Wolfgang Karl Härdle · Sigbert Klinke
Bernd Rönz

Introduction to Statistics

Using Interactive MM*Stat Elements

Springer

Introduction to Statistics

Wolfgang Karl Härdle • Sigbert Klinke •
Bernd Rönz

Introduction to Statistics

Using Interactive MM*Stat Elements

Springer

Wolfgang Karl Härdle
C.A.S.E. Centre f. Appl. Stat. & Econ.
School of Business and Economics
Humboldt-Universität zu Berlin
Berlin, Germany

Sigbert Klinke
Ladislaus von Bortkiewicz Chair of Statistics
Humboldt-Universität zu Berlin
Berlin, Germany

Bernd Rönz
Department of Economics Inst. for Statsitics and Econometrics
Humboldt-Universität zu Berlin
Berlin, Germany

The quantlet codes in Matlab or R may be downloaded from http://www.quantlet.de or via a link on http://springer.com/978-3-319-17703-8

ISBN 978-3-319-17703-8 ISBN 978-3-319-17704-5 (eBook)
DOI 10.1007/978-3-319-17704-5

Library of Congress Control Number: 2015958919

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

Statistics is a science central to many disciplines: modern, big, and smart data analysis can only be performed with statistical scientific tools. This is the reason why statistics is fundamental and is taught in many curricula and used in many applications. The collection and analysis of data changes the way how we observe and understand real data. Nowadays, we are collecting more and more, mostly less structured, data, which require a new analysis method and challenge the classical ones. But even nowadays the ideas used for the development of the classical methods are the foundation to new and future methods.

At the Ladislaus von Bortkiewicz Chair of Statistics, School of Business and Economics in Humboldt-Universität zu Berlin, we are introducing students into this important science with the lectures "Statistics I & II." The structure of these lectures and the methods used have changed over time, especially with the rise of the internet, but the topics taught are still the same.

In the end of the last millennium, we developed a set of hyperlinked web pages on CD, which covered even more than our lectures "Statistics I & II" in English, Spanish, French, Arabic, Portuguese, German, Indonesian, Italian, Polish, and Czech. This gave the students an easy access to data and methods.

An integral and important part of the CD were the interactive examples where the students can learn certain statistical facts by themselves. With wiki, we made a first version in German available in the internet (without interactive examples). But modern web technology nowadays allows much easier, better, and faster development of interactive examples than 15 years ago, which lead to this SmartBook with web-based interactive examples.

> Dicebat Bernardus Carnotensis nos esse quasi nanos gigantum umeris insidentes, ut possimus plura eis et remotiora videre, non utique proprii visus acumine, aut eminentia corporis, sed quia in altum subvehimur et extollimur magnitudine gigantea.
>
> Bernard of Chartres used to compare us to [puny] dwarfs perched on the shoulders of giants. He pointed out that we see more and farther than our predecessors, not because we have keener vision or greater height, but because we are lifted up and borne aloft on their gigantic stature.
>
> Johannes von Salisbury: Metalogicon 3,4,46–50

Therefore, we would like to thank all the colleagues and students who contributed to the development of the current book and its predecessors:

- For the CDs: Gökhan Aydinli, Michal Benko, Oliver Blaskowitz, Thierry Brunelle, Pavel Čížek, Michel Delecroix, Matthias Fengler, Eva Gelnarová, Zděnek Hlávka, Kateřina Houdková, Paweł Jaskólski, Šárka Jakoubková, Jakob Jurdziak, Petr Klášterecký, Torsten Kleinow, Thomas Kühn, Salim Lardjane, Heiko Lehmann, Marlene Müller, Rémy Slama, Hizir Sofyan, Claudia Trentini, Axel Werwatz, Rodrigo Witzel, Adonis Yatchew, and Uwe Ziegenhagen.
- For the Arabic and German wikis: Taleb Ahmad, Paul Giradet, Leonie Schlittgen, Dennis Uieß, and Beate Weidenhammer.
- For the current SmartBook: Sarah Asmah, Navina Gross, Lisa Grossmann, Karl-Friedrich Israel, Wiktor Olszowy, Korbinian Oswald, Darya Shesternya, and Yordan Skenderski.

Berlin, Germany — Wolfgang Karl Härdle
Berlin, Germany — Sigbert Klinke
Berlin, Germany — Bernd Rönz

Structure of the Book

Each chapter covers a broader statistical topic and each topic is categorized in sections, additionally, you may find larger explained, enhanced, and interactive examples:

Explained examples are directly related to the content of the section or chapter.
Enhanced examples may require knowledge from other earlier chapters to understand them.
Interactive examples allow to use different datasets, to choose between analysis methods and/or to play with the parameters of the (chosen) analysis method. The web address of a specific interactive example can be found in the appropriate section.

In the online version of the book under http://www.springer.com/de/book/9783319177038, you can also find at the end of each chapter a set of multiple choice exercises.

Contents

1 Basics 1
- 1.1 Objectives of Statistics 1
  - A Definition of Statistics 1
  - Explained: Descriptive and Inductive Statistics 3
- 1.2 Statistical Investigation 4
  - Conducting a Statistical Investigation 4
  - Sources of Economic Data 4
  - Explained: Public Sources of Data 6
  - More Information: Statistical Processes 6
- 1.3 Statistical Element and Population 8
  - Statistical Elements 8
  - Population 8
  - Explained: Statistical Elements and Population 8
- 1.4 Statistical Variable 10
- 1.5 Measurement Scales 11
- 1.6 Qualitative Variables 11
  - Nominal Scale 11
  - Ordinal Scale 12
- 1.7 Quantitative Variables 13
  - Interval Scale 13
  - Ratio Scale 13
  - Absolute Scale 13
  - Discrete Variable 14
  - Continuous Variable 14
- 1.8 Grouping Continuous Data 14
  - Explained: Grouping of Data 16
- 1.9 Statistical Sequences and Frequencies 16
  - Statistical Sequence 16
  - Frequency 17
  - Explained: Absolute and Relative Frequency 18

2 One-Dimensional Frequency Distributions 21
2.1 One-Dimensional Distribution 21
2.1.1 Frequency Distributions for Discrete Data 21
Frequency Table 21
2.1.2 Graphical Presentation 22
Explained: Job Proportions in Germany 25
Enhanced: Evolution of Household Sizes 25
2.2 Frequency Distribution for Continuous Data 26
Frequency Table 27
Graphical Presentation 27
Explained: Petrol Consumption of Cars 30
Explained: Net Income of German Nationals 31
2.3 Empirical Distribution Function 34
2.3.1 Empirical Distribution Function for Discrete Data 35
2.3.2 Empirical Distribution Function for Grouped Continuous Data 36
Explained: Petrol Consumption of Cars 37
Explained: Grades in Statistics Examination 38
2.4 Numerical Description of One-Dimensional Frequency Distributions 40
Measures of Location 40
Explained: Average Prices of Cars 47
Interactive: Dotplot with Location Parameters 49
Interactive: Simple Histogram 49
2.5 Location Parameters: Mean Values—Harmonic Mean, Geometric Mean 50
Harmonic Average 50
Geometric Average 52
2.6 Measures of Scale or Variation 55
Range 56
Interquartile Range 57
Mean Absolute Deviation 57
The Variance and the Standard Deviation 58
Explained: Variations of Pizza Prices 60
Enhanced: Parameters of Scale for Cars 61
Interactive: Dotplot with Scale Parameters 62
2.7 Graphical Display of the Location and Scale Parameters 64
Boxplot (Box-Whisker-Plot) 64
Explained: Boxplot of Car Prices 66
Interactive: Visualization of One-Dimensional Distributions 67

3 Probability Theory 69
3.1 The Sample Space, Events, and Probabilities 69
Venn Diagram 70

3.2 Event Relations and Operations 70
Subsets and Complements 70
Union of Sets 71
Intersection of Sets 71
Logical Difference of Sets or Events 73
Disjoint Decomposition of the Sample Space 74
Some Set Theoretic Laws 75
3.3 Probability Concepts 75
Classical Probability 76
Statistical Probability 76
Axiomatic Foundation of Probability 78
Addition Rule of Probability 78
More Information: Derivation of the Addition Rule 79
More Information: Implications of the Probability Axioms 80
Explained: A Deck of Cards 81
3.4 Conditional Probability and Independent Events 82
Conditional Probability 82
Multiplication Rule 83
Independent Events 83
Two-Way Cross-Tabulation 84
More Information: Derivation of Rules for Independent Events 85
Explained: Two-Way Cross-Tabulation 85
Explained: Screws 86
3.5 Theorem of Total Probabilities and Bayes' Rule 87
Theorem of Total Probabilities 87
Bayes' Rule 88
Explained: The Wine Cellar 88
Enhanced: Virus Test 90
Interactive: Monty Hall Problem 91
Interactive: Die Rolling Sisters 94

4 Combinatorics 97
4.1 Introduction 97
Different Ways of Grouping and Ordering 97
Use of Combinatorial Theory 98
4.2 Permutation 98
Permutations Without Repetition 98
Permutations with Repetition 99
Permutations with More Groups of Identical Elements 99
Explained: Beauty Competition 100
4.3 Variations 100
Variations with Repetition 100
Variations Without Repetition 101
Explained: Lock Picking 101

4.4 Combinations 102
Combinations Without Repetition 102
Combinations with Repetition 103
Explained: German Lotto 103
4.5 Properties of Euler's Numbers (Combination Numbers) 104
Symmetry 104
Specific Cases 104
Sum of Two Euler's Numbers 104
Euler's Numbers and Binomial Coefficients 105

5 Random Variables 107
5.1 The Definition 107
More Information 107
Explained: The Experiment 108
Enhanced: Household Size I 108
5.2 One-Dimensional Discrete Random Variables 109
Discrete Random Variable 109
Explained: One-Dimensional Discrete Random Variable 110
Enhanced: Household Size II 111
5.3 One-Dimensional Continuous Random Variables 113
Density Function 113
Distribution Function 113
More Information: Continuous Random Variable, Density, and Distribution Function 114
Explained: Continuous Random Variable 116
Enhanced: Waiting Times of Supermarket Costumers 116
5.4 Parameters 119
Expected Value 120
Variance 121
Standard Deviation 121
Standardization 122
Chebyshev's Inequality 122
Explained: Continuous Random Variable 123
Explained: Traffic Accidents 124
5.5 Two-Dimensional Random Variables 124
Marginal Distribution 125
The Conditional Marginal Distribution Function 126
Explained: Two-Dimensional Random Variable 127
Enhanced: Link Between Circulatory Diseases and Patient Age 129
5.6 Independence 131
Conditional Distribution 132
More Information 133
Explained: Stochastic Independence 134
Enhanced: Economic Conditions in Germany 136

5.7 Parameters of Two-Dimensional Distributions 139
Covariance 140
Correlation Coefficient 140
More Information 141
Explained: Parameters of Two-Dimensional Random Variables 144
Enhanced: Investment Funds 146

6 Probability Distributions 149
6.1 Important Distribution Models 149
6.2 Uniform Distribution 149
Discrete Uniform Distribution 149
Continuous Uniform Distribution 150
More Information 151
Explained: Uniform Distribution 152
6.3 Binomial Distribution 154
More Information 155
Explained: Drawing Balls from an Urn 157
Enhanced: Better Chances for Fried Hamburgers 158
Enhanced: Student Jobs 160
Interactive: Binomial Distribution 162
6.4 Hypergeometric Distribution 163
More Information 164
Explained: Choosing Test Questions 166
Enhanced: Selling Life Insurances 167
Enhanced: Insurance Contract Renewal 168
Interactive: Hypergeometric Distribution 169
6.5 Poisson Distribution 170
More Information 171
Explained: Risk of Vaccination Damage 172
Enhanced: Number of Customers in Service Department 173
Interactive: Poisson Distribution 175
6.6 Exponential Distribution 176
More Information 177
Explained: Number of Defects 178
Enhanced: Equipment Failures 180
Interactive: Exponential Distribution 181
6.7 Normal Distribution 181
Standardized Random Variable 183
Standard Normal Distribution 183
Confidence Interval 184
More Information 186
Other Properties of the Normal Distribution 187
Standard Normal Distribution 188

