2, \ldots, k$ its *transpose*, denoted by A^T, is the $k \times m$ matrix with elements a_{ji} for $j = 1, 2, \ldots, k, i = 1, 2, \ldots, m$. That is, A^T is obtained from A by interchanging the rows and columns ($\text{row}_i(A^T) = \text{col}_i(A)$). If A is a $k \times k$ (square) matrix for which $A = A^T$ it is said to be *symmetric*.

Matrices are added element-wise. If A and B are both $m \times k$ matrices, having elements a_{ij} and b_{ij}, for $i = 1, 2, \ldots, m, j = 1, 2, \ldots, k$, then the sum of A and B is an $m \times k$ matrix C, written $C = A + B$, having elements c_{ij} given by

$$c_{ij} = a_{ij} + b_{ij} \quad i = 1, 2, \ldots, m \quad j = 1, 2, \ldots, k.$$

Note that the addition of matrices is defined only for matrices of the same dimensions. If c is a number A is an $m \times k$ matrix with elements a_{ij} then $cA = Ac$ is an $m \times k$ matrix B with elements b_{ij} that satisfy $b_{ij} = ca_{ij}$ for $i = 1, 2, \ldots, m, j = 1, 2, \ldots, k$. If A is an $m \times n$ matrix having elements a_{ij} and B is an $n \times k$ matrix having elements b_{ij} then their product $C = AB$ is the $m \times k$ matrix C whose element c_{ij} is given by

$$\begin{aligned} c_{ij} &= \text{row}_i(A) \cdot \text{col}_j(B) \\ &= \sum_{\ell=1}^{n} a_{i\ell} b_{\ell j} \end{aligned}$$

for all $i = 1, 2, \ldots, m, j = 1, 2, \ldots, k$. For the product AB to be defined, the column dimension of A must equal the row dimension of B. Then the row dimension of AB equals the row dimension of A and the column dimension of AB equals the column dimension of B. If A is an $m \times n$ matrix and B is an $n \times k$ matrix with $m \neq k$ then BA is not defined.

A square matrix A is said to be *diagonal* if its only non-zero entries are on its main diagonal, i.e., $A_{ij} = 0$ when $i \neq j$. The k-dimensional *identity* matrix, denoted by I_k, is the $k \times k$ diagonal matrix having all of its main diagonal elements equal to 1.

A.7 Linear Independence

A pair of vectors v_1 and v_2 is linearly dependent if they are multiples of each other, meaning that $v_2 = kv_1$ for some nonzero number k or, equivalently, if $c_1v_1 + c_2v_2 = 0$ where 0 represents the zero vector (the vector all of whose components are zero) and where $c_1 = k$ and $c_2 = -1$. Otherwise, if v_1 and v_2 are not multiples of each other, and neither is the non-zero vector, then there are no nonzero numbers c_1 and c_2 for which $c_1v_1 + c_2v_2 = 0$ and we say that v_1 and v_2 are linearly independent. More

generally, we say that a set of several vectors $v_1, v_2, \ldots, v_k$ are *linearly independent* if for every set of numbers $c_1, c_2, \ldots, c_k$ that are not all zero,

$$c_1 v_1 + c_2 v_2 + \cdots + c_k v_k \neq 0.$$

Equivalently, $v_1, v_2, \ldots, v_k$ are linearly independent if $c_1 v_1 + c_2 v_2 + \cdots + c_k v_k = 0$ implies that $c_1 = c_2 = \cdots = c_k = 0$. When $v_1, v_2, \ldots, v_k$ are not linearly independent then $c_1 v_1 + c_2 v_2 + \cdots + c_k v_k = 0$ for some nonzero set of coefficients $c_1, c_2, \ldots, c_k$, and the vectors are instead linearly dependent. In this case it becomes possible to write at least two of the vectors as linear combinations of the others for suitably chosen coefficients. For example, assuming $c_1 \neq 0$ we can set $a_i = -c_i/c_1$ for $i = 2, \ldots, k$ and by dividing $c_1 v_1 + c_2 v_2 + \cdots + c_k v_k = 0$ through by c_1 and then subtracting v_1 from both sides we get $v_1 = a_2 v_2 + \cdots + a_k v_k$.

For an $m \times k$ matrix A we may consider the set of m vectors consisting of the rows of A, i.e., the vectors $v_i = \text{row}_i(A)$ for $i = 1, \ldots, m$. The *row rank* of A is the maximum number of these row vectors that can be collected together and still remain linearly independent. Similarly, if we consider the k column vectors $\text{col}_1(A), \text{col}_2(A), \ldots, \text{col}_k(A)$, the *column rank* of A is the maximum number of these vectors that may be collected together and remain linearly independent. It may be shown that the row rank and the column rank of a matrix are equal. Thus, we speak of the *rank* of A, which is both the row rank and the column rank and is written $\text{rank}(A)$. Note that for an $m \times k$ matrix A we must have $\text{rank}(A) \leq \min(m, k)$. If $\text{rank}(A) = \min(m, k)$ then A is said to be of *full rank*. When a square matrix is of full rank it is called *nonsingular*.

Two key characterizations of nonsingular matrices are the following. First, a $k \times k$ matrix A is nonsingular if and only if for every non-zero vector x the vector Ax is also non-zero. Second, a $k \times k$ matrix A is nonsingular if and only if it has an inverse A^{-1} such that

$$AA^{-1} = A^{-1}A = I_k.$$

A third important characterization involves the *determinant* of A, denoted by $|A|$, and defined to be the scalar

$$\begin{array}{ll} |A| = a_{11} & \text{if } k = 1 \\ |A| = \sum_{j=1}^{k} a_{1j}|A_{1j}|(-1)^{1+j} & \text{if } k > 1 \end{array}$$

where A_{1j} is the $(k-1) \times (k-1)$ matrix obtained by deleting the first row and jth column of A. Also, $|A| = \sum_{j=1}^{k} a_{ij}|A_{ij}|(-1)^{i+j}$ using the ith row in place of the first row. We have that A is nonsingular if and only if $|A| \neq 0$. If A is nonsingular then $|A^{-1}| = 1/|A|$.

If A is a $k \times k$ matrix with elements a_{ij}, its *trace*, written $tr(A)$ is the sum of its diagonal elements: $tr(A) = \sum_{i=1}^{k} a_{ii}$.

A.8 Orthogonal Matrices and the Spectral Decomposition

A square matrix A is said to be *orthogonal* if its columns form an orthonormal set of vectors. This means that $\text{col}_i(A) \cdot \text{col}_j(A) = 1$ if $i = j$ and is 0 for $i \neq j$. Another way to say this is that $A^T A = I_k$ and, because $I_k^T = I_k$, we also have $AA^T = I_k$. These relations show that a square matrix A is orthogonal if and only if $A^T = A^{-1}$. As a special case, suppose A is a 2×2 orthogonal matrix. Then $\text{col}_1(A)$ is a unit vector, so it lies on the unit circle, and therefore may be written in the form $(\cos\theta, \sin\theta)^T$ for some θ; by orthogonality $\text{col}_2(A)$ then has the form vector $\pm(-\sin\theta, \cos\theta)^T$. If we take $\text{col}_2(A) = (-\sin\theta, \cos\theta)^T$ then for every two-dimensional vector x, Ax is the rotation of x counter-clockwise by the angle θ. We say that A is a *rotation matrix*. Note that $A^T x$ (which is also $A^{-1}x$) becomes a rotation of x clockwise by the angle θ. If instead $\text{col}_2(A) = -(-\sin\theta, \cos\theta)^T$ then Ax results from first re-orientating the y-axis that it points in the opposite direction (down, instead of up) and then rotating x counter-clockwise by the angle θ. Thus, every 2×2 orthogonal matrix is either a rotation matrix or a combination of rotation and re-orientation of the axes. In higher dimensions every orthogonal matrix is also necessarily a combination of rotation matrix and re-orientation of axes.

If A is a $k \times k$ square matrix, λ is a scalar, and x is a vector satisfying

$$Ax = \lambda x$$

then λ is said to be an *eigenvalue* of A and x is an *eigenvector* corresponding to λ. Suppose A is a symmetric matrix. If for all non-zero x we have $x^T Ax > 0$ then A is *positive definite*; if for all non-zero x we have $x^T Ax \geq 0$ then A is *positive semi-definite*. Note that variance matrices are positive semi-definite (see Section 4.3, p. 91). We now state one of the most powerful and important theorems in matrix algebra.

The Spectral Decomposition Theorem If A is a $k \times k$ symmetric matrix then it has a representation in the form

$$A = PDP^T \tag{A.20}$$

where D is a $k \times k$ diagonal matrix with D_{ii} being an eigenvalue of A, and P is orthogonal with $\text{col}_i(P)$ being an eigenvector corresponding to D_{ii}.

The spectral decomposition of a $k \times k$ symmetric matrix A gives a way of specifying a set of eigenvalues and eigenvectors for A. In general, if $Ax = \lambda x$ and $v = x/c$ for a non-zero scalar c then $Av = (c\lambda)v$, so that $c\lambda$ is also an eigenvalue. If, however, we require each eigenvector to be a unit vector, as in the spectral decomposition, then the corresponding eigenvalue is uniquely determined. When eigenvalues are computed by software they are usually put in descending order: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$. If A is also positive semi-definite then $\lambda_i \geq 0$ for all $i = 1, \ldots, k$ and the number of positive eigenvalues is equal to its rank. We note that a symmetric matrix is positive definite if and only if it is non-singular (which is also stated by saying it is positive definite

if and only if it is of full rank). Thus, a positive semi-definite matrix is non-singular if and only if all its eigenvalues are positive.

The spectral decomposition has a very nice geometrical interpretation. First, the set of two-dimensional points (u_1, u_2) satisfying

$$\frac{u_1^2}{E_{11}} + \frac{u_2^2}{E_{22}} = c^2 \tag{A.21}$$

where E_{11} and E_{22} are positive numbers, forms an ellipse centered at the origin. Furthermore, the ellipse is oriented so that its two axes fall along the u_1 and u_2 coordinate axes, and the lengths of its two axes are $2c\sqrt{E_{11}}$ and $2c\sqrt{E_{22}}$. If we let $u = (u_1, u_2)$ then Eq. (A.21) may be written

$$u^T D u = c^2 \tag{A.22}$$

where D is the diagonal matrix with diagonal elements E_{11}^{-1} and E_{22}^{-1}. Now let R_θ be the 2×2 orthogonal matrix that rotates each vector counter-clockwise through an angle θ. As pointed out above, R_θ^T is the 2×2 orthogonal matrix that rotates each vector clockwise through an angle θ. If we define $x = R_\theta u$ then $u = R_\theta^T x$, and from (A.22) we have

$$x^T R_\theta D R_\theta^T x = c^2 \tag{A.23}$$

so that (A.23) must be the equation of an ellipse whose axes fall along the axes defined by the vectors $\mathrm{col}_1(R_\theta)$ and $\mathrm{col}_2(R_\theta)$ and have lengths $2c\sqrt{E_{11}}$ and $2c\sqrt{E_{22}}$. Because every orthogonal matrix is a rotation followed by a possible re-orientation of the axes, and such a re-orientation of axes defining x would not change the location of the ellipse defined by (A.23), for any 2×2 orthogonal matrix P, the equation

$$x^T P D P^T x = c^2, \tag{A.24}$$

is the equation of an ellipse whose axes fall along the axes defined by the vectors $\mathrm{col}_1(P)$ and $\mathrm{col}_2(P)$ and have lengths $2c\sqrt{E_{11}}$ and $2c\sqrt{E_{22}}$. An analogous interpretation of Eq. (A.24) holds when x is k-dimensional and P and D are $k \times k$ matrices. Thus, for a positive definite matrix A, the equation $x^T A x = 1$ defines an ellipse, and the spectral decomposition of A shows that the axes of this ellipse are oriented along the eigenvectors of A and have lengths equal to twice the square-root of the reciprocal of the corresponding eigenvalues.