Explained: Normal Distributed Random Variable 188
Interactive: Normal Distribution 195
6.8 Central Limit Theorem 196
Central Limit Theorem 197
More Information 197
Explained: Application to a Uniform Random Variable 197
6.9 Approximation of Distributions 199
Normal Distribution as Limit of Other Distributions 199
Explained: Wrong Tax Returns 201
Enhanced: Storm Damage 203
6.10 Chi-Square Distribution 204
More Information 205
6.11 t-Distribution (Student t-Distribution) 206
More Information 207
6.12 F-Distribution 207
More Information 208

7 Sampling Theory 209
7.1 Basic Ideas 209
Population 209
Sample 210
Statistic 211
More Information 213
Explained: Illustrating the basic Principles of Sampling Theory 213
7.2 Sampling Distribution of the Mean 218
Distribution of the Sample Mean 218
More Information 221
Explained: Sampling Distribution 225
Enhanced: Gross Hourly Earnings of a Worker 228
7.3 Distribution of the Sample Proportion 233
Explained: Distribution of the Sample Proportion 237
Enhanced: Drawing Balls from a Urn 239
7.4 Distribution of the Sample Variance 242
Distribution of the Sample Variance S^2 243
Probability Statements About S^2 243
More Information 244
Explained: Distribution of the Sample Variance 247

8 Estimation 251
8.1 Estimation Theory 251
Point Estimation 251
The Estimator or Estimating Function 251
Explained: Basic Examples of Estimation Procedures 252
8.2 Properties of Estimators 253
Mean Squared Error 255

Unbiasedness 255
Asymptotic Unbiasedness 256
Efficiency 256
consistency 257
More Information 257
Explained: Properties of Estimators 262
Enhanced: Properties of Estimation Functions 263
8.3 Construction of Estimators 264
Maximum Likelihood 264
Least Squares Estimation 266
More Information 266
Applications of ML 266
Application of Least Squares 270
Explained: ML Estimation of an Exponential Distribution 271
Explained: ML Estimation of a Poisson Distribution 272
8.4 Interval Estimation 273
8.5 Confidence Interval for the Mean 275
Confidence Interval for the Mean with Known Variance 276
Confidence Interval for the Mean with Unknown Variance 278
Explained: Confidence Intervals for the Average Household Net Income 280
Enhanced: Confidence Intervals for the Lifetime of a Bulb 285
Interactive: Confidence Intervals for the Mean 287
8.6 Confidence Interval for Proportion 288
Properties of Confidence Intervals 290
Explained: Confidence Intervals for the Percentage of Votes 291
Interactive: Confidence Intervals for the Proportion 291
8.7 Confidence Interval for the Variance 292
Properties of the Confidence Interval 293
Explained: Confidence Intervals for the Variance of Household Net Income 294
Interactive: Confidence Intervals for the Variance 295
8.8 Confidence Interval for the Difference of Two Means 295
1. Case: The Variances σ_1^2 and σ_2^2 of the Two Populations Are Known 297
Properties of the Confidence Interval 297
2. Case: The Variances σ_1^2 and σ_2^2 of the Two Populations Are Unknown 298
Properties of Confidence Intervals When Variances Are Unknown 299

Explained: Confidence Interval for the Difference of Car Gas Consumptions 300
Enhanced: Confidence Intervals of the Difference of Two Mean Stock Prices 301
Interactive: Confidence Intervals for the Difference of Two Means 304
8.9 Confidence Interval Length 305
(a) Confidence Interval for μ 306
(b) Confidence Interval for π 306
Explained: Finding a Required Sample Size 307
Enhanced: Finding the Sample Size for an Election Threshold 308
Interactive: Confidence Interval Length for the Mean 309

9 Statistical Tests 311
9.1 Key Concepts 311
Formulating the Hypothesis 313
Test Statistic 314
Decision Regions and Significance Level 314
Non-rejection Region of Null Hypothesis 315
Rejection Region of Null Hypothesis 315
Power of a Test 323
OC-Curve 324
A Decision-Theoretical View on Statistical Hypothesis Testing 324
More Information: Examples 325
More Information: Hypothesis Testing Using Statistical Software 327
9.2 Testing Normal Means 330
Hypotheses 331
Test Statistic, Its Distribution, and Derived Decision Regions 332
Calculating the Test Statistic from an Observed Sample 336
Test Decision and Interpretation 337
Power 338
More Information: Conducting a Statistical Test 342
Explained: Testing the Population Mean 348
Enhanced: Average Life Time of Car Tires 352
Hypothesis 353
1*st* Alternative 354
2*nd* Alternative 355
3*rd* Alternative 357
Interactive: Testing the Population Mean 358
Interactive: Testing the Population Mean with Type I and II Error 359

9.3 Testing the Proportion in a Binary Population 360
Hypotheses 361
Test Statistic and Its Distribution: Decision Regions 361
Sampling and Computing the Test Statistic 363
Test Decision and Interpretation 363
Power Curve $P(\pi)$ 364
Explained: Testing a Population Proportion 364
Enhanced: Proportion of Credits with Repayment Problems 369
Interactive: Testing a Proportion in a Binary Population 376
9.4 Testing the Difference of Two Population Means 377
Hypotheses 377
Test Statistic and Its Distribution: Decision Regions 378
Sampling and Computing the Test Statistic 380
Test Decision and Interpretation 380
Explained: Testing the Difference of Two Population Means 381
Enhanced: Average Age Difference of Female and Male Bank Employees 383
1*st* Dispute 384
2*nd* Dispute 386
3*rd* Dispute 387
Interactive: Testing the Difference of Two Population Means 388
9.5 Chi-Square Goodness-of-Fit Test 389
Hypothesis 390
How Is p_j Computed? 391
Test Statistic and Its Distribution: Decision Regions 391
Approximation Conditions 392
Sampling and Computing the Test Statistic 393
Test Decision and Interpretation 394
More Information 394
Explained: Conducting a Chi-Square Goodness-of-Fit Test 397
Enhanced: Goodness-of-Fit Test for Product Demand 399
1*st* Version 400
2*nd* Version 401
9.6 Chi-Square Test of Independence 404
Hypothesis 405
Test Statistic and Its Distribution: Decision Regions 406
Sampling and Computing the Test Statistic 407
Test Decision and Interpretation 408
More Information 408

Explained: The Chi-Square Test of Independence in Action 411
Enhanced: Chi-Square Test of Independence for Economic Situation and Outlook 413

10 Two-Dimensional Frequency Distribution 419
10.1 Introduction 419
10.2 Two-Dimensional Frequency Tables 419
Realizations $m \cdot r$ 420
Absolute Frequency 420
Relative Frequency 420
Properties 420
Explained: Two-Dimensional Frequency Distribution 421
Enhanced: Department Store 422
Interactive: Example for Two-Dimensional Frequency Distribution 423
10.3 Graphical Representation of Multidimensional Data 423
Frequency Distributions 423
Scatterplots 424
Explained: Graphical Representation of a Two- or Higher Dimensional Frequency Distribution 426
Interactive: Example for the Graphical Representation of a Two- or Higher Dimensional Frequency Distribution 429
10.4 Marginal and Conditional Distributions 429
Marginal Distribution 429
Conditional Distribution 430
Explained: Conditional Distributions 432
Enhanced: Smokers and Lung Cancer 433
Enhanced: Educational Level and Age 434
10.5 Characteristics of Two-Dimensional Distributions 435
Covariance 435
More Information 437
Explained: How the Covariance Is Calculated 437
10.6 Relation Between Continuous Variables (Correlation, Correlation Coefficients) 438
Properties of the Correlation Coefficient 439
Relation of Correlation and the Scatterplot of X and Y Observations 440
Explained: Relationship of Two Metrically Scaled Variables 443
Interactive: Correlation Coefficients 444
10.7 Relation Between Discrete Variables (Rank Correlation) 445
Spearman's Rank Correlation Coefficient 445

Kendall's Rank Correlation Coefficient 447
Explained: Relationship Between Two Ordinally Scaled Variables 448
Interactive: Example for the Relationship Between Two Ordinally Scaled Variables 450
10.8 Relationship Between Nominal Variables (Contingency) 450
Explained: Relationship Between Two Nominally Scaled Variables 452
Interactive: Example for the Relationship Between Two Nominally Scaled Variables 454

11 Regression 455
11.1 Regression Analysis 455
The Objectives of Regression Analysis 455
11.2 One-Dimensional Regression Analysis 457
One-Dimensional Linear Regression Function 457
Quality (Fit) of the Regression Line 463
One-Dimensional Nonlinear Regression Function 466
Explained: One-Dimensional Linear Regression 468
Enhanced: Crime Rates in the US 471
Enhanced: Linear Regression for the Car Data 472
Interactive: Simple Linear Regression 473
11.3 Multi-Dimensional Regression Analysis 474
Multi-Dimensional Regression Analysis 474

12 Time Series Analysis 477
12.1 Time Series Analysis 477
Definition 477
Graphical Representation 477
The Objectives of Time Series Analysis 477
Components of Time Series 479
12.2 Trend of Time Series 479
Method of Moving Average 479
Least-Squares Method 481
More Information: Simple Moving Average 483
Explained: Calculation of Moving Averages 485
Interactive: Test of Different Filters for Trend Calculation 486
12.3 Periodic Fluctuations 487
Explained: Decomposition of a Seasonal Series 489
Interactive: Decomposition of Time Series 491
12.4 Quality of the Time Series Model 492
Mean Squared Dispersion (Estimated Standard Deviation) 493
Interactive: Comparison of Time Series Models 494

A Data Sets in the Interactive Examples 495
A.1 ALLBUS Data 495
A.1.1 ALLBUS1992, ALLBUS2002, and ALLBUS2012: Economics 495
A.1.2 ALLBUS1994, ALLBUS2002, and ALLBUS2012: Trust 496
A.1.3 ALLBUS2002, ALLBUS2004, and ALLBUS2012: General 497
A.2 Boston Housing Data 499
A.3 Car Data 499
A.4 Credit Data 500
A.5 Decathlon Data 501
A.6 Hair and Eye Color of Statistics Students 502
A.7 Index of Basic Rent 502
A.8 Normally Distributed Data 503
A.9 Telephone Data 503
A.10 Titanic Data 504
A.11 US Crime Data 504

Glossary 507

Chapter 1
Basics

1.1 Objectives of Statistics

A Definition of Statistics

Statistics is the science of collecting, describing, and interpreting data, i.e., the tool box underlying empirical research.

In analyzing data, scientists aim to describe our perception of the world. Descriptions of stable relationships among observable phenomena in the form of theories are sometimes referred to as being explanatory. (Though one could argue that science merely describes *how* things happen rather than *why.*) Inventing a theory is a creative process of restructuring information embedded in existing (and accepted) theories and extracting exploitable information from the real world. (We are abstracting from purely axiomatic theories derived by logical deduction.)

A first exploratory approach to groups of phenomena is typically carried out using methods of *statistical description*.

Descriptive Statistics

Descriptive statistics encompasses tools devised to organize and display data in an accessible fashion, i.e., in a way that doesn't exceed the perceptual limits of the human mind. It involves the quantification of recurring phenomena. Various summary statistics, mainly averages, are calculated; raw data and statistics are displayed using tables and graphs.

W.K. Härdle et al., *Introduction to Statistics*, DOI 10.1007/978-3-319-17704-5_1

Statistical description can offer important insights into the occurrence of isolated phenomena and indicate associations among them. But can it provide results that can be considered laws in a scientific context? Statistics is a means of dealing with variations in characteristics of distinct objects. Isolated objects are thus not representative for the population of objects possessing the quantifiable feature under investigation. Yet variability can be the result of the (controlled or random) variation of other, underlying variables. Physics, for example, is mainly concerned with the extraction and mathematical formulation of exact relationships, not leaving much room for random fluctuations. In statistics such random fluctuations are modeled Statistical relationships are thus relationships which account for a certain proportion of stochastic variability.

Inductive Statistics

In contrast to wide areas of physics, empirical relationships observed in the natural sciences, sociology, and psychology (and more eclectic subjects such as economics) are statistical. Empirical work in these fields is typically carried out on the basis of experiments or sample surveys. In either case, the entire population cannot be observed—either for practical or economic reasons. Inferring from a limited sample of objects to characteristics prevailing in the underlying population is the goal of *inferential* or *inductive statistics*. Here, variability is a reflection of variation in the sample and the sampling process.

Statistics and the Scientific Process

Depending on the stage of the scientific investigation, data are examined with varying degrees of prior information. Data can be collected to explore a phenomenon in a first approach, but it can also serve to statistically test (verify/falsify) hypotheses about the structure of the characteristic(s) under investigation.