A.9 Vector Spaces

The n-dimensional vectors $e_1 = (1, 0, 0, \ldots, 0)$, $e_2 = (0, 1, 0, 0, \ldots, 0)$, ..., $e_n = (0, \ldots, 0, 1)$ play a special role because they specify the axes or coordinate directions corresponding to each component of an n-dimensional vector $x = (x_1, x_2, \ldots, x_n)$

(note also that each e_i has length 1). We may write

$$x = x_1 e_1 + x_2 e_2 + \cdots + x_n e_n. \tag{A.25}$$

We think of the set of n-tuples as forming a *vector space*, which we call n-dimensional real space and write as R^n. When we have any set of n linearly independent vectors $v_1, \ldots, v_n$ (such as $e_1, \ldots, e_n$) the vectors $v_1, \ldots, v_n$ are said to form a *basis* for R^n. The basis is the set of vectors, which we write as $\{v_1, \ldots, v_n\}$. Note that every vector x in R^n may be written as a linear combination of these basis vectors, i.e., there are numbers $c_1, \ldots, c_n$ for which $x = c_1 v_1 + \cdots + c_n v_n$. The basis vectors are said to *span* the vector space R^n and R^n is said to be *the span* of $\{v_1, \ldots, v_n\}$. If we have a smaller set of linear independent vectors, say $\{w_1, \ldots, w_k\}$, where $k < n$, then the set of all linear combinations of those vectors (including the zero vector) is also called their *span*; let us denote it by V. Then V is a k-dimensional vector space, which is a subspace of R^n. We may now generalize the notion of *orthogonal projection* given in Section A.5. If $y \in R^n$ the orthogonal projection of y onto V, written $\hat{y}$, is the vector $\hat{y}$ for which

$$\langle v, y - \hat{y} \rangle = 0 \tag{A.26}$$

for all $v \in V$. It may be shown that for any y there is only one vector $\hat{y}$ with this property. If the columns of an $n \times k$ matrix X span a k-dimensional vector space V in R^n then we may write

$$V = \{X\beta \text{ such that } \beta \in R^k\}. \tag{A.27}$$

Equation (A.27) provides an important way to think about linear regression: by (A.27) we may rewrite (A.26) in the form

$$\langle X\beta, y - \hat{y} \rangle = 0 \tag{A.28}$$

for all $\beta \in R^k$. This is the same as Eq. (12.58).

A.10 Complex Numbers

Imaginary numbers were introduced to solve equations that do not have real solutions, like $x^2 = -1$. One solution of this equation is the imaginary[1] number i (sometimes

[1] Imaginary numbers are like real numbers in being abstract constructions that do not represent perfectly any measurement process, and so they live in what might be called a theoretical world (of mathematics, physics, statistics, etc.) rather than our real world of sensations and physical tools. The name "imaginary" (apparently given by Descartes in 1637), is perhaps somewhat misleading in that it seems to imply real numbers are more "real" than imaginary numbers, which they are not. The great mathematician Gauss lamented this name for example, suggesting it might have been better to call square-roots of negative numbers "lateral." (See Dantzig (1954).)

instead denoted by j). The other solution is $-i$. If we multiply i by any real number y we get an imaginary number iy. A complex number is one that may have both real and imaginary components. The usual notation writes a generic complex number as $z = x + iy$, with $x = Re(z)$ being the real part of z and $y = Im(z)$ being the imaginary part of z. A real number $x = x + i0$ is also considered a complex number; similarly, an imaginary number $iy = 0 + iy$ is also considered complex. The number $\bar{z} = x - iy$ is called the complex conjugate of z. The magnitude of z is

$$|z| = \sqrt{x^2 + y^2} = \sqrt{z\bar{z}}.$$

Once we allow complex numbers, every polynomial equation can be solved.

Some amazing properties of complex numbers are derived fairly easily[2] by representing them in the form of two-dimensional vectors (x, y), where again x and y are the real and imaginary components, and then also using the polar coordinate form (R, θ), where $x = R\cos\theta$ and $y = R\sin\theta$. Here, $R = \sqrt{x^2 + y^2}$ is the length of the vector (x, y) and θ is the angle between (x, y) and the x-axis. In this representation the real number 1 becomes (1, 0), -1 becomes (-1, 0) and i becomes (0, 1). Consider the product $z = z_1 z_2$ of two complex numbers $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$. Applying the addition formulas for cosine and sine we have

$$\begin{aligned} z &= x_1x_2 - y_1y_2 + i(x_1y_2 + x_2y_1) \\ &= R_1R_2(\cos(\theta_1 + \theta_2) + i\sin(\theta_1 + \theta_2)). \end{aligned}$$

Let us specialize to the case in which $|z_1| = |z_2| = 1$ so that z_1 and z_2 become vectors on the unit circle, and we have

$$z = z_1 z_2 = \cos(\theta_1 + \theta_2) + i\sin(\theta_1 + \theta_2). \tag{A.29}$$

This is illustrated in Fig. A.3. Equation (A.29) says that multiplication of complex unit vectors corresponds to addition of the corresponding angles. We thus have an instance of addition (of angles) being transformed to multiplication (of complex unit vectors). But conversion of addition to multiplication is carried out by the exponential function. Apparently, there is some kind of exponentiation going on here. This exponential transformation is revealed in Euler's Formula, given by Eq. (A.31).

In Eq. (A.29), let us set $\theta_1 = \theta_2 = \theta/2$, where $z = \cos\theta + i\sin\theta$. We then have

$$z = \left(\cos(\frac{\theta}{2}) + i\sin(\frac{\theta}{2})\right)^2.$$

Repeating this multiplication for n vectors each having angle θ/n we obtain

[2] A rigorous argument would require additional details about convergence. In particular, Euler's formula (A.31) follows immediately from a comparison of the infinite Taylor series expansions of the complex exponential, cosine, and sine functions—but that requires proof of convergence of these series.

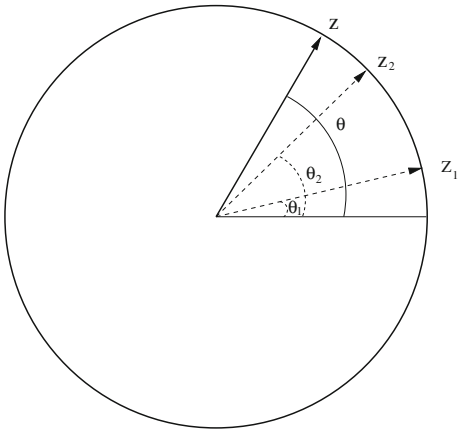

Fig. A.3 Multiplication of complex unit vectors. The complex numbers z_1 and z_2 are pictured as vectors with coordinates $x_i = \cos\theta_i$ and $y_i = \sin\theta_i$ for $i = 1, 2$, θ_i being the angle between z_i and the x-axis. Their product $z = z_1 z_2$ is a new complex number which, when pictured as a unit vector, has coordinates $x = \cos\theta$ and $y = \sin\theta$ where $\theta = \theta_1 + \theta_2$.

$$z = \left(\cos(\frac{\theta}{n}) + i\sin(\frac{\theta}{n})\right)^n,$$

or,

$$\cos\theta + i\sin\theta = \left(\cos(\frac{\theta}{n}) + i\sin(\frac{\theta}{n})\right)^n \tag{A.30}$$

for every positive integer n. Now consider what happens as we make n indefinitely large. Applying Eqs. (A.12) and (A.13) we get

$$\cos(\frac{\theta}{n}) + i\sin(\frac{\theta}{n}) \approx 1 + \frac{i\theta}{n}$$

and then, inserting this in the right-hand side of (A.30), letting $n \to \infty$, and applying (A.6) we get

$$\left(\cos(\frac{\theta}{n}) + i\sin(\frac{\theta}{n})\right)^n \to e^{i\theta}.$$

In other words, (A.30) together with (A.6) gives

$$\cos\theta + i\sin\theta \to e^{i\theta}$$

which, because the left-hand side does not involve n, can only be true if these quantities are equal; we thereby obtain Euler's formula:

$$e^{i\theta} = \cos\theta + i\sin\theta. \tag{A.31}$$

This formula is the foundation for Fourier analysis. On the one hand, it provides a kind of "book-keeping" of cosine and sine terms within a complex exponential while, on the other hand, it simplifies many manipulations because multiplication becomes addition of exponents. We also have

$$\cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{2} \tag{A.32}$$

and

$$\sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{2} \tag{A.33}$$

which are used to convert results involving complex exponentials to results involving sines and cosines. Using Euler's formula (A.31) we may represent any complex number z, in an exponential polar co-ordinate form,

$$z = Re^{i\theta}$$

where $R = |z| = \sqrt{x^2 + y^2}$ and $\theta = \arctan(y/x)$, with $x = Re(z)$ and $y = Im(z)$.

Just as the cosine and sine functions are periodic with period 2π, the complex exponential function is periodic with period $2\pi i$, i.e., $e^z = e^{z+i2k\pi}$ for every integer k. Special values of e^z include $1 = e^0$ (and thus $1 = e^{i2k\pi}$ for every integer k), $i = e^{i\pi/2}$, and $-1 = e^{i\pi}$. The latter may be written

$$e^{i\pi} - 1 = 0,$$

which appeals to many people's sense of mathematical aesthetics because it combines the five most fundamental numbers in a single equation. It is often called Euler's equation.

References

Abeles, M. (2009). "Synfire chains." *Scholarpedia*, 4, 1441.

Adolph, K. (2002). "Babies steps make giant strides toward a science of development." *Infant Behavior and Development*, 25, 86–90.

Adrian, E. and Lucas, K. (1912). "On the summation of propagated disturbances in nerve and muscle." *The Journal of Physiology*, 44, 68–124.

Adrian, E. and Zotterman, Y. (1926). "The impulses produced by sensory nerve endings: Part II: The response of a single end organ." *Journal of Physiology*, 61, 151–171.

Agresti, A. (1990). *Categorical data analysis*. Wiley.

Agresti, A. and Caffo, B. (2000). "Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures." *The American Statistician*, 54, 280–288.

Akaike, H. (1974). "A new look at the statistical model identification." *Automatic Control, IEEE Transactions on*, 19, 716–723.

Amirikian, B. and Georgopoulos, A. (2000). "Directional tuning profiles of motor cortical cells." *Neuroscience Research*, 36, 73–79.

Anderson, C. and Stevens, C. (1973). "Voltage clamp analysis of acetylcholine produced end-plate current fluctuations at frog neuromuscular junction." *Journal of Physiology*, 235, 655–691.

Anderson, J. (1990). *Cognitive Psychology and its Implications*. MacMillan.

Anderson, J. and Schooler, L. (1991). "Reflections of the environment in memory." *Psychological Science*, 2, 396–408.

Anscombe, F. (1973). "Graphs in statistical analysis." *The American Statistician*, 27, 1, 17–21.

Arlot, S. and Celisse, A. (2010). "A survey of cross-validation procedures for model selection." *Statistics Surveys*, 4, 40–79.

Bar-Gad, I., Ritov, Y., Vaadia, E., and Bergman, H. (2001). "Failure in identification of overlapping spikes from multiple neuron activity causes artificial correlations." *Journal of Neuroscience Methods*, 107, 1–13.

Bates, D. and Watts, D. (1988). *Nonlinear Regression Analysis and Its Applications*. New York: Wiley.

Behrmann, M., Ghiselli-Crippa, T., Sweeney, J., DiMatteo, I., and Kass, R. (2002). "Mechanisms underlying spatial representation revealed through studies of hemispatial neglect." *Journal of Cognitive Neuroscience*, 14, 272–290.

Behseta, S., Berdyyeva, T., Olson, C.R., and Kass, R.E. (2009). "Bayesian correction for attenuation of correlation in multi-trial spike count data." *Journal of Neurophysiology*, 101, 2186–2193.

Behseta, S. and Kass, R. E. (2005). "Testing equality of two functions using BARS." *Statistics in Medicine*, 24, 3523–3534.

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1,

Behseta, S., Kass, R. E., Moorman, D., and Olson, C. R. (2007). "Testing equality of several functions: Analysis of single-unit firing rate curves across multiple experimental conditions." *Statistics in Medicine*, 26, 21, 3958–3975.

Behseta, S., Kass, R., and Wallstrom, G. (2005). "Hierarchical models for assessing variability among functions." *Biometrika*, 92, 419–434.

Bengio, Y. and Granvalet, Y. (2004). "No unbiased estimator of the variance of K-fold cross-validation." *Journal of Machine Learning Research*, 5, 1089–1105.

Benjamini, Y. and Yekutieli, D. (2001). "The control of the false discovery rate in multiple testing under dependency." *Annals of Statistics*, 29, 1165–1188.

Bernoullli, J. (1713). *Ars conjectandi*. Basel; Thurnisiorum.

Bickel, P. and Doksum, K. (2001). *Mathematical Statistics: Basic Ideas and Selected Topics*, vol. 1. Prentice Hall.

Billingsley, P. (1995). *Probability and Measure*. 3rd ed. New York: Wiley.

Bliss, C. (1936). "The size factor in the action of arsenic upon silkworm larvae." *Journal of Experimental Biology*, 13, 95–110.

Bloomfield, P. (2000). *Fourier Analysis of Time Series*. Wiley.

Bollen, K. (2002). "Latent variables in psychology and the social sciences." *Annual Review of Psychology*, 53, 605–634.

Box, G., Jenkins, G., and Reinsel, G. (2008). *Time Series Analysis: Forecasting and Control*. 4th ed. Wiley.

Box, G. E. P. (1979). "Robustness in the strategy of scientific model building." In *Robustness in Statistics*, eds. R. Launer and G. Wilkinson. New York: Academic Press.

Brillinger, D. (1972). "The spectral analysis of stationary interval functions." *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 483–513.

Brillinger, D. (1981). *Time Series: Data Analysis and Theory*. Expanded ed. Holden Day.

Brillinger, D. (2002). "John W. Tukey: His life and professional contributions." *The Annals of Statistics*, 30, 1535–1575.

Brion, M., Lawlor, D., Matijasevich, A., Horta, B., Anselmi, L., Araújo, C., Menezes, A., Victora, C., and Smith, G. (2011). "What are the causal effects of breastfeeding on IQ, obesity and blood pressure? Evidence from comparing high-income with middle-income cohorts." *International Journal of Epidemiology*, 40, 670–680.

Brockwell, A., Kass, R., and Schwartz, A. (2007). "Statistical signal processing and the motor cortex." *Proceedings of the IEEE*, 95, 881–898.

Brovelli, A., Ding, M., Ledberg, A., Chen, Y., Nakamura, R., and Bressler, S. (2004). "Beta oscillations in a large-scale sensorimotor cortical network: Directional influences revealed by Granger causality." *Proceedings of the National Academy of Sciences*, 101, 9849–9854.

Brown, E. N., Barbieri, R., Eden, U., and Frank, L. (2003). "Likelihood methods for neural data analysis." In *Computational Neuroscience: A comprehensive approach*, ed. J. Feng, chap. 9, 253–286. London: CRC.

Brown, E. N., Frank, L. M., Tang, D., Quirk, M. C., and Wilson, M. A. (1998). "A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells." *Journal of Neuroscience*, 18: 7411–7425.

Brown, E. N. and Kass, R. E. (2009). "What is statistics? (with discussion)." *American Statistician*, 63, 105–123.

Brown, E. N., Kass, R. E., and Mitra, P. (2004). "Multiple neural spike trains analysis: state-of-the-art and future challenges." *Nature Neuroscience*, 7, 456–461.

Brownlee, K. (1965). *Statistical Theory and Methodology in Science and Engineering*. Wiley.

Bundesen, C. (1998). "A computational theory of visual attention." *Philosophical Transactions of the Royal Society B, London*, 353, 1271–1281.

Button, K., Ioannidis, J., Mokrysz, C., Nosek, B., Flint, J., Robinson, E., and Munafó, M. (2013). "Power failure: Why small sample size undermines the reliability of neuroscience." *Nature Reviews Neuroscience*, 14, 365–376.

Casey, B., Somerville, L., Gotlib, I., Ayduk, O., Franklin, N., Askren, M., Jonides, J., Berman, M., Wilson, N., Teslovich, T., Glover, G., Zayas, V., Mischel, W., and Shoda, Y. (2011). "Behavioral and neural correlates of delay of gratification 40 years later." *Proceedings of the National Academy of Sciences*, 108, 14998–15003.

Chaumon, M., Schwartz, D., and Tallon-Baudry, C. (2009). "Unconscious learning versus visual perception: Dissociable roles for gamma oscillations revealed in MEG." *Journal of Cognitive Neuroscience*, 21, 2287–2299.

Chen, C. (1985). "On asymptotic normality of limiting density functions with Bayesian implications." *Journal of the Royal Statistical Society. Series B*, 47, 540–546.

Churchland, A.K. Kiani, R., Chaudhuri, R., Wang, X., Pouget, A., and Shadlen, M. (2011). "Variance as a signature of neural computations during decision-making." *Neuron*, 69, 818–831.

Cleveland, W., Diaconis, P., and McGill, R. (1982). "Variables on scatterplots look more highly correlated when the scales are increased." *Science*, 216, 1138–1141.

Colquhoun, D. (2007). "Classical perspective: What have we learned from single ion channels?" *The Journal of Physiology*, 581, 425–427.

Colquhoun, D. and Sakmann, B. (1985). "Fast events in single-channel currents activated by acetylcholine and its analogues at the frog muscle end-plate." *Journal of Physiology*, 369, 501–557.

Courant, R. and Robbins, H. (1996). *What is Mathematics?*. 2nd ed. revised by Ian Stewart: Oxford.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. New York: Wiley.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.

Dantzig, T. (1954). *Number: The Language of Science*. 4th ed. Doubleday.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer Texts in Statistics. Springer.

Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and Their Applications*. 2nd ed. Cambridge University Press.

Dayan, P. and Abbott, L. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press.

del Castillo, J. and Katz, B. (1954). "Quantal components of the end-plate potential." *The Journal of Physiology*, 124, 560–573.

Devlin, B., Fienberg, S., Resnick, D., and Roeder, K., eds. (1997). *Intelligence, Genes, & Success: Scientists Respond to The Bell Curve*. New York: Copernicus (Springer-Verlag).

DiCiccio, T. and Efron, B. (1996). "Bootstrap confidence intervals." *Statistical Science*, 11, 189–228.

DiMatteo, I., Genovese, C., and Kass, R. E. (2001). "Bayesian curve-fitting with free-knot splines." *Biometrika*, 88, 1051–1077.

Dinstein, I. (2008). "Human cortex: Reflections of mirror neurons." *Current Biology*, 18, R956–R959.

Dinstein, I., Thomas, C., Humphreys, K., Minshew, N., Behrmann, M., and Heeger, D. (2010). "Normal movement selectivity in autism." *Neuron*, 66, 461–469.

Edwards, W., Lindman, H., and Savage, L. (1963). "Bayesian statistical inference for psychological research." *Psychological Review*, 70, 193–242.

Efron, B. (1979a). "Bootstrap methods: Another look at the jackknife." *The Annals of Statistics*, 7, 1–26.

Efron, B. (1979b). "Computers and the theory of statistics: Thinking the unthinkable." *SIAM Review*, 21, 460–480.

Efron, B. (2004). "The estimation of prediction error: Covariance penalties and cross-validation (with discussion)." *Journal of the American Statistical Association*, 99, 619–642.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall.

Ernst, M. and Banks, M. (2002). "Humans integrate visual and haptic information in a statistically optimal fashion." *Nature*, 415, 429–433.

Faber, D. and Korn, H. (1991). "Applicability of the coefficient of variation method for analyzing synaptic plasticity." *Biophysical Journal*, 60, 1288–1294.

Fan, J. and Kreutzberger, E. (1998). "Automatic local smoothing for spectral density estimation." *Scandinavian Journal of Statistics*, 25, 359–369.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*. 3rd ed. New York: Wiley.

Feynman, R. P., Leighton, R. B., and Sands, M. (1963). "Lectures on Physics, Vol. I." *Addison-Wesley*, 49–1.

Fienberg, S. E. (2006). "When did Bayesian inference become "Bayesian?"." *Bayesian Analysis*, 1, 1–40.

Fisher, R. A. (1922). "On the mathematical foundations of theoretical statistics." *Philosophical Transactions of the Royal Society, A*, 222, 309–368.

Fisher, R. A. (1924). "On a distribution yielding the error functions of several well known statistics." In *Proceedings of the international congress of mathematics*, vol. 2, 805–813. Toronto.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Hafner Press.

Fisher, R. (1935). "Statistical tests'." *Nature*, 136, 474.

Formisano, E., Di Salle, F., and Goebel, R. (2006). "Fundamentals of data analysis methods in fMRI." In *Advanced imaging processing in magnetic resonance imaging*, eds. L. Landini, V. Positano, and M. Santarelli, 481–503. Boca Raton (FL): CRC Taylor & Francis.

Fox, M., Snyder, A., Vincent, J., Corbetta, M., Van Essen, D., and Raichle, M. (2005). "The human brain is intrinsically organized into dynamic, anticorrelated functional networks." *Proc. National Acad. Sciences*, 102, 9673–9678.

Francq, C. and Zakoiän, J.-M. (2005). "A central limit theorem for mixing triangular arrays of variables whose dependence is allowed to grow with the sample size." *Econometric Theory*, 21, 1165–1171.

Frank, L. M., Eden, U. T., Solo, V., Wilson, M. A., and Brown, E. N. (2002). "Contrasting patterns of receptive field plasticity in the hippocampus and the entorhinal cortex: An adaptive filtering approach." *The Journal of Neuroscience*, 22, 3817–3830.

Freedman, D., Pisani, R., and Purves, R. (2007). *Statistics*. 4th ed. New York: W.W. Norton.

Frezza, M., di Padova, C., Pozzato, G., Terpin, M., Baraona, E., and Lieber, C. S. (1990). "High blood alcohol levels in women." *New England Journal of Medicine*, 322, 95–99.

Gagliardo, A., Ioaleé, P., Odetti, F., Bingman, V., Siegel, J., and Vallortigara, G. (2001). "Hippocampus and homing in pigeons: left and right hemispheric differences in navigational map learning." *Eur J Neruosci*, 13, 1617–1624.

Gasser, T. and Muller, H. (1984). "Estimating regressive functions and their derivatives by the kernel method." *Scandinavian Journal of Statistics*, 11, 171–185.

Gaunt, P. and Lambert, B. (1987). "Single dose ciprofloxacin for the eradication of pharyngeal carriage of Neisseria meningitidis." *Journal of Antimicrobial Chemotherapy*, 21, 489–496.

Geisler, W. S. (2011). "Contributions of ideal observer theory to vision research." *Vision Research*, 51, 771–781.

Geman, S. and Geman, D. (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 6, 721–741.

Genovese, C. R., Lazar, N. A., and Nichols, T. (2002). "Thresholding of statistical maps in functional neuroimaging using the false discovery rate." *NeuroImage*, 15, 870–878.

Georgopoulos, A. P. and Ashe, J. (2000). "One motor cortex, two different views." *Nature Neuroscience*, 3, 963–965.

Georgopoulos, A. P., Kalaska, J. F., Caminiti, R., and Massey, J. T. (1982). "On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex." *The Journal of Neuroscience*, 2, 1527–1537.

Gerstein, G. and Mandelbrot, B. (1964). "Random walk models for the spike activity of a single neuron." *Biophysical Journal*, 4, 41–68.

Geurts, H. M., Verté, S., Oosterlaan, J., Roeyers, H., and Sergeant, J. A. (2004). "How specific are executive functioning deficits in attention deficit hyperactivity disorder and autism?" *Journal of Child Psychology and Psychiatry*, 45, 836–854.

Geweke, J. (1982). "Measurement of linear dependence and feedback between multiple time series." *Journal of the American Statistical Association*, 77, 304–313.

Gilovich, T., Vallone, R., and Tversky, A. (1985). "The hot hand in basketball: On the misperception of random sequences." *Cognitive Psychology*, 17, 295–314.

Glover, G. H. (1999). "Deconvolution of impulse response in event-related BOLD fMRI." *NeuroImage*, 9, 416–429.

Goebel, R., Esposito, R., and Formisano, E. (2006). "Analysis of FIAC data with BrainVoyager QX: From single-subject to cortically aligned group GLM analysis and self-organizing group ICA." *Human Brain Mapping*, 27, 392–401.