Thus, statistics is applied at all stages of the scientific process wherever quantifiable phenomena are involved.

Here, our concept of quantifiability is sufficiently general to encompass a very broad range of scientifically interesting propositions. Take, for example, a proposition such as “a bumble bee is flying by.” By counting the number of such occurrences in various settings we are quantifying the occurrence of the phenomenon. On this basis we can try to infer the likelihood of coming across a bumble bee under specific circumstances (e.g., on a rainy summer day in Berlin).

Table 1.1 Absolute frequencies of numbers in National Lottery

1	*2*	*3*	*4*	*5*	*6*	*7*
311	337	345	316	321	335	322
8	*9*	*10*	*11*	*12*	*13*	*14*
309	324	331	315	302	276	310
15	*16*	*17*	*18*	*19*	*20*	*21*
322	319	337	331	326	312	334
22	*23*	*24*	*25*	*26*	*27*	*28*
322	319	304	325	337	323	285
29	*30*	*31*	*32*	*33*	*34*	*35*
321	311	333	378	340	291	330
36	*37*	*38*	*39*	*40*	*41*	*42*
340	320	357	326	329	335	335
43	*44*	*45*	*46*	*47*	*48*	*49*
311	314	304	327	311	337	361

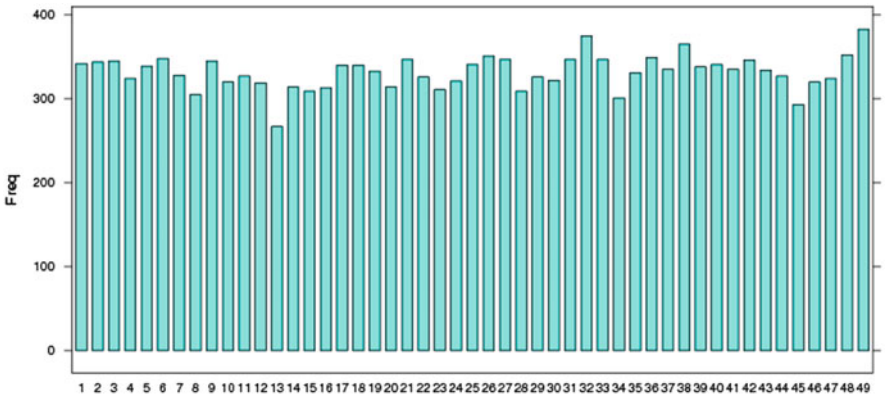


Fig. 1.1 Absolute frequencies of numbers in the National Lottery from 1955 to 2007

Explained: Descriptive and Inductive Statistics

Descriptive statistics provide the means to summarize and visualize data. Table 1.1, which contains the frequency distribution of numbers drawn in the National Lottery, provides an example of a such a summary. Cursory examination suggests that some numbers occur more frequently than others (Fig. 1.1). Does this suggest bias in the way numbers are selected? As we shall see, statistical methods can also be used to test such propositions.

1.2 Statistical Investigation

Conducting a Statistical Investigation

Statistical investigations often involve the following steps:

1. Designing the investigation: development of the objectives, translation of theoretical concepts into observable phenomena (i.e., variables), environmental setting (e.g., determining which parameters are held constant), cost projection, etc.
2. Obtaining data
   - Primary data: data collected by the institution conducting the investigation
     - Surveys:
       * recording data without exercising control over environmental conditions which could influence the observations
       * observing all members of the population (*census*) or taking a sample (*sample survey*)
       * collecting data by interview or by measurement
       * documentation of data via questionnaires, protocols, etc.
       * personal vs. indirect observation (e.g., personal interview, questionnaires by post, telephone, etc.)
     - Experiments: actively controlling variables to capture their impact on other variables
     - Automated recording: observing data as it is being generated, e.g., within a production process
   - Secondary data: using readily available data, either from internal or external sources.
3. Organizing the data
4. Analysis: applying statistical tools
5. Interpretation: which conclusions do the quantitative information generated by statistical procedures support?

Sources of Economic Data

- Public Statistics
- Private Statistics
- International Organizations

Figure 1.2 illustrates the sequence of steps in a statistical investigation.

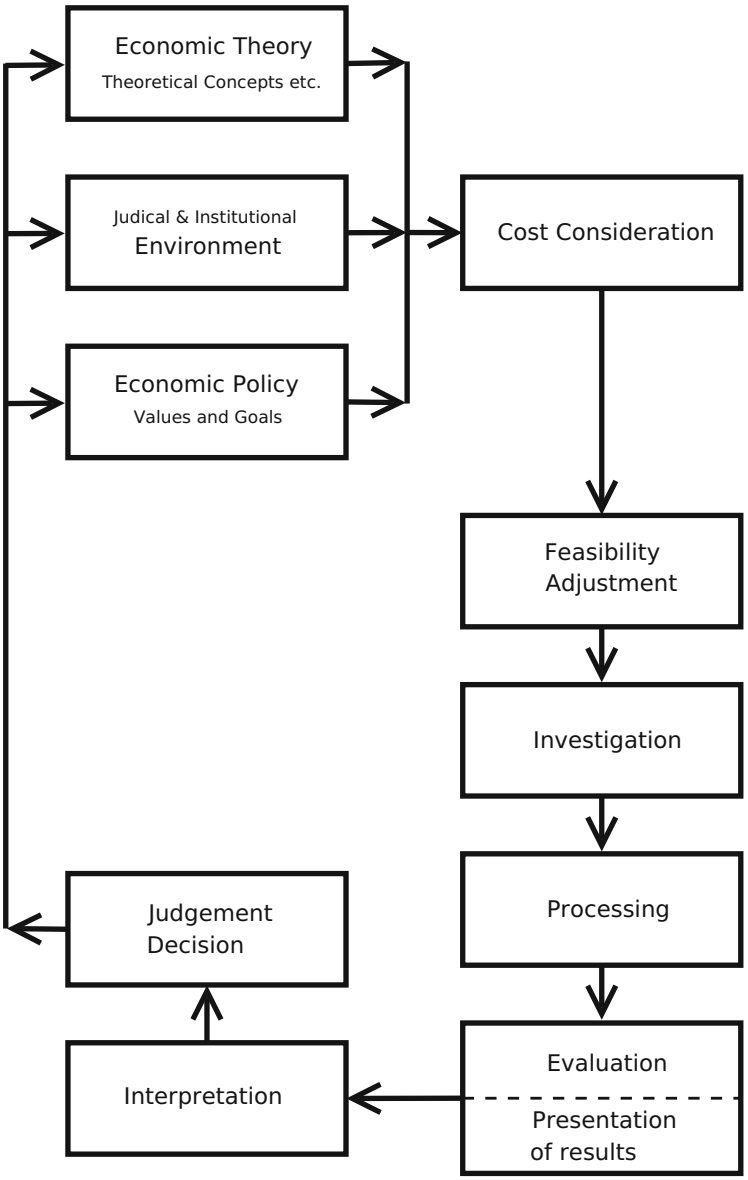


Fig. 1.2 Overview of steps in statistical investigation

Table 1.2 Data on animal populations of Berlin's three major zoos

Animals in Berlin, 2013	Zoo and Aquarium	Tierpark
Mammals		
Total population	1044	1283
Species	169	199
Birds		
Total population	2092	2380
Species	319	356
Snakes		
Total population	357	508
Species	69	103
Lizards		
Total population	639	55
Species	54	3
Fish		
Total population	7629	938
Species	562	106
Invertebrate		
Total population	8604	2086
Species	331	79
Visitors	3059136	1035899

Data from: Financial reports 2013 of Zoologischer Garten Berlin AG, Tierpark Berlin GmbH and Amt für Statistik Berlin-Brandenburg

Explained: Public Sources of Data

The official body engaged in collecting and publishing Berlin-specific data is the *Amt für Statistik Berlin-Brandenburg*. For example, statistics on such disparate subjects as the animal populations of Berlin's three major zoos and voter participation in general elections, are available (Table 1.2). The data on voter participation covers the elections into the 8th European Parliament in Berlin (25.05.2014). The map displays the **election participation** in election districts of Berlin (Fig. 1.3).

More Information: Statistical Processes

A common objective of economic policy is to reduce the overall duration of unemployment in the economy.

An important theoretical question is, to what extent can the level of unemployment benefits account for variations in unemployment duration.

In order to make this question suitable for a statistical investigation, the variables must be translated into directly observable quantities. (For example, the number

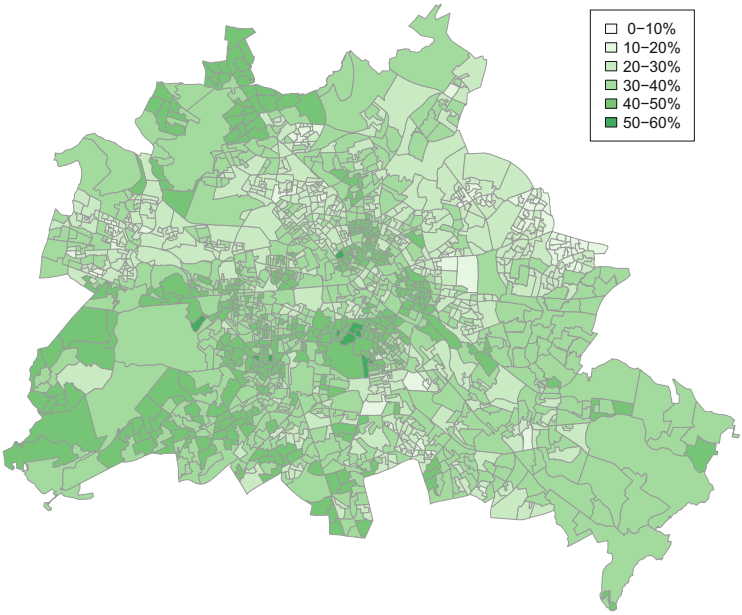


Fig. 1.3 Voter participation in Berlin (Data from: Amt für Statistik Berlin-Brandenburg 2014)

of individuals who are registered as unemployed is a quantity available from government statistics. While this may not include everyone who would like to be working, it is usually used as the unemployment variable in statistical analyses.)

By examining government unemployment benefit payments in different countries, we can try to infer whether more or less generous policies have an impact on the unemployment rate.

Prior to further investigation, the collected raw data must be organized in a fashion suitable for the statistical methods to be performed upon them.

Exploring the data for extractable information and presenting the results in an accessible fashion by means of statistical tools lies at the heart of statistical investigation.

In interpreting quantitative statistical information, keys to an answer to the initial scientific questions are sought.

Analogous to the general scientific process, conclusions reached in the course of statistical interpretation frequently give rise to further propositions—triggering the next iteration of the statistical process.

1.3 Statistical Element and Population

Statistical Elements

Objects whose attributes are observed or measured for statistical purposes are called statistical elements.

In order to identify all elements relevant to a particular investigation, one must specify their defining characteristics as well as temporal and spatial dimensions.

Example Population Census in Germany

- defining characteristic: citizen of Germany
- spatial: permanent address in Federal Republic of Germany
- temporal: date of census

Population

The universe of statistical elements covered by a particular set of specifications is called population. In general, increasing the number of criteria to be matched by the elements will result in a smaller and more homogeneous population. Populations can be finite or infinite in size.

In a census, all elements of the population are investigated. Recording information from a portion of the population yields a sample survey.

The stock of elements constituting a population may change over time, as some elements leave and others enter the population. This sensitivity of populations to time flow has to be taken into account when carrying out statistical investigations.

Explained: Statistical Elements and Population

We use the following questionnaire, developed at the Department of Statistics, to clarify the notions of statistical unit and population. This questionnaire was filled out by all participants in Statistics 1. The investigation was carried out on the first lecture of the summer semester in 1999.

The **population** consists of all students taking part in Statistics 1 at Humboldt University during the summer semester 1999. The **statistical unit** is one student.

HUMBOLDT UNIVERSITY BERLIN
Department of Statistics - Statistics 1

QUESTIONNAIRE

Welcome to Statistics. Before we start, we would like to ask a you few questions. Your answers will help us to optimize the lectures. Furthermore, your answers will be statistically analyzed during the lecture. Everybody, who fills in this questionnaire, will have a chance to win the multimedia version of Statistics 1 which is worth 200,-DM.