Gold, G., Giannakopoulos, P., Montes-Paixao, C., Herrman, F., Mulligan, R., Michel, J., and Bouras, C. (1997). "Sensitivity and specificity of newly proposed clinical criteria for possible vascular dementia." *Neurology*, 49, 690–694.

Goodman, S. N. (1999a). "Toward evidence-based medical statistics. 1: The *P* value fallacy." *Annals of Internal Medicine*, 130, 995–1004.

Goodman, S. N. (1999b). "Toward evidence-based medical statistics. 2: The Bayes factor". *Annals of Internal Medicine*, 130, 1005–1013.

Gordon, A., Glazko, G., Qiu, X., and Yakovlev, A. (2007). "Control of the mean number of false discoveries, Bonferroni and stability of multiple testing." *Annals of Applied Statistics*, 1, 179–190.

Gould, S. (1996). *The Mismeasure of Man*. Norton.

Grace, A. A., Floresco, S. B., Goto, Y., and Lodge, D. J. (2007). "Regulation of firing of dopaminergic neurons and control of goal-directed behaviors." *Trends in Neurosciences*, 220–227.

Granger, C. (1969). "Investigating causal relations by econometric models and cross-spectral methods". *Econometrica*, 37, 424–438.

Greenhouse, J. B., Kass, R. E., and Tsay, R. S. (1987). "Fitting nonlinear models with ARMA errors to biological rhythm data." *Statistics in Medicine*, 6, 167–183.

Griffiths, T. L., Chater, N., Norris, D., and Pouget, A. (2012). "How the Bayesians got their beliefs (and what those beliefs actually are)." *Psychological Bulletin*, 138, 415–422.

Harrison, M., Amarasingham, A., and Kass, R. (2013). "Statistical identification of synchronous spiking." In *Spike Timing: Mechanisms and Function*, eds. P. DiLorenzo and J. Victor. Taylor and Francis, pp. 77–120.

Hartline, H. and Graham, C. (1932). "Nerve impulses from single receptors in the eye." *Journal of Cellular Comparative Physiology*, 1, 227–295.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer.

Hastings, W. K. (1970). "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika*, 57, 97–109.

Hawkins, D. (1989). "Using U statistics to derive the asymptotic distribution of Fisher's Z statistic." *The American Statistician*, 43, 235–237.

Hawkins, T. (2001). *Lebesgue's Theory of Integration: Its Origins and Development*. American Mathematical Society.

Hébert, R. and Brayne, C. (1995). "Epidemiology of vascular dementia." *Neuroepidemiology*, 14, 240–257.

Hecht, S., Shlaer, S., and Pirenne, M. H. (1942). "Energy, quanta, and vision." *The Journal of General Physiology*, 25, 819–840.

Hill, A. (1971). *Principles of medical statistics*. 9th ed. Oxford University Press.

Hoeft, F., McCandliss, B., Black, J., Gantman, A., Zakerani, N., Hulme, C., Lyytinen, H., Whitfield-Gabrieli, S., Glover, G., Reiss, A., and Gabrieli, J. (2011). "Neural systems predicting long-term outcome in dyslexia." *Proceedings of the National Academy of Sciences*, 108, 361–366.

Hursh, J. B. (1939). "Conduction velocity and diameter of nerve fibers." *American Journal of Physiology*.

Ikegaya, Y., Aaron, G., Cossart, R., Aronov, D., Lampl, I., Ferster, D., and Yuste, R. (2004). "Synfire chains and cortical songs: Temporal modules of cortical activity." *Science*, 304, 559–564.

Ikegaya, Y., Matsumoto, W., Chiou, H.-Y., Yuste, R., and Aaron, G. (2008). "Statistical significance of precisely repeated intracellular synaptic patterns." *PloS ONE*, 3, 12, e3983.

Iyengar, S. and Liao, Q. (1997). "Modeling neural activity using the generalized inverse Gaussian distribution." *Biological Cybernetics*, 77, 289–295.

Jacobs, R. A. and Kruschke, J. K. (2010). "Bayesian learning theory applied to human cognition." *Wiley Interdisciplinary Reviews: Cognitive Science*, 2, 8–21.

Jarosiewicz, B., Chase, S. M., Fraser, G. W., Velliste, M., Kass, R. E., and Schwartz, A. B. (2008). "Functional network reorganization during learning in a brain-computer interface paradigm." *Proceedings of the National Academy of Sciences*, 105, 19486–19491.

Jeffreys, H. (1931). *Scientific Inference*. Cambridge University Press.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press.

Jeffreys, H. and Wrinch, D. (1921). "On certain fundamental principles of scientific inquiry." *Philosophical Magazine*, 42, 369–390.

Kalman, R. E. (1960). "A new approach to linear filtering and prediction problems." *Journal of Basic Engineering*, 82, 35–45.

Karpicke, J. D. and Roediger, H. L. (2008). "The critical importance of retrieval for learning." *Science*, 319, 966–968.

Kass, R. E. (2011). "Statistical inference: the big picture (with discussion)." *Statistical Science*, 26, 1–20.

Kass, R. E., Kelly, R., and Loh, W.-L. (2011). "Assessment of synchrony in multiple neural spike trains using loglinear point process models." *Annals of Applied Statistics*, 5, 1262–1292.

Kass, R. E. and Natarajan, R. (2006). "A default conjugate prior for variance components in generalized linear mixed models (comment on article by Browne and Draper)." *Bayesian Analysis*, 1, 535–542.

Kass, R. E. and Raftery, A. E. (1995). "Bayes factors." *Journal of the American Statistical Association*, 90, 773–795.

Kass, R. E. and Steffey, D. (1989). "Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models)." *Journal of the American Statistical Association*, 84, 717–726.

Kass, R. E. and Ventura, V. (2001). "A spike-train probability model." *Neural Computation*, 13, 8, 1713–1720.

Kass, R. E., Ventura, V., and Brown, E. (2005). "Statistical issues in the analysis of neuronal data." *Journal of Neurophysiology*, 94, 8–25.

Kass, R. E., Ventura, V., and Cai, C. (2003). "Statistical smoothing of neuronal data." *Network-Computation in Neural Systems*, 14, 5–16.

Kass, R. E. and Vos, P. W. (1997). *Geometrical Foundations of Asymptotic Inference*. Wiley.

Kass, R. E. and Wasserman, L. (1995). "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion." *Journal of the American Statistical Association*, 90, 928–934.

Kass, R. E. and Wasserman, L. (1996). "The selection of prior distributions by formal rules." *Journal of the American Statistical Association*, 91, 1343–1370.

Kaufman, C. G., Ventura, V., and Kass, R. E. (2005). "Spline-based non-parametric regression for periodic functions and its application to directional tuning of neurons." *Statistics in Medicine*, 24, 2255–2265.

Kay, K. N., Naselaris, T., Prenger, R. J., and Gallant, J. L. (2008). "Identifying natural images from human brain activity." *Nature*, 452, 352–355.

Kelly, R. and Kass, R. E. (2012). "A framework for evaluating pairwise and multiway synchrony among stimulus-driven neurons." *Neural Computation*, 24, 2007–2032.

Kelly, R. C., Smith, M. A., Kass, R. E., and Lee, T. S. (2010). "Local field potentials indicate network state and account for neuronal response variability." *Journal of Computational Neuroscience*, 29, 567–579.

Kelly, R. C., Smith, M. A., Samonds, J. M., Kohn, A., Bonds, A., Movshon, J. A., and Lee, T. S. (2007). "Comparison of recordings from microelectrode arrays and single electrodes in the visual cortex." *The Journal of Neuroscience*, 27, 261–264.

Kempthorne, O. (1957). *An Introduction to Genetic Statistics*. Wiley.

Kent, L., Middle, F., Hawi, Z., Fitzgerald, M., Gill, M., Feehan, C., and Craddock, N. (2001). "Nicotinic acetylcholine receptor [alpha] 4 subunit gene polymorphism and attention deficit hyperactivity disorder." *Psychiatric Genetics*, 11, 37–40.

Klingsberg, T., Fernell, E., Olesen, P., Johnson, M., Gustafsson, P., Dahlstrom, K., Gillberg, C., Forssberg, H., and Westerberg, H. (2005). "Computerized training of working memory in children with ADHD - A randomized controlled trial." *Journal of the American Academy of Child and Adolescent Psychiatry*, 44, 177–186.

Knill, D. C. and Pouget, A. (2004). "The Bayesian brain: The role of uncertainty in neural coding and computation." *TRENDS in Neurosciences*, 27, 712–719.

Kolers, P. A. (1976). "Reading a year later." *Journal of Experimental Psychology: Human Learning and Memory*, 2, 554–565.

Kolmogorov, A. N. (1933). *Grundbegriffe der wahrscheinlichkeitsrechnung*. Springer-Verlag.

Konishi, S. and Kitagawa, G. (2007). *Information Criteria and Statistical Modeling*. Springer.

Körding, K. (2007). "Decision theory: What "should" the nervous system do?" *Science*, 318, 606–610.

Körding, K. P. and Wolpert, D. M. (2004). "Bayesian integration in sensorimotor learning." *Nature*, 427, 244–247.

Koyama, S., Chase, S. M., Whitford, A. S., Velliste, M., Schwartz, A. B., and Kass, R. E. (2010). "Comparison of brain-computer interface decoding algorithms in open-loop and closed-loop control." *Journal of Computational Neuroscience*, 29, 73–87.

Koyama, S. and Kass, R. E. (2008). "Spike train probability models for stimulus-driven leaky integrate-and-fire neurons." *Neural Computation*, 20, 1776–1795.

Kriegeskorte, N., Lindquist, M. A., Nichols, T. E., Poldrack, R. A., and Vul, E. (2010). "Everything you never wanted to know about circular analysis, but were afraid to ask." *J. Cereb. Blood Flow Metab.*, 30, 1551–1557.

Kullingsbaek, S. (2006). "Modeling visual attention." *Behavioral Research Methods*, 38, 123–133.

Kwon, H., Reiss, A. L., and Menon, V. (2002). "Neural basis of protracted developmental changes in visuo-spatial working memory." *Proceedings of the National Academy of Sciences*, 99, 13336–13341.

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). "Penalized regression, standard errors, and Bayesian lassos." *Bayesian Analysis*, 5, 369–411.

Lahiri, S. (2003a). "A necessary and sufficient condition for asymptotic independence of discrete Fourier transforms under short-and long-range dependence." *The Annals of Statistics*, 31, 613–641.

Lahiri, S. N. (2003b). *Resampling Methods for Dependent Data*. Springer.

Lanczos, C. (1966). *Discourse on Fourier series*, vol. 3. Edinburgh: Oliver & Boyd.

Levine, M. (1991). "The distribution of intervals between neural impulses in the maintained discharges of retinal ganglion cells." *Biological Cybernetics*, 65, 459–467.

Lewicki, M. (1998). "A review of methods for spike sorting: The detection and classification of neural action potentials." *Network: Computation in Neural Systems*, 9, R53–R78.

Lewicki, M. (2002). "Efficient coding of natural sounds." *Nature Neuroscience*, 5, 356–363.

Lewis, S. M., Jerde, T. A., Tzagarakis, C., Gourtzelidis, P., Georgopoulos, M.-A., Tsekos, N., Amirikian, B., Kim, S.-G., Uğurbil, K., and Georgopoulos, A. P. (2005). "Logarithmic transformation for high-field BOLD fMRI data data." *Experimental Brain Research*, 165, 447–453.

Li, D., Held, U., Petkau, J., Daumer, M., Barkhof, F., Fazekas, F., Frank, J., Kappos, L., Miller, D., Simon, J., et al. (2006). "MRI T2 lesion burden in multiple sclerosis: A plateauing relationship with clinical disability." *Neurology*, 66, 1384–1389.

Loader, C. (1999). *Local regression and likelihood*. New York: Springer.

Logothetis, N., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). "Neurophysiological investigation of the basis of the fMRI signal." *Nature*, 412, 150–157.

Lucas, A., Morley, R., Cole, T., Lister, G., and Leeson-Payne, C. (1992). "Breast milk and subsequent intelligence quotient in children born preterm." *The Lancet*, 339, 261–264.

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2012). *The BUGS Book: A Practical Introduction to Bayesian Analysis*, Chapman and Hall, CRC Press.