1. **Do you have access to the internet?**

   □ **Yes** □ **No**

   If yes:

2. **From where do you connect to the internet?**

   □ **home** □ **university**
   □ **internet cafe** □ **friends**
   □ **other (please, specify):**

3. **Which internet browser do you usually use?**

   □ **Netscape 4.5 or newer**
   □ **older version of Netscape**
   □ **Netscape, I do not know the version number**
   □ **Internet Explorer 4 or newer**
   □ **older version of Internet Explorer**
   □ **Internet Explorer, I do not know the version number**

4. **Do you have access to a multimedia computer (i.e., computer which can be used to play audio and video files?)**

   □ **Yes** □ **No**

5. **Have you previously studied Theory of Probability or Stochastics?**

   □ **Yes** □ **No**

6. **What is the probability that the sum of numbers of two dice is seven?**

7. **In which state did you attend secondary school?**

   □ **Baden-Württemberg** □ **Bayern**
   □ **Berlin** □ **Brandenburg**
   □ **Bremen** □ **Hamburg**
   □ **Hessen** □ **Mecklenburg-Vorpommern**
   □ **Niedersachsen** □ **Nordrhein-Westfalen**
   □ **Rheinland-Pfalz** □ **Saarland**
   □ **Sachsen** □ **Sachsen-Anhalt**
   □ **Schleswig-Holstein** □ **Thüringen**

Thank you. If you would like to win the multimedia CD, please complete the following entries:
Name:
ID number:

1.4 Statistical Variable

An observable characteristic of a statistical element is called statistical variable. The actual values assumed by statistical variables are called observations, measurements, or data. The set of possible values a variable can take is called sample space.

Variables are denoted by script capitals $X, Y, \ldots$, whereas corresponding realizations are written in lowercase: $x_1, x_2, \ldots$, the indices reflect the statistical elements sampled.

Variable	Observations
X	$x_1, x_2, x_3, \ldots$
Y	$y_1, y_2, y_3, \ldots$

It is useful to differentiate between variables used for identification and target variables.

Identification variables In assigning a set of fixed values the elements of the population are specified. For example, restricting a statistical investigation to female persons involves setting the identification variable "sex" to "female."

Target variables These are the characteristics of interest, the phenomena that are being explored by means of statistical techniques, e.g., the age of persons belonging to a particular population.

Example Objective of the statistical investigation is to explore Berlin's socio-economic structure as of December 21, 1995. The identification variables are chosen to be:

- legal: citizen
- spatial: permanent address in Berlin
- temporal: 31 December 1995

Statistical element: a registered citizen of Berlin on 31 December 1995
Population: all citizens of Berlin on 31 December 1995
Possible target variables:

Symbol	Variable	Sample space
X	Age (rounded to years)	$\{0, 1, 2, \ldots\}$
S	Sex	{female, male}
T	Marital status	{single, married, divorced}
Y	Monthly income	$[0, \infty)$

1.5 Measurement Scales

The values random variables take can differ distinctively, as can be seen in the above table. They can be classified into quantitative, i.e., numerically valued (age and income) and qualitative, i.e., categorical (sex, marital status) variables. As numerical values are usually assigned to observations of qualitative variables, they may appear quantitative. Yet such synthetic assignments aren't of the same quality as numerical measurements that naturally arise in observing a phenomenon. The crucial distinction between quantitative and qualitative variables lies in the properties of the actual scale of measurement, which in turn is crucial to the applicability of statistical methods. In developing new tools statisticians make assumptions about permissible measurement scales.

A measurement is a numerical assignment to an observation. Some measurements appear more natural than others. By measuring the height of persons, for example, we apply a yardstick that ensures comparability between observations up to almost any desired precision—regardless of the units (such as inches or centimeters). School grades, on the other hand, represent a relatively rough classification indicating a certain ranking, yet putting many pupils into the same category. The values assigned to qualitative statements like "very good," "average," etc. are an arbitrary yet practical shortcut in assessing people's achievements. As there is no conceptual reasoning behind a school grade scale, one should not try to interpret the "distances" between grades.

Clearly, height measurements convey more information than school marks, as distances between measurements can consistently be compared. Statements such as "Tom is twice as tall as his son" or "Manuela is 35 centimeters smaller than her partner" are permissible.

As statistical methods are developed in mathematical terms, the applicable scales are also defined in terms of mathematical concepts. These are the transformations that can be imposed on them without loss of information. The wider the range of permissible transformations, the less information the scale can convey. Table 1.3 lists common measurement scales in increasing order of information content. Scales carrying more information can always be transformed into less informative scales.

1.6 Qualitative Variables

Nominal Scale

The most primitive scale, one that is only capable of expressing whether two values are equal or not, is the nominal scale. It is purely qualitative.

If an experiment's sample space consists of categories without a natural ordering, the corresponding random variable is nominally scaled. The distinct numbers assigned to outcomes merely indicate whether any two outcomes are equal or not.

Table 1.3 Measurement scales of random variables

Variable	Measurement scale	Statements	Permissible transformations
Qualitative	Nominal scale	Equivalence	Any equivalence preserving mapping
Categorical	Ordinal scale	Equivalence, order	Any order preserving mapping
Quantitative metric	Interval scale	Equivalence, order, distance	$y = \alpha x + \beta, \alpha > 0$
	Ratio scale	Equivalence, order, distance, ratio	$y = \alpha x, \alpha > 0$
	Absolute scale	Equivalence, order, distance, ratio, absolute level	Identity function

For example, numbers assigned to different political opinions may be helpful in compiling results from questionnaires. Yet in comparing two opinions we can only relate them as being of the same kind or not. The numbers do not establish any ranking.

Variables with exactly two mutually exclusive outcomes are called binary variables or dichotomous variables. If the indicator numbers assigned convey information about the ranking of the categories, a binary variable might also be regarded as ordinally scaled.

If the categories (events) constituting the sample space are not mutually exclusive, i.e., one statistical element can correspond to more than one category, we call the variable cumulative. For example, a person might respond affirmatively to different categories of professional qualifications. But there cannot be more than one current full-time employment (by definition).

Ordinal Scale

If the numbers assigned to measurements express a natural ranking, the variable is measured on an ordinal scale.

The distances between different values cannot be interpreted—a variable measured on an ordinal scale is thus still somehow nonquantitative. For example, school marks reflect different levels of achievement. There is, however, usually no reason to regard a work receiving a grade of "4" as twice as good as one that achieved a grade of "2."

As the numbers assigned to measurements reflect their ranking relatively to each other, they are called rank values.

There are numerous examples for ordinally scaled variables in psychology, sociology, business studies, etc. Scales can be designed attempting to measure such vague concepts as "social status," "intelligence," "level of aggression," or "level of satisfaction."

1.7 Quantitative Variables

Apart from possessing a natural ordering, measurements of quantitative variables can also be interpreted in terms of distances between observations.

Interval Scale

If distances between measurements can be interpreted meaningfully, the variable is measured on an interval scale. In contrast to the ratio scale, ratios of measurements don't have a substantial meaning, for the interval scale doesn't possess a natural zero value. For example, temperatures measured in degrees centigrade can be interpreted in order of higher or lower levels. Yet, a temperature of 20 degrees centigrade cannot be regarded to be twice as high as a temperature of 10 degrees. Think of equivalent temperatures measured in Fahrenheit. Converting temperatures from centigrade to Fahrenheit and vice versa involves shifting the zero point.

Ratio Scale

Values of variables measured on a ratio scale can be interpreted both in terms of distances *and* ratios. The ratio scale thus conveys even more information than the interval scales, in which only intervals (distances between observations) are quantitatively meaningful.

The phenomena to be measured on a ratio scale possess a natural zero element, representing total lack of the attribute. Yet there isn't necessarily a natural measurement unit. Prominent examples are weight, height, age, etc.

Absolute Scale

The absolute scale is a metric scale with a natural unit of measurement. Absolute scale measurement is thus simply counting. It is the only measurement without

alternative. *Example:* All countable phenomena such as the number of people in a room or number of balls in an urn.

Discrete Variable

A metric variable that can take a finite or countably infinite set of values is called discrete. *Example:* Monthly production of cars or number of stars in the universe.

Continuous Variable

A metric variable is called continuous, if it can take on an uncountable number of values in any interval on the number line. *Example:* Petrol sold in a specific period of time.

In practice, many theoretically continuous variables are measured discretely due to limitations in the precision of physical measurement devices. Measuring a person's age can be carried out to a certain fraction of a second, but not infinitely precisely.

We regard a theoretically continuous variable, which we can measure with a certain sufficient precision, as effectively continuous. Similar reasoning applies to discrete variables, which we sometimes regard as quasi-continuous, if there are enough values to suggest the applicability of statistical methods devised for continuous variables.

1.8 Grouping Continuous Data

Consider height data on 100 school boys. In order to gain an overview of the distribution of heights you start "reading" the raw data. But the typical person will soon discover that making sense of more than, say, 10 observations without some process of simplification is not useful. Intuitively, one starts to group individuals with similar heights. By focusing on the size of these groupings rather than on the raw data itself one gains an overview of the data. Even though one has set aside detailed information about exact heights, one has created a clearer overall picture.

Data sampled from continuous or quasi-continuous random variables can be condensed by partitioning the sample space into mutually exclusive classes. Counting the number of realizations falling into each of these classes is a means of providing a descriptive summary of the data. Grouping data into classes can greatly enhance our ability to "see" the structure of the data, i.e., the distribution of the realizations over the sample space.

Table 1.4 Example for grouping of continuous variables

1^{st} alternative	
Less than 10	< 10
10 to less than 12	$\geq 10, < 12$
12 to less than 15	$\geq 12, < 15$
15 or greater	≥ 15
2^{nd} alternative	
Less than or equal to 10	≤ 10
Greater than 10 to less than or equal to 12	$> 10, \leq 12$
Greater than 12 to less than or equal to 15	$> 12, \leq 15$
Greater than 15	> 15

Classes are nonoverlapping intervals specified by their upper and lower limits (class boundaries). Loss of information arises from replacing the actual values by the sizes and location of the classes into which they fall. If one uses too few classes, then useful patterns may be concealed. Too many classes may inhibit the expositional value of grouping.

Class boundaries The upper and lower values of a class are called class boundaries. A class j is fully specified by its lower boundary x_j^l and upper boundary x_j^u $(j = 1, \ldots, k)$, where

$x_j^u = x_{j+1}^l \quad (j = 1, \ldots, k-1)$, i.e., upper boundary of the jth class and lower boundary of the $(j+1)$th class coincide.

$x_j^l < x \leq x_j^u$ or $x_j^l \leq x < x_j^u \quad (j = 1, \ldots, k)$, i.e., the class boundary can be attributed to either of the classes it separates (Table 1.4).

When measurements of (theoretically) unbounded variables are being classified, left- and/or right-most classes extend to $-\infty$, $+\infty$, respectively, i.e., they form a semi-open interval.

Class width Taking the difference between two boundaries of a class yields the class width (sometimes referred to as the class size). Classes need not be of equal width:

$$\Delta x_j = x_j^u - x_j^l \quad (j = 1, \ldots, k)$$

Class midpoint The class midpoint x_j can be interpreted as a representative value for the class, if the measurements falling into it are evenly or symmetrically distributed.

$$x_j = \frac{x_j^l + x_j^u}{2} \quad (j = 1, \ldots, k)$$

Table 1.5 Income distribution in Germany

Taxable income			Persons (1000)	Consolidated gross income (mio. marks)
1	–	4000	1445.2	2611.3
4000	–	8000	1455.5	8889.2
8000	–	12000	1240.5	12310.9
12000	–	16000	1110.7	15492.7
15000	–	25000	2762.9	57218.5
25000	–	30000	1915.1	52755.4
30000	–	50000	6923.7	270182.7
50000	–	75000	3876.9	234493.1
75000	–	100000	1239.7	105452.9
100000	–	250000	791.6	108065.7
250000	–	500000	93.7	31433.8
500000	–	1 Mio.	26.6	17893.3
1 Mio.	–	2 Mio.	8.6	11769.9
2 Mio.	–	5 Mio.	3.7	10950.8
5 Mio.	–	10 Mio.	0.9	6041.8
$\geq$ 10 Mio.			0.5	10749.8

Data from: Datenreport 1992, p. 255; Statistisches Jahrbuch der Bundesrepublik Deutschland 1993, p. 566

Explained: Grouping of Data

Politicians and political scientists are interested in the income distribution. In Germany, a large portion of the population has taxable income The 1986 data, compiled from various official sources, displays a concentration in small and medium income brackets. Relatively few individuals earned more than one million marks. Greater class widths have been chosen for higher income brackets to retain a compact exposition despite the skewness in the data (Table 1.5).