MacKay, D. J. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Madiman, M. and Barron, A. (2007). "Generalized entropy power inequalities and monotonicity properties of information." *Information Theory, IEEE Transactions on*, 53, 2317–2329.

Makris, S. L., Raffaele, K., Allen, S., Bowers, W. J., Hass, U., Alleva, E., Calamandrei, G., Sheets, L., Amcoff, P., Delrue, N., et al. (2009). "A retrospective performance assessment of the developmental neurotoxicity study in support of OECD Test Guideline 426." *Environmental Health Perspectives*, 117, 17–25.

Manly, B. F. (2007). *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman & Hall/CRC.

Marshall, J. C. and Halligan, P. W. (1988). "Blindsight and insight in visuo-spatial neglect." *Nature*, 336, 766–767.

Marshel, J., Garrett, M., Nauhaus, I. and Callaway. E. (2011). "Functional specialization of seven mouse visual cortical areas." *Neuron*, 72, 1040–1054.

Matsuzaka, Y., Picard, N., and Strick, P. L. (2007). "Skill representation in the primary motor cortex after long-term practice." *Journal of Neurophysiology*, 97, 1819–1832.

Mayo, D. G. and Spanos, A. (2010). *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*. Cambridge University Press.

McCullagh, P. and Nelder, J. (1989). *General linear models. Chapman and Hall*.

McGrayne, S. (2011). *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*. Yale University Press.

Metropolis, N. (1987). "The beginning of the Monte Carlo method." *Los Alamos Science (Special Issue dedicated to Stanislaw Ulam)*, 15, 125–130.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). "Equation of state calculations by fast computing machines." *The Journal of Chemical Physics*, 21, 1087–1091.

Mitra, P. P. and Pesaran, B. (1999). "Analysis of dynamic brain imaging data." *Biophysical Journal*, 76, 691–708.

Mosteller, F. and Tukey, J. (1968). *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley.

Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression*. Addison-Wesley.

Mudholkar, G. S. and Tian, L. (2002). "An entropy characterization of the inverse Gaussian distribution and related goodness-of-fit test." *Journal of Statistical Planning and Inference*, 102, 211–221.

Mullins, P. G., Rowland, L. M., Jung, R. E., and Sibbitt, W. L. (2005). "A novel technique to study the brain's response to pain: Proton magnetic resonance spectroscopy." *NeuroImage*, 26, 642–646.

Nagelkerke, N. (1991). "A note on a general definition of the coefficient of determination." *Biometrika*, 78, 691–692.

Nelder, J. A. and Wedderburn, R. W. (1972). "Generalized linear models." *Journal of the Royal Statistical Society*. Series A, 135, 370–384.

Neyman, J. (1937). "Outline of a theory of statistical estimation based on the classical theory of probability." *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236, 333–380.

Nielsen, T. A., DiGregorio, D. A., and Silver, R. A. (2004). "Modulation of glutamate mobility reveals the mechanism underlying slow-rising AMPAR EPSCs and the diffusion coefficient in the synaptic cleft." *Neuron*, 42, 757–771.

Nieuwenhuis, S., Forstmann, B., and Wagenmakers, E.-J. (2011). "Erroneous analyses of interactions in neuroscience: A problem of significance." *Nature Neuroscience*, 14, 1105–1107.

Olson, C. R., Gettner, S. N., Ventura, V., Carta, R., and Kass, R. E. (2000). "Neuronal activity in macaque supplementary eye field during planning of saccades in response to pattern and spatial cues." *Journal of Neurophysiology*, 84, 1369–1384.

Ombao, H. and Van Bellegem, S. (2008). "Evolutionary coherence of nonstationary signals." *Signal Processing, IEEE Transactions on*, 56, 2259–2266.

Optican, L. M. and Richmond, B. J. (1987). "Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. III. Information theoretic analysis." *Journal of Neurophysiology*, 57, 162–178.

Pearson, K. and Lee, A. (1903). "On the laws of inheritance in man." *Biometrika*, 2, 357–462.

Percival, D. and Walden, A. (2000). *Wavelet Methods for Time Series Analysis*. Cambridge University Press.

Percival, D. B. and Constantine, W. L. (2006). "Exact simulation of Gaussian time series from nonparametric spectral estimates with application to bootstrapping." *Statistics and Computing*, 16, 25–35.

Percival, D. B. and Walden, A. T. (1993). *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques*. Cambridge University Press.

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E., and Simoncelli, E. P. (2008). "Spatio-temporal correlations and visual signalling in a complete neuronal population." *Nature*, 454, 995–999.

Platt, M. L. and Glimcher, P. W. (1999). "Neural correlates of decision variables in parietal cortex." *Nature*, 400, 233–238.

Qian, A., Buller, A. L., and Johnson, J. W. (2005). "NR2 subunit-dependence of NMDA receptor channel block by external Mg2+." *The Journal of Physiology*, 562, 319–331.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer.

Roediger, H. L. and Karpicke, J. D. (2006). "Test-Enhanced Learning." *Psychological Science*, 17, 249–255.

Roesch, M. R. and Olson, C. R. (2004). "Neuronal Activity Related to Reward Value and Motivation in Primate Frontal Cortex." *Science*, 304, 307–310.

Rollenhagen, J. and Olson, C. (2005). "Low-frequency oscillations arising from competitive interactions between visual stimuli in macaque inferotemporal cortex." *J. Neurophysiology*, 94, 3368–3387.

Ross, S. M. (1996). *Stochastic Processes*. Wiley Series in Probability and Statistics. Wiley.

Rutherford, E., Chadwick, J., and Ellis, C. D. (1920). *Radiations from Radioactive Substances*. Cambridge University Press.

Santello, M., Flanders, M., and Soechting, J. (1998). "Postural hand synergies for tool use." *The Journal of Neuroscience*, 18, 10105–10115.

Sarma, S. V., Cheng, M. L., Eden, U., Williams, Z., Brown, E. N., and Eskanda, E. (2012). "The effects of cues on neurons in the basal ganglia in Parkinson's disease." *Frontiers in Integrative Neuroscience*, 6, 1–12.

Saygin, Z. M., Osher, D. E., Koldewyn, K., Reynolds, G., Gabrieli, J. D., and Saxe, R. R. (2011). "Anatomical connectivity patterns predict face selectivity in the fusiform gyrus." *Nature Neuroscience*, 15, 321–327.

Schervish, M. (1995). *Theory of Statistics*. Springer-Verlag.

Schreiber, T. and Schmitz, A. (2000). "Surrogate time series." *Physica D*, 142, 346–382.

Schultz, H. (1929). "Applications of the theory of error to the interpretation of trends." *J. Amer. Statist. Assoc., Supp: Proc. Amer. Statisti. Assoc.*, 24, 86–89.

Schwarz, G. (1978). "Estimating the dimension of a model." *The Annals of Statistics*, 6, 2, 461–464.

Scott, C. C. and Chen, K. (1944). "Comparison of the action of 1-ethyl theobromine and caffeine in animals and man." *Journal of Pharmacology and Experimental Therapeutics*, 82, 89–97.

Scott, D. W. (1992). *Multivariate Density Estimation*, vol. 1. Wiley.

Sellke, T., Bayarri, J., and Berger, J. (2001). "Calibration of p-values for testing precise hypotheses." *American Statistician*, 55, 62–71.

Shadlen, M. N. and Newsome, W. T. (1998). "The variable discharge of cortical neurons: Implications for connectivity, computation, and information coding." *The Journal of Neuroscience*, 15, 3870–3896.

Shruti, S., Clem, R. L., and Barth, A. L. (2008). "A seizure-induced gain-of-function in BK channels is associated with elevated firing activity in neocortical pyramidal neurons." *Neurobiology of disease*, 30, 323–330.

Shumway, R. H. and Stoffer, D. S. (2006). *Time Series Analysis and Its Applications, with R Examples*. Springer New York.

Simmons, J., Nelson, L., and Simonsohn, U. (2011). "False-positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant." *Psychological Science*, 22, 1359–1366.

Sklar, R. and Strauss, B. (1980). "Role of the *uvrE* gene product and of inducible *O6* -methylguanine removal in the induction of mutations by N-methyl- N-nitro- N-nitrosoguanidine in *Escherichia coli*." *Journal of molecular biology*, 143, 343–362.

Smith, A. C., Stefani, M. R., Moghaddam, B., and Brown, E. N. (2005). "Analysis and design of behavioral experiments to characterize population learning." *Journal of Neurophysiology*, 93, 1776–1792.

Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., Ramsey, J. D., and Woolrich, M. W. (2011). "Network modelling methods for fMRI." *Neuroimage*, 54, 875–891.

Solt, K., Cotten, J., Cimenser, A., Wong, K., Chemali, J., and Brown, E. (2011). "Methylphenidate actively induces emergence from general anesthesia." *Anesthesiology*, 115, 791–803.

Sperling, G. (1967). "Successive approximations to a model for short term memory." *Acta psychologica*, 27, 285–292.

Stefani, M. R., Groth, K., and Moghaddam, B. (2003). "Glutamate receptors in the rat medial prefrontal cortex regulate set-shifting ability." *Behavioral Neuroscience*, 117, 728–737.

Stevens, S. (1961). "To Honor Fechner and Repeal His Law." *Science*, 133, 80–86.

Stevens, S. (1970). "Neural events and the psychophysical law." *Science*, 170, 1043–1050.

Stigler, S. M. (1986). *The History of Statistics. The Measurement of Uncertainty before 1900*. Cambridge, Mass.: Harvard.

Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions (with discussion)." *Journal of the Royal Statistical Society. Series B (Methodological)*, 36, 111–147.

Stopfer, M., Jayaraman, V., and Laurent, G. (2003). "Intensity versus identity coding." *Neuron*, 39, 991–1004.

Teich, M. C., Prucnal, P. R., Vannucci, G., Breton, M. E., and McGill, W. J. (1982). "Multiplication noise in the human visual system at threshold: 3. The role of non-Poisson quantum fluctuations." *Biological Cybernetics*, 44, 157–165.

Thompson, J. A., Wu, W., Bertram, R., and Johnson, F. (2007). "Auditory-dependent vocal recovery in adult male zebra finches is facilitated by lesion of a forebrain pathway that includes the basal ganglia." *The Journal of Neuroscience*, 27, 12308–12320.

Thomson, D. J. (1982). "Spectrum estimation and harmonic analysis." *Proceedings of the IEEE*, 70, 1055–1096.

Tibshirani, R. (2011). "Regression shrinkage and selection via the lasso: A retrospective (with discussion)." *J. Royal Statist. Soc. B*, 73, 273–282.

Tokdar, S., Xi, P., Kelly, R. C., and Kass, R. E. (2010). "Detection of bursts in extracellular spike trains using hidden semi-Markov point process models." *Journal of Computational Neuroscience*, 29, 203–212.

Tuckwell, H. (1988). *Introduction to Theoretical Neurobiology*, vol. 2: Nonlinear and Stochastic Theories. Cambridge.

Tukey, J. (1987). *The Collected Works of John W. Tukey*, vol. 4. Wadsworth.

Turner, R. and DeLong, M. (2000). "Corticostriatal activity in primary motor cortex of the macaque." *Journal of Neuroscience*, 20, 7096–7198.

Uhlhaas, P. J., Pipa, G., Lima, B., Melloni, L., Neuenschwander, S., Nikolić, D., and Singer, W. (2009). "Neural synchrony in cortical networks: History, concept and current status." *Frontiers in Integrative Neuroscience*, 3.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

Vandenbergh, R., Nelissen, N., Salmon, E., Ivanoiu, A., Hasselbalch, S., Andersen, A., Korner, A., Minthon, L., Brooks, D., Van Laere, K., and Dupont, P. (2013). "Binary classification of ^{18}F-flutemetamol PET using machine learning: Comparison with visual reads and structural MRI". Neuroimage, 64, 57–25.

Ventura, V., Cai, C., and Kass, R. (2005a). "Statistical assessment of time-varying dependence between two neurons." *J. Neurophys.*, 94, 2940–2947.

Ventura, V., Cai, C., and Kass, R. E. (2005b). "Trial-to-trial variability and its effect on time-varying dependency between two neurons." *Journal of neurophysiology*, 94, 2928–2939.