1.9 Statistical Sequences and Frequencies

Statistical Sequence

In recording data we generate a statistical sequence. The original, unprocessed sequence is called *raw data*. Given an appropriate scale level (i.e., at least an ordinal scale), we can sort the raw data, thus creating an *ordered sequence*.

Data collected at the same point in time or for the same period of time on different elements are called cross-section data. Data collected at different points in time or for different periods of time on the same element are called time series data. The sequence of observations is ordered along the time axis.

Frequency

The number of observations falling into a given class is called the frequency. Classes are constructed to summarize continuous or quasi-continuous data by means of frequencies.

In discrete data one regularly encounters so-called ties, i.e., two or more observations taking on the same value. Thus, discrete data may not require grouping in order to calculate frequencies.

Absolute Frequency

Counting the number of observations taking on a specific value yields the absolute frequency:

$$h\left(X = x_j\right) = h\left(x_j\right) = h_j$$

When data are grouped, the *absolute frequencies of classes* are calculated as follows:

$$h\left(x_j\right) = h\left(x_j^l \leq X < x_j^u\right)$$

Properties:

$$0 \leq h\left(x_j\right) \leq n$$

$$\sum_j h\left(x_j\right) = n$$

Relative Frequency

The proportion of observations taking on a specific value or falling into a specific class is called the relative frequency, the absolute frequency standardized by the total number of observations.

$$f\left(x_j\right) = \frac{h\left(x_j\right)}{n}$$

Properties:

$$0 \leq f\left(x_j\right) \leq 1$$
$$\sum_j f\left(x_j\right) = 1$$

Frequency Distribution

By standardizing class frequencies for grouped data by their respective class widths, frequencies for differently sized classes are made comparable. The resulting frequencies can be compiled to form a frequency distribution.

$$\hat{h}\left(x_j\right) = \frac{h\left(x_j\right)}{x_j^u - x_j^l}$$
$$\hat{f}\left(x_j\right) = \frac{f\left(x_j\right)}{x_j^u - x_j^l},$$

where x_j^l, x_j^u are the upper and lower class boundaries with $x_j^l < x \leq x_j^u$.

Explained: Absolute and Relative Frequency

150 persons have been asked for their marital status: 88 of them are married, 41 single, and 21 divorced.

The four conceivable responses have been assigned categories as follows:

- single: x_1
- married: x_2
- divorced: x_3
- widowed: x_4

The number of statistical elements is $n = 150$. The absolute frequencies given above are:

- $h\left(x_1\right) = 41$
- $h\left(x_2\right) = 88$
- $h\left(x_3\right) = 21$
- $h\left(x_4\right) = 0$

Dividing by the sample size $n = 150$ yields the relative frequencies:

- $f(x_1) = 41/150 = 0.27$
- $f(x_2) = 88/150 = 0.59$
- $f(x_3) = 21/150 = 0.14$
- $f(x_4) = 0/150 = 0.00$

Thus, 59 % of the persons surveyed are married, 27 % are single, and 15 % divorced. No one is widowed.

Chapter 2
One-Dimensional Frequency Distributions

2.1 One-Dimensional Distribution

The collection of information about class boundaries and relative or absolute frequencies constitutes the frequency distribution. For a single variable (e.g., height) we have a one-dimensional frequency distribution. If more than one variable is measured for each statistical unit (e.g., height and weight), we may define a two-dimensional frequency distribution. We use the notation X to denote the observed variable.

2.1.1 Frequency Distributions for Discrete Data

Suppose the variable X can take on k distinct values $x_j, j = 1, .., k$. Note that we index these distinct values or classes using the subscript j. We will denote n observations on the random variable by $x_i, i = 1, \ldots, n$. The context will usually make it clear whether we are referring to the k distinct values or the n observations. We will assume that $n > k$.

Frequency Table

For a discrete variable X, the frequency table displays the distribution of frequencies over the given categories. From now on we will speak of discrete variables to encompass categorical variables and discrete metric variables with few possible observations. Note that the sum of the frequencies across the various categories equals the number of observations, i.e., $\sum_{j=1}^{k} x_j = n$ (Table 2.1).

W.K. Härdle et al., *Introduction to Statistics*, DOI 10.1007/978-3-319-17704-5_2

Table 2.1 A frequency table

Values	Absolute frequencies	Relative frequencies
x_1	$h(x_1)$	$f(x_1)$
x_2	$h(x_2)$	$f(x_2)$
$\vdots$	$\vdots$	$\vdots$
x_j	$h(x_j)$	$f(x_j)$
$\vdots$	$\vdots$	$\vdots$
x_k	$h(x_k)$	$f(x_k)$
Total	n	1

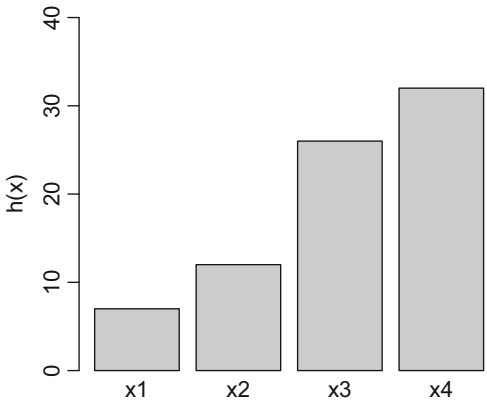


Fig. 2.1 Example of a bar graph

2.1.2 Graphical Presentation

Several graph types exist for displaying frequency distributions of discrete data.

Bar Graph

In a bar graph, frequencies are represented by the height of bars vertically drawn over the categories depicted on the horizontal axis. Since the categories do not represent intervals as in the case of grouped continuous data, the width of the bars cannot be interpreted meaningfully. Consequently, the bars are drawn with equal width (Fig. 2.1).

Stacked Bar Chart

Sometimes one wants to compare relative frequencies in different samples (different samples may arise at different points in time or from different populations). This can be done by drawing one bar graph for each sample. An alternative is the stacked bar

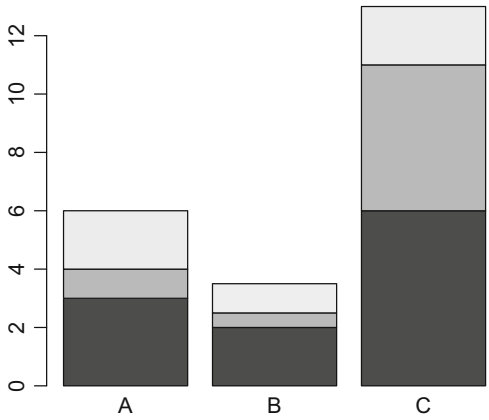


Fig. 2.2 Example of a stacked bar chart

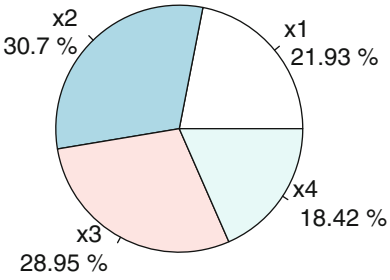


Fig. 2.3 Example of a pie chart

chart. It consists of as many segmented bars as there are samples. Each segment of a bar chart represents a relative frequency (Fig. 2.2).

Pie Chart

In pie charts, frequencies are displayed as segments of a pie. The area of each segment is proportional to the corresponding relative frequency (Fig. 2.3).

Pictograph

In a pictograph, the size or number of pictorial symbols is proportional to observed frequencies (Fig. 2.4).

Statistical Map

Different relative frequencies in different areas are visualized by different colors, shadings, or patterns (Fig. 2.5).

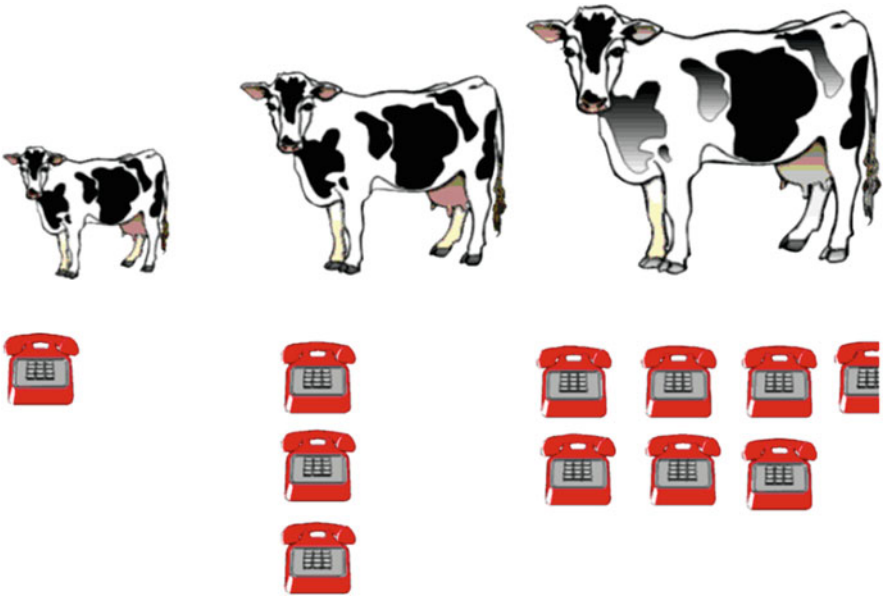


Fig. 2.4 Two examples of pictographs

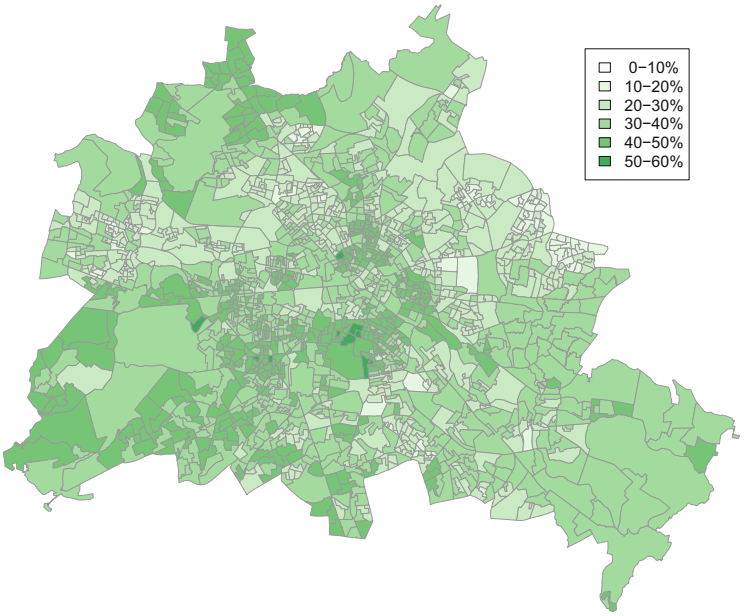


Fig. 2.5 Example of a statistical map

Table 2.2 Frequency table on employed population in Germany

j	Status x_j	$h(x_j)$ $(1000's)$	$f(x_j)$
1	Wage-earners	14568	0.389
2	Salaried	16808	0.449
3	Civil servants	2511	0.067
4	Self employed	3037	0.081
5	Family employed	522	0.014
	Total	37466	1.000

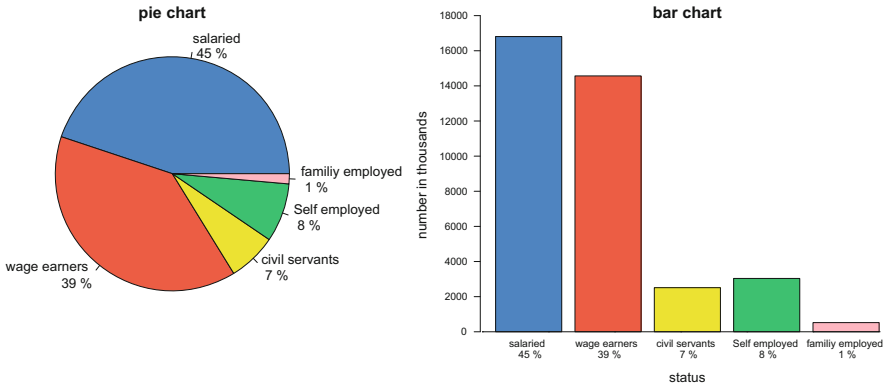


Fig. 2.6 Pie chart and bar graph on employed population in Germany

Explained: Job Proportions in Germany

In April 1991, Germany's employed population was surveyed with respect to type of employment. Table 2.2 summarizes the data. Visualizing the proportions helps us to analyze the data. In Fig. 2.6 you can clearly see the high proportion of wage-earners and salaried in contrast to the other categories.

Enhanced: Evolution of Household Sizes

The evolution of household sizes over the twentieth century can be studied using data compiled at various points in time.

Statistical elements: households

Statistical variable: size of household (metric, discrete)

Table 2.3 contains relative frequencies measured in percent for various years.

The structural shift in the pattern of household sizes towards the end of the century becomes visible if we draw bar charts for each year. The graphics in Fig. 2.7 display a clear shift towards smaller families during the twentieth century.

Table 2.3 Frequency table on the evolution of household sizes over the twentieth century

Household size X	1900	1925	1950	1990
1	7.1	6.7	19.4	35.0
2	14.7	17.7	25.3	30.2
3	17.0	22.5	23.0	16.7
4	16.8	19.7	16.2	12.8
≥ 5	44.4	33.3	16.1	5.3
Total	100	100	100	100

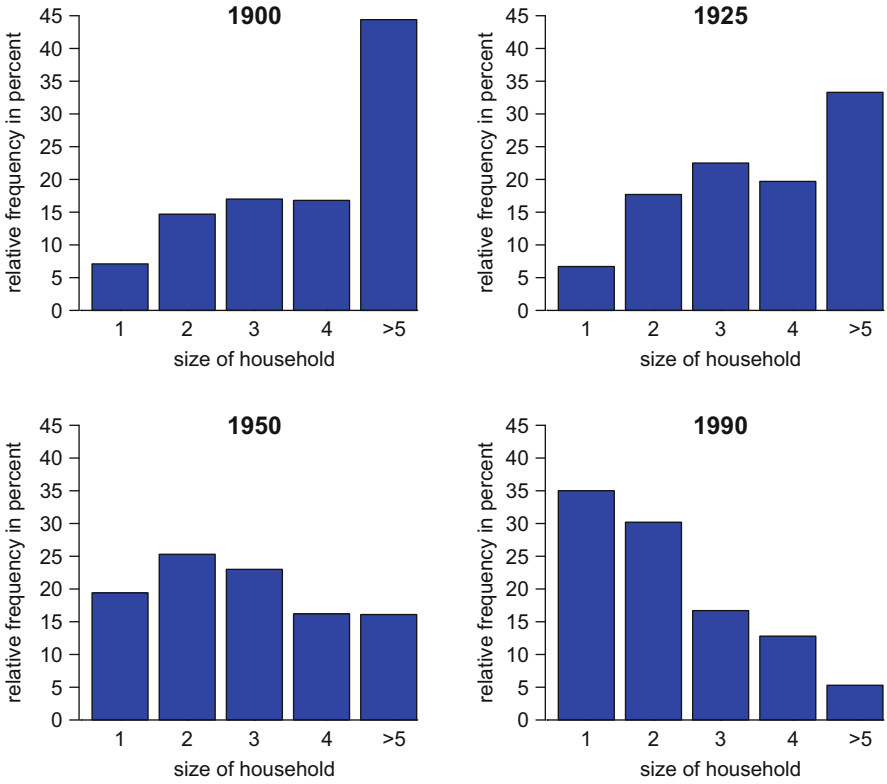


Fig. 2.7 Histograms on the evolution of household sizes over the twentieth century

2.2 Frequency Distribution for Continuous Data

Given a sample $x_1, x_2, \ldots, x_n$ on a continuous variable X, we may group the data into k classes with class boundaries denoted by $x_1^l, x_1^u = x_2^l, x_2^u = x_3^l, \ldots, x_k^u$ and class widths $\Delta x_j = x_j^u - x_j^l$ $(j = 1, \ldots, k)$. Note that the upper boundary for a given class is equal to the lower boundary for the succeeding class.

An observation x_i belongs to class j, if $x_j^l \leq x_i < x_j^u$. Since within a category, there are a range of possible values we will focus on the midpoint and denote it

Table 2.4 Structure of a frequency table

Class #	Classes	Absolute frequencies	Relative frequencies
1	$x_1^l \leq X < x_1^u$	$h(x_1)$	$f(x_1)$
2	$x_2^l \leq X < x_2^u$	$h(x_2)$	$f(x_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	$x_j^l \leq X < x_j^u$	$h(x_j)$	$f(x_j)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$x_k^l \leq X < x_k^u$	$h(x_k)$	$f(x_k)$
	Total	n	1

by x_j. (Contrast this with the discrete data case where x_j denotes the value for the category.) Once again the subscript j corresponds to categories x_j, $j = 1, \ldots, k$ and the subscript i denotes observations x_i, $i = 1, \ldots, n$.

Frequency Table

A *frequency table* for continuous data provides the distribution of frequencies over the given classes. The structure of a frequency table is shown in Table 2.4.

Graphical Presentation

Histogram

In a histogram, continuous data that have been grouped into categories are represented by rectangles. Class boundaries are marked on the horizontal axis. As they can be of varying width, we cannot simply represent frequencies by the heights of bars as we did for bar graphs. Rather, we must correct for class widths. The rectangles are constructed so that their areas are equal to the corresponding absolute or relative frequencies.

$$\hat{h}(x_j) \cdot \Delta x_j = \frac{h(x_j)}{x_j^u - x_j^l} \cdot \left(x_j^u - x_j^l\right) = h(x_j)$$

or

$$\hat{f}(x_j) \cdot \Delta x_j = \frac{f(x_j)}{x_j^u - x_j^l} \cdot \left(x_j^u - x_j^l\right) = f(x_j)$$

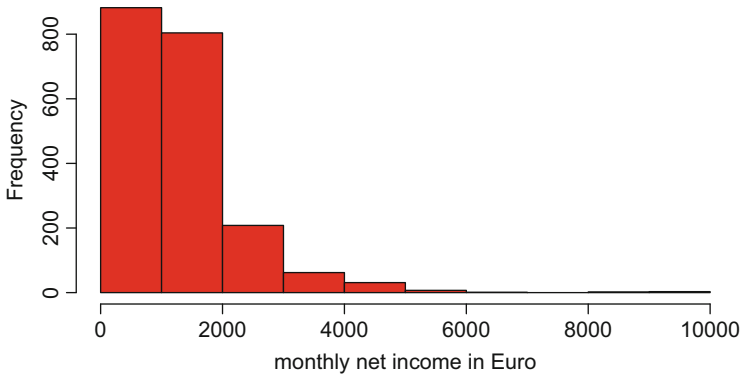


Fig. 2.8 Example of histogram—716 observations on monthly income (Euro)

If the class widths are identical, then the frequencies are also proportional to the heights of the rectangles. The rectangles are drawn contiguous to each other, reflecting common class boundaries $x_j^u = x_{j+1}^l$ (Fig. 2.8).

Stem-and-Leaf Display

In stem-and-leaf displays (plots), the data are not summarized using geometric objects. Rather, the actual values are arranged to give a rough picture of the data structure. The principle is similar to that of the bar chart, but values belonging to a particular class are recorded horizontally rather than being represented by vertical bars. Classes are set up by splitting the numerical observations into two parts: One or more of the leading digits make up the stem, the remaining (trailing) digits are called leaves. All observations with the same leading digits, i.e., the same stem, belong to one class. Typically, class frequencies are proportional to the lengths of the lines.

The principle is best understood by applying it to real data. Consider the following collection of observations :

$$32, 32, 35, 36, 40, 44, 47, 48, 53, 57, 57, 100, 105$$

The "stems" consist of the following "leading digits": $3, 4, 5, 10$. They correspond to the number of times that "ten" divides into the observation. The resulting stem-and-leaf diagram is displayed below.

Frequency	Stems	Leaves
4	3	2256
4	4	0478
3	5	377
2	10	05

Displaying data graphically (or, as is the case here, quasi-graphically), we can extract more relevant information than we could otherwise. (The human brain is comparatively efficient at storing and comparing visual patterns.)

The above stem-and-leaf plot appears quite simple. We can refine this by splitting the lines belonging to one stem in two, the first one for the trailing digits in the range one to four, the second for five to nine. We label the first group with *l* for low, the second with *h* for high. In the resulting stem-and-leaf plot the data appears approximately evenly distributed:

Frequency	Stems		Leaves
2	3	l	22
2	3	h	56
2	4	l	04
2	4	h	78
1	5	l	3
2	5	h	77
1	10	l	0
1	10	h	5

Yet there is an apparent gap between stems 5 and 10. It is indeed one of the advantages of stem-and-leaf plots that they are helpful in both giving insights into concentration of data in specific regions and spotting extraordinary or extreme observations. By labeling 100 and 105 as outliers we obtain a useful enhancement to the stem-and-leaf plot:

Frequency	Stems		Leaves
2	3	l	22
2	3	h	56
2	4	l	04
2	4	h	78
1	5	l	3
2	5	h	77
2	Extremes: 100, 105		

For an example with data conveying a richer structure of concentration and a more detailed stem structure have a look at the following examples for grouped continuous data.

Dotplots

Dotplots are used to graphically display small datasets. For each observation, a “dot” (a point, a circle or any other symbol) is plotted. Some data will take on the same values. Such ties would result in “overplotting” and thus would distort the display of the frequencies.

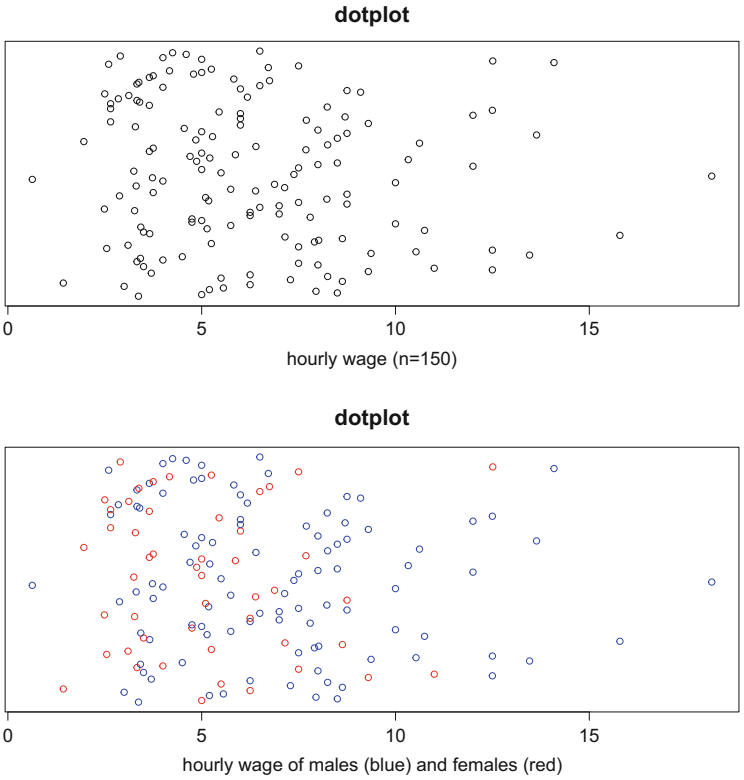


Fig. 2.9 Example of dotplot—student salaries in the USA

The dots are therefore spread out into the vertical dimension in a random fashion. The y-axis thus contains uniformly spread random numbers over the $[0, 1]$ interval. Provided, the size of each symbol is sufficiently small for a given sample size, the dots are then unlikely to overlap each other.

Example The data in Fig. 2.9 consist of 150 observations on student salaries in the USA. In the upper part panel, we display a dot plot for all 150 observations. In the lower part, we use color to distinguish the gender of the students. Since the random perturbations in the vertical dimension are different for the two panels, the points are located in slightly different positions.