Ventura, V., Carta, R., Kass, R., Gettner, S., and Olson, C. (2002). "Statistical analysis of temporal evolution in single-neuron firing rates." *Biostatistics*, 3, 1–20.

Vu, V. Q., Ravikumar, P., Naselaris, T., Kay, K. N., Gallant, J. L., and Yu, B. (2011). "Encoding and decoding V1 fMRI responses to natural images with sparse nonparametric models." *The Annals of Applied Statistics*, 5, 1159–1182.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., and Grasman, R. (2010). "Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method." *Cognitive psychology*, 60, 158–189.

Wallstrom, G. L., Kass, R. E., Miller, A., Cohn, J. F., and Fox, N. A. (2002). "Correction of ocular artifacts in the EEG using Bayesian adaptive regression splines." In *Case Studies in Bayesian, Statistics*, 351–366. Springer-Verlag.

Wallstrom, G. L., Kass, R. E., Miller, A., Cohn, J. F., and Fox, N. A. (2004). "Automatic correction of ocular artifacts in the EEG: A comparison of regression-based and component-based methods." *International Journal of Psychophysiology*, 53, 105–119.

Wang, W., Sudre, G. P., Xu, Y., Kass, R. E., Collinger, J. L., Degenhart, A. D., Bagic, A. I., and Weber, D. J. (2010). "Decoding and cortical source localization for intended movement direction with MEG." *Journal of Neurophysiology*, 104, 2451–2461.

Wasserman, L. (2004). *All of Statistics*. Springer.

Watson, R. and Tang, D. (1980). "The predictive value of prostatic acid phosphates as a screening test for prostatic cancer." *New England Journal of Medicine*, 303, 497–499.

Weinberg, S. (2002). *It Must Be Beautiful: Great Equations of Modern Science*, chap. Afterword: How great equations survive. Granta Press.

Welch, P. (1967). "The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms." *Audio and Electroacoustics, IEEE Transactions on*, 15, 70–73.

Whitmore, G. and Seshadri, V. (1987). "A heuristic derivation of the inverse Gaussian distribution." *The American Statistician*, 41, 280–281.

Wolpert, D. M., Diedrichsen, J., and Flanagan, J. R. (2011). "Principles of sensorimotor learning." *Nature Reviews Neuroscience*, 12, 739–751.

Wood, F., Black, M. J., Vargas-Irwin, C., Fellows, M., and Donoghue, J. P. (2004). "On the variability of manual spike sorting." *Biomedical Engineering, IEEE Transactions on*, 51, 912–918.

Wu, C.-F. (1981). "Asymptotic theory of nonlinear least squares estimation." *The Annals of Statistics*, 9, 501–513.

Wu, W., Gao, Y., Bienenstock, E., Donoghue, J. P., and Black, M. J. (2006). "Bayesian population decoding of motor cortical activity using a Kalman filter." *Neural computation*, 18, 80–118.

Xu, Y., Sudre, G. P., Wang, W., Weber, D. J., and Kass, R. E. (2011). "Characterizing global statistical significance of spatio-temporal hot spots in MEG/EEG source space via excursion algorithms." *Statistics in Medicine*, 30, 2854–2866.

Yonelinas, A. (2001). "Consciousness, control, confidence: The 3 Cs of recognition memory." *Journal of Experimental Psychology: General*, 130, 361–379.

Yu, B., Cunningham, J., Santhanam, G., Ryu, S., Shenoy, K., and Sahani, M. (2009). "Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity." *Journal of Neurophysiology*, 102, 614–635.

Yu, C.-E., Seltman, H., Peskind, E. R., Galloway, N., Zhou, P. X., Rosenthal, E., Wijsman, E. M., Tsuang, D. W., Devlin, B., and Schellenberg, G. D. (2007). "Comprehensive analysis of APOE and selected proximate markers for late-onset Alzheimer's disease: Patterns of linkage disequilibrium and disease/marker association." *Genomics*, 89, 655–665.

Yuille, A. and Kersten, D. (2006). "Vision as Bayesian inference: Analysis by synthesis?" *Trends in Cognitive Sciences*, 10, 301–308.

Zelazo, P., Zelazo, N., and Kolb, S. (1972). "'Walking in the newborn." *Science*, 176, 314–315.

Example Index

A

ACT-R theory of procedural memory, 103
Action potential width and the preceding inter-spike interval, 193, 241, 347, 408, 431, 433, 436
Alcohol metabolism among men and women, 382, 384, 389
Allele frequencies in fruit flies, 251, 253
Alzheimer's and APOE (Apolipoprotein E), 254–256, 258
Auditory-dependent vocal recovery in zebra finches, 93

B

Beta oscillations during a sensorimotor task, 553, 557, 561
Beta oscillations in Parkinson's disease, 569, 587
Blindsight in patient P.S., 9, 13, 158, 171, 174, 175, 257, 261, 267, 268, 272, 285, 289–291
BOLD hemodynamic response in fMRI, 313
Brain-machine interface perturbation, 194
Burst detection from spike trains, 458, 470

C

Circadian rhythm in core temperature, 519, 521, 525, 532, 541, 542

D

Decision-making and trial-to-trial variability of spike counts from LIP neurons, 86
Decoding hand movement from cortical activity, 471, 474
Decoding intended movement using MEG, 100, 306, 371, 432, 494
Decoding natural images from V1 fMRI, 426
Decoding of saccade direction from SEF spike counts, 45
Development of motor control, 361, 366, 368, 372, 374
Developmental change in working memory from fMRI, 333, 338
Dual-process theory of memory, 280

E

Ebbinghaus on human memory, 117
EEG spectrogram under general anesthesia, 27, 514, 548
Efficient coding of natural sounds, 504
Electrooculogram smoothing for EEG artifact removal, 16
EMG in frog movement, 34
Emission of alpha particles, 111, 253
EPSC (Excitatory post-synaptic current), 14

F

Finger tapping in response to stimulants, 363, 369, 370
fMRI adaptation among autistic and control subjects, 303, 306
fMRI BOLD hemodynamic response, 340
fMRI BOLD signal and neural activity, 518, 549, 558
fMRI face selectivity prediction using anatomical connectivity, 356

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1,

fMRI in a visuomotor experiment, 6
fMRI network models, 135
Functional specialization of mouse visual areas, 494

G
Gamma oscillations in MEG during learning, 518
Genetic linkage across multiple related strains, 462, 477
Glutamate increase in response to pain, 263
Gratification delay, 379

H
High-field BOLD signal, 29
Hippocampal hemispheric differences among homing pigeons, 269
Hippocampal place cell, 410
Hippocampal place field plasticity, 472, 592

I
Integrate-and-fire model fit to cochlear neuron inter-spike intervals, 127
Intracellular synaptic patterns, 301, 302
Ion channel activation duration, 58, 452
IQ and breast milk, 385, 386, 388
IT neural response, 413

L
Learning impairment following NMDA antagonist injection, 108
Local field potential in primary visual cortex, 421, 516, 528, 547, 551

M
Magical number seven, 97
MEG background noise, 5, 54, 65, 345
Membrane conductance, 109
Methylphenidate-induced emergence from general anesthesia, 451, 478
Miniature excitatory post-synaptic currents, 571
Motor cortical directional tuning, 348
Motor cortical neuron non-cosine directional tuning, 406
Motor cortical spike counts, 46, 138, 164

N
Neural conduction velocity, 11, 324, 326
Neural firing rate selectivity index, 236, 243
Nicotinic acetylcholine receptor and ADHD, 107, 249, 250
NMDA receptor magnesium block, 406

O
Olfactory bulb spike trains, 597
Onset latency in a basal ganglia neuron, 409
Optimal integration of sensory information, 192

P
Perception of light, 112, 213, 392, 394, 396–398
Postural hand synergies, 501
Predicting reading improvement in dyslexic children from fMRI, 510

Q
Quantal response in synaptic transmission, 113

R
Regression of son's height on father's height, 87
Retinal ganglion cell under constant conditions, 296, 569, 586, 596
Reward and parietal cortex neural activity, 310, 326, 333, 338

S
Saccadic reaction time in hemispatial neglect, 24, 28, 69, 378, 393
SEF neuron directional sensitivity, 400, 401
SEF neuronal activity under two conditions, 3, 187, 244, 258, 422, 576, 589, 593
SEF selectivity indices, 459, 462, 464, 468
Sensorimotor learning, 445
Set shifting in ADHD, 481
Skill acquisition power law, 32
Spatiotemporal correlations in visual signaling, 569, 594
Spike count correlation could limit fidelity, 141
Spike sorting forebrain recordings, 502, 511
Square-root transformation of spike counts in motor cortex, 234
Stimulus-response power laws, 32
Synchronous firing of V1 neurons, 284

T
Temporal coding in inferotemporal cortex, 97
Test-enhanced learning, 167, 258, 265, 274, 298, 300, 301
Tetrode spike sorting, 71
Time-varying dependence between spike trains, 277
Time-varying firing rates, 496, 500
Toxicity as a function of dose and weight, 334, 338
Two neurons from primary visual cortex, 38, 39, 42

V
Vascular dementia diagnostic test, 44
Vision as Bayesian decision-making, 102
Visual attention model, 150

W
Working memory in children with ADHD, 380

Index

Symbols
α particles, 111
χ^2_{obs}, 250
$\frac{2}{3}$ rule, 116
95% rule, 116

A
Absolute refractory period, 583
ACF (Autocorrelation function), 322, 515, 531
ACT-R, 103
Action potential, 3, 193, 564
ADHD (Attention deficit hyperactivity disorder), 107, 248, 380, 451, 481, 640
Adrian, Edgar, 563
AIC (Akaike information criterion), 295, 354, 357
Akaike information criterion, *see* AIC
Aliasing, 545
Alignment of theoretical and real worlds, 175
Alternative hypothesis, 248, 276
Alzheimer's disease, 254
Analysis of covariance, *see* ANCOVA
Analysis of variance, *see* ANOVA
ANCOVA (Analysis of covariance), 379, 380
Anesthesia, 27, 451
ANOVA (Analysis of variance), 284, 361
  assumptions, 364
  decomposition, 342, 365
Aperiodic, 453
APOE (Apolipoprotein E), 254
Apolipoprotein E, *see* APOE
Approximate 95 % confidence interval, 160, 166
Approximate coverage probability, 276
AR(p), 530
ARMA model, 540, 558
Association, 8, 328, 387
Asymptotic normality, 180
Asymptotic normality of least squares estimators, 324, 344
Attention, 150
Attention deficit hyperactivity disorder, *see* ADHD
Attenuation of correlation, 330, 465
Augmented data, 216
Autocorrelation coefficient, 322
Autocorrelation function, *see* ACF
Autocovariance function, 515
Autoregressive model, 322, 530
Autoregressive moving average, 540
Axioms of probability, 38

B
B-splines, 420
Backward elimination, 354
Band-pass, 543
BARS (Bayesian Adaptive Regression Splines), 16, 413, 415, 424, 532, 541
Basal ganglia, 409
Basis, 417, 523, 617
Basis functions, 414
Batch of numbers, 25
Bayes classifier, 98
Bayes factor, 297, 476
Bayes sufficient, 442
Bayes' rule, 102
Bayes' theorem, 43, 45, 98, 173, 174
Bayesian, 14
Bayesian Adaptive Regression Splines, *see* BARS
Bayesian decision-making, 102

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1,

Bayesian decoding, 46
Bayesian information criterion, *see* BIC
Bayesian interpretation, 174
Bayesian methods, 175
Bayesian sufficiency, 200
Bell-shaped curve, 25
Bernoulli random variable, 105
Bernoulli trials, 107
Bernoulli, Jacob, 37
Beta distribution, 52, 124, 174
Beta oscillations, 553, 569, 588
Bias, 182
Bias variance trade-off, 183, 432, 434
BIC (Bayesian information criterion), 295, 354, 357
Bimodal, 24
Binary data, 22
Binary events, 115
Binomial distribution, 48, 106
Bivariate dependence, 76
Bivariate normal distribution, 82
Blindsight, 9
Blocks, 370
Blood-oxygen-level dependent, *see* BOLD
BOLD (Blood-oxygen-level dependent), 6, 29, 232, 313, 518
Bonferroni correction, 304, 373
Bootstrap, 145, 222, 237
Bootstrap BC_a confidence intervals, 241, 243
Bootstrap in regression, 344
Bootstrap sampling, 274
Bootstrap test, 274, 300, 385
Box, George, 17, 18
Brain–computer interface, 471
Brain–machine interface, 194
Brownian motion, 126
BUGS, 468
Burn-in, 455
Burst activity, 458, 566
Burst detection, 458

C

Canonical link, 404
Canonical normal hierarchical model, 460
Canonical parameter, 404
Case-control study, 256
Cauchy, Augustin-Louis, 523
Cauchy–Schwartz inequality, 78
Causal effect, 387
Causation, 385
cBARS (circular Bayesian Adaptive Regression Splines), 407
cdf (Cumulative distribution function), 48, 54
Central limit theorem, *see* CLT
Central tendency, 26
Change of variables formula, 62
Change point, 194, 241, 408, 410
Characteristic function, 146
Chi-squared distribution, 124, 131
Chi-squared statistics, 248
CI (Confidence interval), 160, 164
Circadian rhythm, 519
circular Bayesian Adaptive Regression Splines, *see* cBARS
Circular data, 269
Classification, 99
CLT (Central limit theorem), 30, 137, 145, 162, 163, 537
Clustering, 492, 502, 510
Coefficient of variation, 56, 126, 580
Coherence, 555
Combining multiple independent p-values, 301
Common log, 31
Complete covariance function, 600
Complex numbers, 619
Conditional and marginal intensities, 586
Conditional density, 84
Conditional expectation, 85
Conditional maximum likelihood, 534
Conditional probability, 40
Conditionally independent hierarchical model, 460
Conduction velocity, 11
Confidence interval, interpretation, 170
Confidence interval, *see* CI
Confidence intervals and tests, 274
Confounding, 359, 385
Conjugate prior, 442, 468
Consistency of least squares estimators, 318
Consistent, 196
Consistent and asymptotically normal, 196
Continuity theorem, 146
Continuous data, 22
Continuous distribution, 48
Continuous random variable, 48, 52
Contour, 82
Convergence in distribution, 142
Convergence in probability, 143
Convolution, 314
Cooley, James, 527
Correction for attenuation, 331
Correlation, 20, 77
Correlation coefficient, 78, 327
Cortex, 236
Cosine regression, 346

Count data, 22
Counting process, 566
Covariance, 77
Covariance matrix, 91
Covariate, 332, 380, 591
Coverage probability, 276
Cramér-Rao lower bound, 200
Credible interval, 174, 439
Critical value of test, 271
Cross-correlation function, 554
Cross-covariance function, 553
Cross-validation, 20, 355
Cubic spline, 418
Cumulative distribution function, *see* cdf
Curve fitting, 414

D
Data analysis, 1, 23
Data augmentation, 216
Data category, 251
Decibels, 28
Decision rule, 99, 102
Decision theory, 102, 195
Decoding, 100, 426, 471
Deductive reasoning, 13
Degenerate distribution, 142
Degrees of freedom, 124, 128, 176
Delta method, 229
Density estimation, 435
Derivative, 607
Descriptive probability, 13
Design matrix, 341
Determinant, 616
Detrending, 529
Development, 368
Development of motor control, 372, 374
Deviance, 396
Digamma function, 211
Dimensionality reduction, 492
Dirac delta function, 600
Direction of maximal variation, 499
Dirichlet kernel, 546
Discrete data, 22
Discrete distribution, 48
Discrete Fourier transform, 527
Discrete random variable, 48, 52
Discrete-time stochastic process, 515
Disjoint, 38
Distribution, 47
Distribution function, 54
Distribution of a random variable, 48
Distribution of data, 24
Double-blind experiment, 387
Doubly stochastic point process, 592
Dynamic range, 547

E
Ebbinghaus, Hermann, 117
EDA (Exploratory data analysis), 17, 23, 26
EEG (Electroencephalogram), 16, 27
Efficient estimator, 201
Efron, Bradley, 180, 222
Eigenvalue, 131, 617
Eigenvector, 131, 617
Electroencephalogram, *see* EEG
Electromyogram, *see* EMG
Electrooculogram, *see* EOG
Ellipse, 82
Elliptical contours, 130
EM (expectation maximization) algorithm, 215, 468, 475, 511
EMG (Electromyogram), 34
Empirical Bayes, 464
Empirical cumulative distribution function, 64
Entropy, 95
EOG (Electrooculogram), 16
Epistemic probability, 13
EPSC (Excitatory post-synaptic current), 14, 571
Errors in variables, 359
Estimation and learning, 492
Estimator, 151, 179
Estimators, asymptotically normal, 180
Euler's equation, 622
Euler's formula, 526, 621
Euler, Leonhard, 523
Event times, 564
Events, 38
Evidence in favor of a hypothesis, 477
Excitatory post-synaptic current, *see* EPSC
Expectation, 50
Expectation maximization (EM) algorithm, *see* EM (expectation maximization) algorithm
Expected information, 199
Expected value, 50, 55
Exploratory data analysis, *see* EDA
Exponential distribution, 52, 56, 120
Exponential family, 200, 402, 443
Exponential function, 608

F
F distribution, 129
F-ratio, 336, 366

F-test for regression, 337
Factor analysis, 503
False discovery rate, *see* FDR
Family-wise error rate, 304
Fano factor, 580
Fast Fourier transform, 527
FDR (False discovery rate), 305
Feynman, Richard, 522
Filtering, 16
Filtering equation, 472
Firing rate, 563
Fisher information, 199
Fisher's z transformation, 329
Fisher, Ronald, 149, 179, 247
Fitted value, 12
fMRI (Functional magnetic resonance imaging), 1, 6, 29, 135, 303, 313, 340, 356
Forward selection, 354
Fourier analysis, 27, 522
Fourier coefficients, 524, 528
Fourier frequencies, 528
Fourier, Joseph, 523
Frequency domain, 518
Frequentist, 14, 172, 175
Frontal lobe, 108
Full conditional distributions, 468
Full rank, 616
Fully Bayes, 464
Function, 607
Function of a random variable, 62
Functional magnetic resonance imaging, *see* fMRI
Fundamental frequency, 521

G

Gabor wavelet, 426, 429
Gamma distribution, 52, 58, 123
Gamma oscillations, 518
Gauss, Karl Friedrich, 523
Gaussian distribution, 25, 116
Gaussian filter, 422, 431, 435, 436, 539, 578
Gaussian state-space model, 473
General linear model, 340, 374
Generalized cross-validation, 424
Generalized linear model, *see* GLM
Generalized maximum likelihood, 424
Generalized nonlinear model, 409
Geometric distribution, 120, 453
Gibbs phenomenon, 546
Gibbs sampling, 218, 466, 467
GLM (Generalized linear model), 392, 402, 405, 421
Glutamate, 263
Goodness-of-fit, 247, 248
Granger causality, 559
Gratification, 379
Guiding principles of science, 440, 450

H

h (hours), 52, 93, 519–521
Hand synergies, 501
Hardy-Weinberg model, 108
Harmonic frequencies, 521
Harmonic regression, 521
Hartline, Keffer, 563
Hat matrix, 415
Hazard function, 61, 121, 583
Heavy-tailed distribution, 69
Hemispatial neglect, 24
Hemodynamic response function, 314, 340
Hessian, 213
Hidden Markov model, 470
Hidden states, 439
Hierarchical model, 459
High-pass, 543
Highly significant, 253
Hippocampal place cell, 410, 472, 592
Histograms, 25
History of spiking, 564, 582
Homogeneity assumption, 106, 107
Homogeneous Poisson process, 570
Hotelling's T^2, 496
Human memory, 117
Hypothesis, 247
Hypothesis test, 248

I

ICA (Independent components analysis), 504
Ideal observer, 102
Identity, 615
i.i.d (Independent and identically distributed), 137
Imaginary number, 620
Imagined movement, 100
IMI (Inhomogeneous Markov interval), 589
Improper prior, 447
Impulse response function, 544
Increment, 566
Independence, 106
Independence assumption, 107

Independent and identically distributed, *see* i.i.d
Independent components analysis, *see* ICA
Independent events, 41
Independent random variables, 75
Indicator variable, 145
Inductive reasoning, 13, 154, 481
Inferential principle of equivalence, 441
Inferotemporal cortex, 97
Infinitesimal interval, 84
Information, 20, 94
Information in an estimator, 198
Inhomogeneous Markov interval, *see* IMI
Inhomogeneous Poisson process, 573
Inhomogeneous variances, 371
Initial values, 215
Instantaneous firing rate, 563
Integrate-and-fire neuron, 126, 127, 141, 589
Integrated likelihood, 211
Intelligence, 504
Inter-spike intervals, *see* ISIs
Interaction effects, 352, 377
Interquartile range, 26
Interspike interval distribution, 579
Inverse Gaussian distribution, 125
Inverse-gamma distribution, 468
Inverse-Wishart distribution, 468
Ion channel activation, 58
IQ, 385, 504
Irreducible, 453
ISIs (Inter-spike intervals), 127, 566
IT neural response, 167, 413

J
Jeffreys, Harold, 40, 149, 476
Joint distribution, 73
Joint pdf, 73

K
K-fold cross-validation, 355
K-means clustering, 511
Kalman filter, 473
Kalman smoother, 475
Kernel density estimate, 435
Kernel density estimator, 422, 578
Kernel function, 509
Kernel regression, 430
Kernel trick, 509
Knots, 418
Knowledge, 13
Kolmogorov-Smirnov test, 270
Kruskal-Wallis test, 384
KS statistic, 270
Kullback-Leibler (KL) divergence, 92

L
L1-penalized regression, 358, 469
L2-penalized regression, 358, 469
Lag, 531
Laplace distribution, 469
Large-sample optimality, 180
LASSO, 358
Latent factors, 503
Latent variables, 216, 399, 457
Lateral intraparietal cortex, *see* LIP
Law of cosines, 610
Law of large numbers, *see* LLN
Law of total expectation, 85
Law of total probability, 43, 86
Law of total variance, 86
LDA (Linear discriminant analysis), 506
Leakage, 546
Learning, 108
Learning a hyperparameter, 464
Learning and estimation, 492
Learning trials, 108
Least upper bound, 242
Least-squares estimates, 311
Least-squares regression, 11, 212
Leave-one-out cross-validation, 101, 356
Lebesgue integration, 59, 417
LFP (Local field potential), 1, 421, 518
Likelihood function, 155
Likelihood ratio test, 287
Limulus, 32
Lindeberg condition, 146
Linear association, 327
Linear discriminant analysis, *see* LDA
Linear discriminant function, 506
Linear filters, 539
Linear independence, 615
Linear prediction, 80
Linear regression, 89
Linear regression assumptions, 315
Linear smoother, 415
Linear trend, 323
Linearity of expectation, 74
LIP (Lateral intraparietal cortex), 86, 310
LIP neuron, 310
LLN (Law of large numbers), 137, 143
Local field potential, *see* LFP
Local fitting, 414, 429
Local polynomial regression, 432

Loess, 433
Log odds, 394
Log transformation, 28, 232
Logarithm, 28, 608
Logic, 394
Logistic distribution, 398
Logistic regression, 214, 392
Logistic regression classifier, 507
Logit transformation, 394
Loglikelihood function, 156
Long-range dependence, 516
Long-run frequency, 172
Loss function, 102
Low-pass, 542