Explained: Petrol Consumption of Cars

Petrol consumption of 74 cars has been measured in miles per gallon (MPG). The measurements are displayed in a frequency table shown in Table 2.5. Using the same

Table 2.5 Petrol consumption of 74 cars in miles per gallon (MPG)

X: Petrol consumption (MPG)	Absolute frequencies $h(x_j)$	Relative frequencies $f(x_j)$
$12 \leq X < 15$	8	0.108
$15 \leq X < 18$	10	0.135
$18 \leq X < 21$	20	0.270
$21 \leq X < 24$	13	0.176
$24 \leq X < 27$	12	0.162
$27 \leq X < 30$	4	0.054
$30 \leq X < 33$	3	0.041
$33 \leq X < 36$	3	0.041
$36 \leq X < 39$	0	0.000
$39 \leq X < 42$	1	0.013
Total	74	1.000

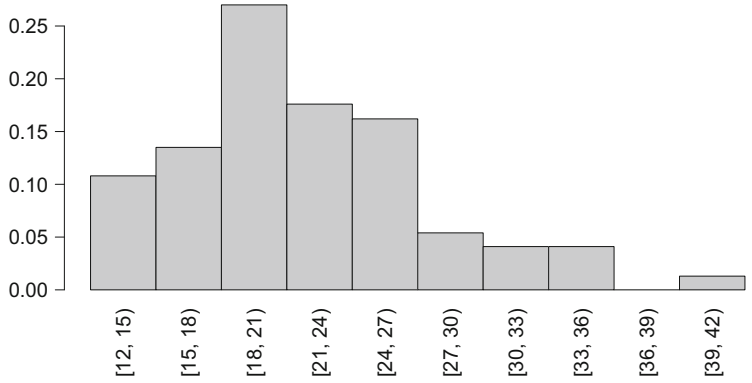


Fig. 2.10 Histogram for petrol consumption of 74 cars in miles per gallon (MPG)

constant class width of 3 MPG, the frequency distribution is displayed in a histogram in Fig. 2.10. As is evident from both, the frequency table and the histogram, the largest proportion of cars lies in the category 18–21 MPG.

Explained: Net Income of German Nationals

Data

Statistical elements:	German nationals, residing in private households, minimum age 18
Statistical variable:	monthly net income
sample size n	716

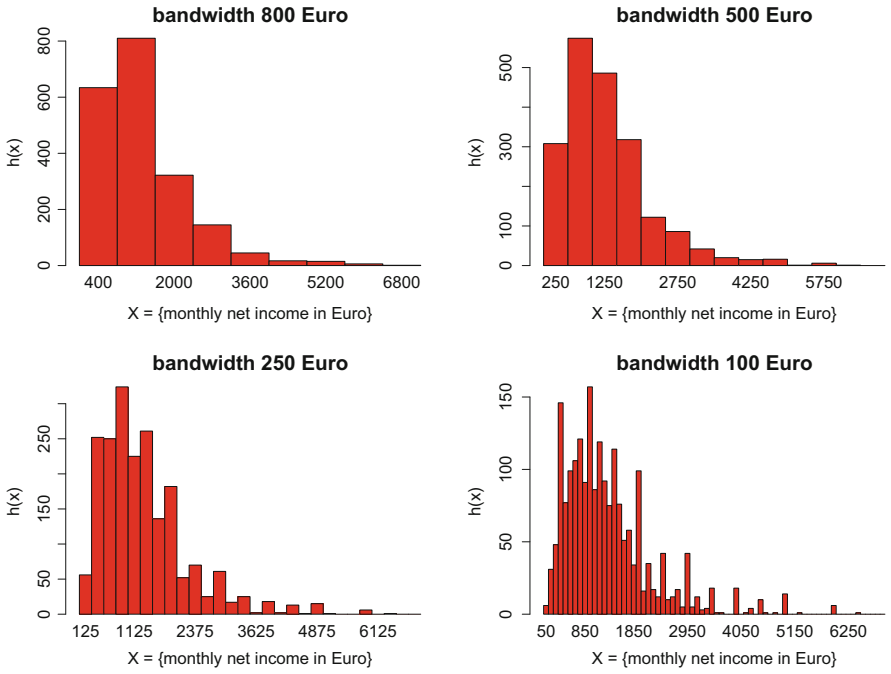


Fig. 2.11 Histograms of monthly net income in Euro for different bandwidths

Histogram

In the histograms shown in Fig. 2.11, the classes are income brackets of equal width. Reducing the common class size (and hence increasing the number of classes) yields a more detailed picture of the income distribution. Observe how the absolute frequencies decline as the class widths become more narrow.

Furthermore, increasing the number of classes decreases the smoothness of the graph. Additional gaps become visible as more information about the actual data is displayed. In choosing a class width we are striking a balance between two criteria: the essential information about the population which might be more strikingly conveyed in a smoother graph, and greater detail contained in a histogram with a larger number of classes.

We can also separate histograms by gender, using a bin width of 500 Euro, as shown in Fig. 2.12.

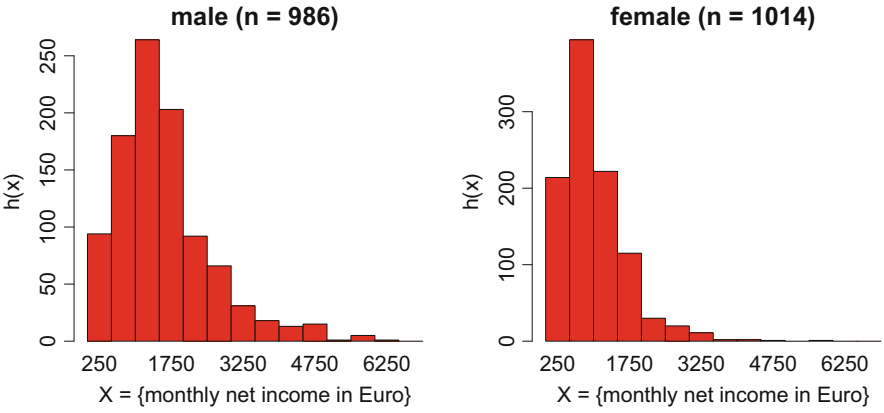


Fig. 2.12 Histograms of monthly net income in Euro for males and females

Stem-and-Leaf Display

The stem-and-leaf plot provided in Table 2.6 displays all 716 income figures. It is more detailed than the stem-and-leaf plots we have previously drawn. The stems, specified by the first leading digit, are divided into five subclasses corresponding to different values in the first trailing, i.e., leaf digit: The first line of each stem, denoted by *, lists all leaves starting with 0 or 1, the second (t) those starting with 2 or 3, and so on. As the stem width is specified to be 1000, the first leaf digit counts the hundreds. To condense exposition, each two observations belonging to the same class (i.e., being the same leaf) are represented by just one number (leaf). For example, six of the 716 surveyed persons earn between 2400 and 2500 Euros, denoted by "444" in the "2 f" line.

The ampersand (&) denotes pairs of observations covering both leaves represented by one line. For example, 4 persons earn between 4200 and 4400 Euros. Following the convention of each leaf representing two cases, there are two persons with net earnings in the interval [4200, 4300). The other two persons, symbolized by &, would be displayed by the sequence "23," if one leaf represented one observation. Thus, one of the two persons belongs to the income bracket [4200, 4300), the other to the [4300, 4400)-bracket.

Observe, that the 17 "extreme" values are displayed separately to highlight their distance from the other more heavily populated classes.

Table 2.6 Stem-and-leaf plot

Frequency	Stem and Leaf		
2	0	*	1
21	0	t	2233333333
35	0	f	44444444555555555
47	0	s	66666666666666667777777
41	0	.	88888888888899999999
45	1	*	000000000000000111111
38	1	t	2222222222222233333
63	1	f	44444444444555555555555555555
45	1	s	6666666666667777777777
72	1	.	88888888888888888888888889999999999
78	2	*	00000000000000000000000000000001111111
46	2	t	22222222222222333333333
32	2	f	444555555555555
28	2	s	66666667777777
23	2	.	88888889999
28	3	*	00000000000011
10	3	t	2233
16	3	f	44555555
8	3	s	6677
5	3	.	88
12	4	*	00000&
4	4	t	2&
10	Extremes:		(4400), (4500), (5000),(5500), (5600),(5900), (6400), (6500), (7000), (15000)
Stem width: 1000			
Each leaf: 2 case(s), & denotes fractional leaves			

2.3 Empirical Distribution Function

Empirical distribution functions can be constructed for data that have a natural numerical ordering. If $h\left(x_j\right)$ is the absolute frequency of observations on a discrete variable, then the absolute frequency (or number) of observations not exceeding that value is called the *absolute cumulated frequency*:

$$H\left(x_j\right) = \sum_{s=1}^{j} h\left(x_s\right), \qquad j = 1, \ldots, k$$

The *relative cumulative frequency* is calculated as:

$$F\left(x_j\right) = \frac{H\left(x_j\right)}{n} = \sum_{s=1}^{j} f\left(x_s\right), \qquad j = 1, \ldots, k$$

If the variable is continuous and the data are grouped into k classes, then the above definitions apply except that we interpret $H(x_j)$ as the frequency of observations not exceeding the upper boundary of the j-th class.

2.3.1 Empirical Distribution Function for Discrete Data

For the *relative cumulative frequency* we have

$$F(x) = \begin{cases} 0 & \text{if } x < x_1 \\ \sum_{s=1}^{j} f\left(x_s\right) & \text{if } x_j \leq x < x_{j+1}, \quad j = 2, \ldots, k \\ 1 & \text{if } x_k \leq x \end{cases}$$

The graph of an empirical distribution function is a monotonically increasing step function, the step size corresponds to the relative frequency at the "jump" points x_j (Table 2.7; Fig. 2.13).

In creating empirical distribution functions we are not losing information about relative frequencies of observations, as we can always reverse the cumulation process:

$$f\left(x_j\right) = F\left(x_j\right) - F\left(x_{j-1}\right), \quad \text{for } j = 1, \ldots, k; F\left(x_0\right) = 0$$

Suppose $x_l < x_u$ are two values that the discrete variable can take. Then the number or frequency of observations taking on values between x_l and x_u can be calculated as follows:

$$F\left(x_{u-1}\right) - F\left(x_l\right)$$

Table 2.7 Example of cumulative frequencies for number of persons in a household—data from 1990

# persons per household	$f\left(x_j\right)$	$F\left(x_j\right)$
1	0.350	0.350
2	0.302	0.652
3	0.167	0.819
4	0.128	0.947
≥ 5	0.053	1.000

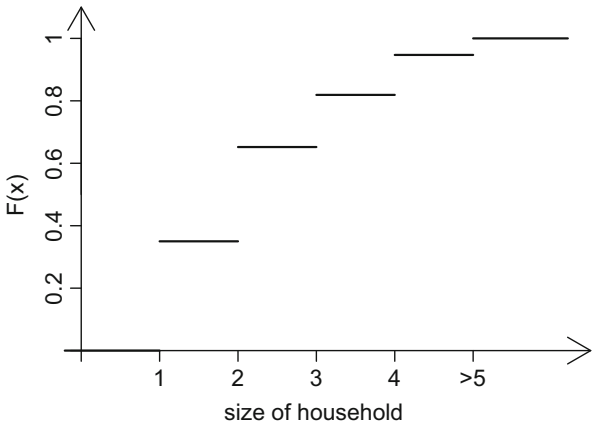


Fig. 2.13 Distribution function for the number of persons in a household—data from 1990

2.3.2 *Empirical Distribution Function for Grouped Continuous Data*

As for discrete data, the empirical distribution function for grouped continuous data is a function of relative cumulative frequencies. But in this case, rather than using a step function, one plots the cumulative frequencies against the upper boundaries of each class, then joints the points with straight lines. Mathematically, the empirical distribution function may be written as:

$$F(x) = \begin{cases} 0 & \text{if } x < x_1^l \\ \sum_{i=1}^{j-1} f(x_i) + \frac{x - x_j^l}{x_j^u - x_j^l} \cdot f(x_j) & \text{if } x_j^l \le x < x_j^u, \quad j = 1, \ldots, k \\ 1 & \text{if } x_k^u \le x \end{cases}$$

The rationale for interpolating with straight lines is that one might expect the distribution of points within classes to be approximately uniform.

An Example is provided in Table 2.8. The corresponding distribution function is given in Fig. 2.14.

As mentioned earlier, the straight lines connecting class boundaries reflect linear interpolations motivated by the assumption that observations are evenly distributed within classes. We will illustrate this by drawing the variable part of the distribution function for $x_j^l \le x < x_j^u$, $\sum_{i=1}^{j-1} f(x_i) + \frac{x - x_j^l}{x_j^u - x_j^l} f(x_j)$, for a fixed interval (class) $\left[x_j^l, x_j^u\right)$.