M

Machine learning, 492
Magnetoencephalography, *see* MEG
Mallow's C_p, 354
Manifold learning, 503
Mann-Whitney test, 384
MANOVA (Multivariate analysis of variance), 491, 493
MAP estimate, 440
Marginal distribution, 73
Marginal intensity, 584
Marginal pdf, 73
Markov chain, 453
Markov chain Monte Carlo, *see* MCMC
Markov's inequality, 144
Matrix, 614
Maximum entropy, 120, 147
Maximum likelihood, *see* ML
Maximum likelihood estimator, *see* MLE
Maximum *a posteriori* estimate, 440
MCMC (Markov chain Monte Carlo), 452
Mean, 25, 50
Mean integrated squared error, 188
Mean squared error, 80, 180, 181
Mean squared error, minimum, 80
Mean vector, 90
Median, 25
MEG (Magnetoencephalography), 1, 5, 100, 358, 518, 550
Membrane conductance, 109
Memory, 151
Memoryless, 120
Method of moments, 153
Methylphenidate, 451
Metropolis-Hastings algorithm, 454, 455
Milner, Brenda, 1
Minimal signaling unit, 141
Missing data, 218
Mixture model, 216, 510, 511
Mixture of Gaussians, 216, 511
ML (Maximum likelihood), 149, 154
MLE (Maximum likelihood estimator), 152, 155
Mode, 25
Model comparison, 353
Model selection, 353
Models, scientific and statistical, 17
Modern regression, 310, 391, 413
Monte Carlo, 452
Morlet wavelet, 429
Mosteller, Frederick, 356
Motor cortical neuron, 46, 348, 406
Multimodal, 24
Multinomial distribution, 119
Multinomial logistic regression, 507
Multiple regression, 310
Multiple factors, 371
Multiple hypotheses, 302
Multiple regression, 332
Multiple testing problem, 287
Multiplication rule, 40
Multi-taper estimation, 548
Multivariate analysis, 491
Multivariate analysis of variance, *see* MANOVA
Multivariate central limit theorem, 148
Multivariate data analysis, 130
Multivariate normal distribution, 129, 130
Mutual information, 20, 92
Mutual information versus correlation, 94
Mutually exclusive, 38

N

Näive Bayes classifier, 507
Nagelkerke R^2, 397
Natural log, 31
Natural parameter, 404
Natural splines, 420
Neural network model, 508
Neuromuscular junction, 110, 113
Newton's method, 404
Neyman, Jerzy, 179, 248
Neyman-Pearson lemma, 294
NMDA antagonist, 108
NMDA receptor, 406
Noise, 10, 13
Noise variability, 317
Nominal criterion, 304
Non-informative prior, 447

Non-nested models, 295
Non-significant test, 283
Non-stationary series, 516
Nonlinear least squares, 406
Nonlinear regression, 16, 405
Nonparametric, 220
Nonparametric bootstrap, 237, 241
Nonparametric methods, 381
Nonparametric regression, 16, 413
Nonparametric statistical model, 14
Nonsingular, 616
Normal approximation to binomial, 118
Normal approximation to Poisson, 118
Normal distribution, 25, 52, 116, 118
Normal equations, 342
Normal hierarchical model, 460
Nuisance parameter, 292
Null deviance, 396
Null hypothesis, 247, 256
Nyquist frequency, 545

O

Observable variables, 457
Observational studies, 385
Observed information, 199, 203, 449
Observed information matrix, 213
Odds, 394
Olfactory bulb, 597
One-sided test, 285
Optimal decision rule, 102
Optimal integration of sensory information, 192
Optimality of MLE, 202
Optimality, large-sample, 180
Orbitofrontal, 236
Orderly, 583
Orthogonal matrix, 617
Orthogonal projection, 337, 613, 619
Orthogonalize, 420
Oscillatory, 27
Outcomes, 38
Outliers, 25

P

P-P plot, 65
p-value, 247, 281, 283
p-value calibration, 482
p-value fallacy, 281, 475
p-value, uniform distribution of, 273
p-values, combining, 301
PACF (Partial autocorrelation function), 531
Parameter, 14
Parameterization, 215, 411
Parametric bootstrap, 237, 239
Parametric regression, 16
Parametric statistical model, 14
Parkinson's disease, 569
Partial autocorrelation function, *see* PACF
Patch-clamp methods, 58
PCA (Principal component analysis), 503
pdf (Probability density function), 48, 52
Pearson correlation, 78
Pearson, Egon, 248
Pearson, Karl, 78
Penalized least squares, 357
Penalized regression, 357, 469
Percentile, 56
Percentile, sample and theoretical, 67
Perception of light, 112
Perceptron, 508
Perceptron learning rule, 508
Peri-stimulus time histogram, *see* PSTH
Periodogram, 525
Permutation test, 298, 385
PET (Positron emission tomography), 1
pH, 28
Place cells, 410
Place field, 472
Point process, 22, 564, 566
Point process regression, 392, 562
Poisson approximation to binomial, 114
Poisson distribution, 110, 115
Poisson process, 122, 570
Poisson regression, 400
Poisson regression splines, 576
Poisson spike counts, 184
Polynomial regression, 346
Polytomous regression, 507
Pooled sample variance matrix, 494, 495
Population, 49
Population mean, 51
Positive definite, 82, 91, 131, 617
Positive predictive value, 44, 282
Positive semi-definite, 91, 617
Positron emission tomography, *see* PET
Post hoc selection, 478
Posterior distribution, 173, 440
Posterior normality, 448
Posterior odds, 476
Posterior probability, 99
Posterior simulation, 451
Power basis, 420
Power law, 31, 32
Power of a test, 276

Power transfer function, 542
Pre-whitening, 322, 504, 557
Precision matrix, 133
Prediction, 89
Prediction equation, 472
Prevalence, 44, 282
Principal component, 499
Principal component analysis, *see* PCA
Principle of equivalence, inferential, 441
Prior distribution, 173
Probabilistic, 38
Probability density function, *see* pdf
Probability distribution, 24, 47, 48, 105
Probability integral transform, 63
Probability mass function, 48
Probability, descriptive, 441
Probability, epistemic, 441
Probability-probability plot, 65
Probit regression, 398
Procedural memory, 103
Propagation of uncertainty, 221
Proportional effects, 30
Proportionality symbol, 156, 232
Proposal pdf, 454
Prostatic acid phosphatase, *see* PSA
Protected least-significant difference, 373
Protocol, 387
PSA (Prostatic acid phosphatase), 44
Pseudo-data, 224, 239, 267, 552
Pseudo-random numbers, 59
PSTH (Peri-stimulus time histogram), 3, 188

Q

Q–Q plot, 65
Quadratic discriminant function, 506
Quadratic regression, 350
Quantal response, 113
Quantile, 56
Quantile, sample and theoretical, 67
Quantile-quantile plot, 65
Quartile, 24, 26

R

R^2, 316
R^2 generalization, 397
Random number, 59
Random sample, 137
Random sequences, 137
Random variable, 46, 48
Random vector, 71
Random walk, 126, 474, 530
Randomization, 385
Rank-sum test, 384
Rare events, 111
Raster plots, 3
Rayleigh test, 268
Reaction time, 24
Real world, 18
Recovery function, 591
Recurrent, 453
Refractory effects, 566, 568
Regression, 85, 87
Regression splines, 421
Regression with time series errors, 346, 535
Regress toward the mean, 88
Regularity and variability, 9
Reject a null hypothesis, 271
Relative frequencies, 49
REML estimator, 464
Renewal process, 579, 580
Repeated measures, 371
Resampling, 242
Residual, 12
Residual analysis, 319
Residual deviance, 396
Residual mean squared error, 316
Residual sum of squares, 316
Resting state, 5, 135
Restricted maximum likelihood, 424, 464
Resultant vector, 269
Retinal ganglion cell, 296
Reward, 236, 310
Ridge regression, 358
Risk, 195
Ritalin, 451
ROC curve, 278
Rotation matrix, 617
Roughness penalty, 424

S

s (seconds), 7, 8, 24, 26, 111, 112, 164, 188, 189, 280, 314, 324, 333, 421, 451, 458, 516, 551, 563
Sample mean vector, 90
Sample ACF (Sample autocorrelation function), 517
Sample autocorrelation function, *see* Sample ACF
Sample autocovariance function, 517
Sample correlation, 78
Sample covariance, 78
Sample mean, 51, 137
Sample Pearson correlation, 78

Sample percentile, 67
Sample quantile, 67
Samples, 51
Sample space, 38
Sample standard deviation, 51
Sample variance, 183
Sample variance matrix, 91
Sampling with replacement, 242, 300
Scalar, 607
Scatterplots, 26
Scheffé test, 373
Scientific models, 17
Scientific progress, 480
SEF (Supplementary eye field), 3, 187, 331, 400, 401
Selectivity index, 236, 331, 459, 462
Sensitivity, 43, 282
Separating hyperplane, 508
Sequential Bayesian estimation, 472
Set shifting, 481
Shannon, Claude, 95
Short-range dependence, 516
Shrinkage, 357, 445, 462
Sigmoidal curve, 213
Signal, 10, 13
Signal detection theory, 279
Signal variability, 317
Signal-to-noise ratio, 317
Significance level, 271
Significance test, 248
Simple linear regression, 310
Simulated data, 224, 267
Simulation sample size, 229
Simulation-based propagation of uncertainty, 225
Skewness, 24
Slutsky's theorem, 163
Smoothing, 188, 414
Smoothing splines, 423
Source localization, 306, 358
Specificity, 43
Spectral analysis, 521
Spectral decomposition, 131, 617
Spectral density function, 535
Spectrogram, 27, 514, 549
Spike, 3, 563
Spike count correlation, 141
Spike sorting, 21, 71, 502
Spike train, 3, 564
Splines, 418
Square-root of n law, 139
Square-root transformation, 234
Squared-error loss, 195
Standard deviation, 26, 50, 51, 55
Standard error, 158, 159
Standard error of the mean, 162
Standard normal, 117
Standardized, 117
Standardized residuals, 319
Starting values, 411
State, 452
State-space model, 470, 593
Stationarity, 6, 147
Stationary increments, 570
Statistic, 137
Statistical model, 10, 13, 17
  nonparametric, 14
  parametric, 14
Statistical paradigm, 2, 9
Statistical procedures, 19
Statistical reasoning, 13
Statistical thinking, 2
Statistically significant, 253
Steady-state, 5
Stepwise regression, 354
Stimulus-response, 7, 32
Stochastic, 38
Stochastic process, 564, 567
Strictly stationary, 515, 553
Student's t, 129
Studentization, 319
Sufficient statistic, 200, 404
Sum of squares due to regression, 316
Sum of squares for error, 316
Superposition, 581
Supervised learning, 510
Supplementary eye field, *see* SEF
Support vector machine, *see* SVM
Supremum, 242, 270
Surrogate data, 553
SVM (Support vector machine), 508
Symmetric, 25, 615
Synaptic transmission, 113
Synchrony, 284

T

t distribution, 128, 176
t-ratio, 325
t-test, 258, 263, 265
Tapering, 547
Taylor series, 608
Temporal coding, 97
Test data, 355
Test of independence, 254
Test-enhanced learning, 167

Tests and confidence intervals, 274
Tetrode, 71
Theobromine, 363
Theoretical distribution, 51
Theoretical world, 19
Theta rhythm, 593
Time domain, 518
Time series, 22, 315, 514
Time-frequency analysis, 548
Time-rescaling theorem, 595, 603
Time-varying firing rates, 496
Total sum of squares, 316
Training data, 355, 506
Transfer function, 542
Transformations, 33, 69
Transition probability, 453
Transpose, 615
Treatment effect, 369
Trial, 3, 4
Trigonometric polynomial, 524
Tukey, John, 23, 247, 356, 527
Two-by-two table, 255
Two-sample *t*-test, 265
Two-way ANOVA, 361
Type one error, 248, 276, 283
Type two error, 248, 276

U
Unbiased estimator, 51, 183
Uncertain inference, 14
Uncertainty, 13, 94
Unequal variance *t*-test, 266
Uniform distribution, 52
Uniformity test, 268
Unimodal, 24
Unit information prior, 481
Unsupervised learning, 510
Utility function, 102

V
Variability, 26
Variance matrix, 90, 129, 130, 133
Variance of a sum of independent random variables, 76
Variance of a sum of random variables, 77
Variance-stabilizing transformation, 34, 232
Variation in data, 14
Vascular dementia, 44
Vector, 606
Vector space, 619
Visual attention, 150

W
Wavelets, 428
Weak law of large numbers, 143
Weakly stationary, 515, 553
Weber–Fechner law, 32
Weighted least squares, 345, 535
Weighted mean, 190, 209
Weighted overlapping segment averaging, 540
Weirstrass approximation theorem, 416
Weirstrass, Karl, 523
Welch's method of spectral density estimation, 540
Welch's *t*-test, 266
White noise, 38
Whittle likelihood, 540
Wilcoxon rank-sum test, 384
Wilks' lambda, 494
Working memory, 333

Z
Z_{ons}, 261
z-score, 118, 261
z-test, 259, 261