Evaluating at a lower class boundary we obtain $F\left(x_j^l\right) = \sum_{i=1}^{j-1} f(x_i) + \frac{x_j^l - x_j^l}{x_j^u - x_j^l} f(x_j) = \sum_{i=1}^{j-1} f(x_i)$. We can thus substitute $F\left(x_j^l\right)$ for $\sum_{i=1}^{j-1} f(x_i)$ in the

Table 2.8 Example—lives of 100 light bulbs

Statistical elements	Light bulbs			
Statistical variable	Life in hours, metric variable			
sample size n	100			
X: Life (hours)	$h(x_j)$	$f(x_j)$	$H(x_j)$	$F(x_j)$
$0 \leq X < 100$	1	0.01	1	0.01
$100 \leq X < 500$	24	0.24	25	0.25
$500 \leq X < 1000$	45	0.45	70	0.70
$1000 \leq X < 2000$	30	0.30	100	1.00
Total	100	1.00		

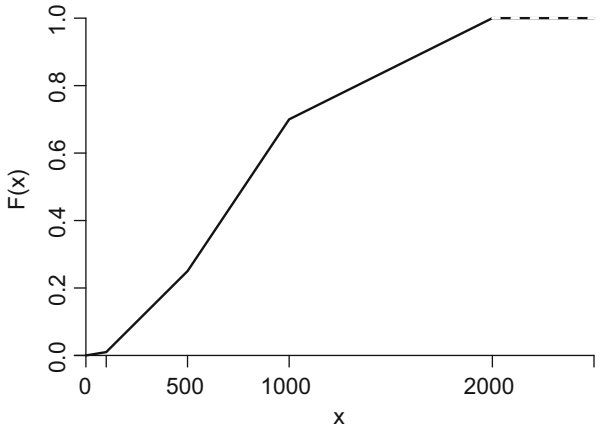


Fig. 2.14 Cumulative distribution function for lives of 100 light bulbs

formula for the distribution function and get

$$F(x) = F\left(x_j^l\right) + \frac{x - x_j^l}{x_j^u - x_j^l} \quad \text{if } x_j^l \leq x < x_j^u\,, \quad j = 1, \ldots, k$$

Figure 2.15 depicts the linear intra-class segment.

Explained: Petrol Consumption of Cars

The petrol consumption of 74 cars has been measured in miles per gallon (MPG). The measurements are displayed in an augmented frequency table shown in Table 2.9. The corresponding empirical distribution function is given in Fig. 2.16.

Again, the linear interpolation of lower class boundaries follows from the assumption of an even distribution of observations within classes. Class widths and boundaries are in turn constructed to approximate this assumption as closely as possible. This allows us to retain as much information as possible about the shape of the data.

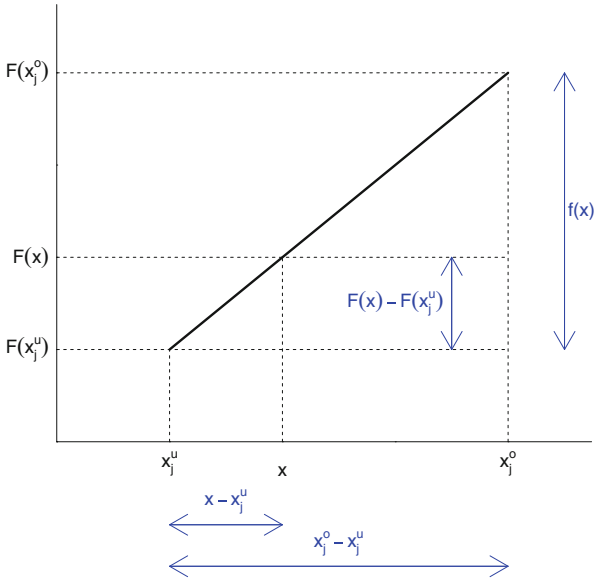


Fig. 2.15 Linear intra-class segment for distribution function

Table 2.9 Augmented frequency table for petrol consumption of 74 cars measured in miles per gallon (MPG)

X: Petrol consumption (MPG)	Absolute frequencies $h(x_j)$	Relative frequencies $f(x_j)$	Relative cumulative frequencies $F(x_j)$
$12 \leq X < 15$	8	0.108	0.108
$15 \leq X < 18$	10	0.135	0.243
$18 \leq X < 21$	20	0.270	0.513
$21 \leq X < 24$	13	0.176	0.689
$24 \leq X < 27$	12	0.162	0.851
$27 \leq X < 30$	4	0.054	0.905
$30 \leq X < 33$	3	0.041	0.946
$33 \leq X < 36$	3	0.041	0.987
$36 \leq X < 39$	0	0.000	0.987
$39 \leq X < 41$	1	0.013	1.000
Total	74	1.000	

Various statements can be extracted from Table 2.9, e.g.: 68.9 % of cars cannot travel more than 24 miles per gallon.

Explained: Grades in Statistics Examination

These are the grades 20 students have achieved in a Statistics examination:

$$\{2, 2, 4, 1, 3, 2, 5, 4, 2, 4, 3, 2, 5, 1, 3, 2, 2, 3, 5, 4\}$$

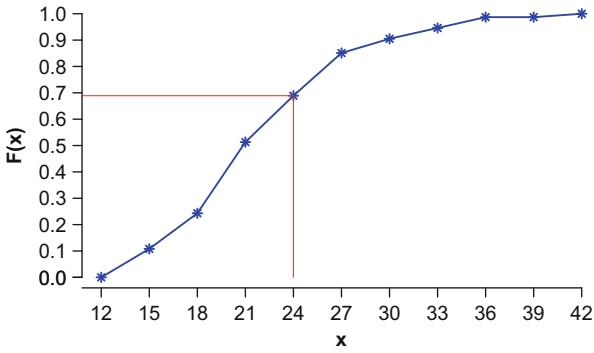


Fig. 2.16 Empirical distribution function for petrol consumption of 74 cars measured in miles per gallon (MPG)

Table 2.10 Frequency table of grades in statistics examination

X: Mark	Absolute frequency $h(x_j)$	Relative frequency $f(x_j)$	Relative cumulative frequency $F(x_j)$
1	2	0.10	0.10
2	7	0.35	0.45
3	4	0.20	0.65
4	4	0.20	0.85
5	3	0.15	1.00

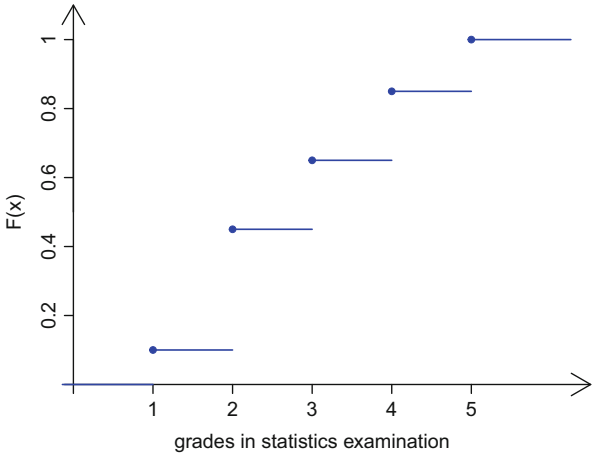


Fig. 2.17 Relative cumulative frequencies of grades in statistics examination

Table 2.10 summarizes the information about the distribution of the given data. The graph of the relative cumulative frequencies is depicted in Fig. 2.17. We observe that the graph of the relative cumulative frequency (and hence the function) is continuous from the right. Each bullet indicates the value of the distribution function at a jump point. In the figure, the x-axis covers all real numbers within the grade range, even though the random variable cannot take other values than $\{1,\ 2,\ 3,\ 4,\ 5\}$. For theoretical reasons, the definition of the distribution function

also assigns numbers (zero and one, respectively) to values outside $[1, 5]$. Various statements can be deduced from the data summarized in the frequency table, e.g.

- 65 % of students have achieved a grade of at least 3.
- 15 % $(1.00 - 0.85)$ of students achieved a grade of 5.

2.4 Numerical Description of One-Dimensional Frequency Distributions

Statistics are numbers which summarize particular features of the data. Formally, a statistic is a function of the data. They can be used to measure different features, such as where the data are generally located (measures of location), the degree to which they are dispersed (measures of dispersion or scale), whether they are symmetrically distributed, the degree to which they are correlated, and so on. In the following sections we will consider various measures of location and dispersion. These measures can then be used to compare different datasets.

Measures of Location

In addition to summarizing where the data are located or concentrated, location measures provide a benchmark against which individual observations can be assessed.

Mode

The value occurring most frequently in a dataset is called the mode or the modal value. If the variable is discrete, the mode is simply the value with the greatest frequency. For continuous data measured with sufficient accuracy, however, most observations are likely to be distinct, rendering the idea meaningless. However, by grouping the data, we can determine the *modal class*, i.e., the class with the highest frequency.

Mode for qualitative or discrete data is given by

$$\arg\max_{x_j} \; \{f\left(x_j\right)\}$$

Mode for Grouped Continuous Data The modal class is the class with the highest class frequency. As a class interval consists of infinitely many numbers, we have to introduce a convention according to which a single number within this class is determined to represent the mode. The simplest convention is to use the midpoint of

the modal class. An alternative and more technical adjustment involves selecting a point which moves towards the neighboring cell with the higher density of observations. It is defined as follows:

$$x_D = x_j^l + \frac{\hat{f}\left(x_j\right) - \hat{f}\left(x_{j-1}\right)}{2 \cdot \hat{f}\left(x_j\right) - \hat{f}\left(x_{j-1}\right) - \hat{f}\left(x_{j+1}\right)} \cdot \left(x_j^u - x_j^l\right),$$

where

x_j^l,x_j^u	lower/upper boundary of modal class
$\hat{f}\left(x_j\right)$	frequency distribution for modal class
$\hat{f}\left(x_{j-1}\right)$	frequency distribution for class preceding modal class
$\hat{f}\left(x_{j+1}\right)$	frequency distribution for class succeeding modal class

The modal class is given by: $[500, 1000)$. We can calculate the mode approximated by the midpoint of the modal class which is just the arithmetic average of the class boundaries: $0.5 \cdot \left(x_j^u + x_j^l\right) = 750$ h. Using the above formula which moves the mid-point in the direction of the neighboring cell with the higher density of observations one obtains: $x_D = 500 + \frac{9-6}{18-6-3} \cdot 500 = 666\frac{2}{3}$ (Table 2.11).

Quantiles

Given data $x_1, x_2, \ldots, x_n$, suppose we order or rank the data in increasing order to obtain the ordered sequence $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$..We call the elements of this sequence the order statistics of the data. From the order statistics we can immediately read off the third largest value, the smallest value, and so on.

Let p be a number between zero and one and think of p as a proportion of the data. A value which divides the sequence of order statistics into the two subsequences containing the first $(p \cdot n)$ and the last $((1-p) \cdot n)$ observations is called the p-quantile. We will denote it by x_p. Equivalently, we may think of x_p as a value such that $100p\,\%$ of the data lie below it and $100(1-p)\,\%$ of the data lie above.

Table 2.11 Example—Lives of 100 light bulbs

j	X: Life (hours)	$h\left(x_j\right)$	$f\left(x_j\right)$	$\hat{f}\left(x_j\right) \cdot 10^{-4}$	$F\left(x_j\right)$
1	$0 \leq X < 100$	1	0.01	1	0.01
2	$100 \leq X < 500$	24	0.24	6	0.25
3	$500 \leq X < 1000$	45	0.45	9	0.70
4	$1000 \leq X < 2000$	30	0.30	3	1.00
	Total	100	1.00		

Quantiles for Ungrouped Data

- If $n \cdot p$ is not an integer and k the smallest integer satisfying $k > n \cdot p$, then we define $x_p = x_{(k)}$. The quantile is thus the observation with rank k, $x_{(k)}$.
- If, $k = n \cdot p$ is an integer, we will take x_p to be the midpoint between $x_{(k)}$.and $x_{(k+1)}$.

Quantiles for Grouped Data For data that are grouped in classes, we will carry out interpolations between class boundaries to obtain a p-quantile:

$$x_p = x_j^l + \frac{p - F\left(x_j^l\right)}{f\left(x_j\right)} \cdot \left(x_j^u - x_j^l\right)$$

Here, x_j^l, x_j^u and $f\left(x_j\right)$ are the lower boundary, upper boundary, and the relative frequency of the class containing the p-th quantile. The cumulative relative frequency up to and including the class preceding the quantile class is denoted by $F\left(x_j^l\right)$.

The quantile x_p can be defined using interpolation. The principle of interpolation for the quantity $p = F(x_p)$ can be easily understood from Fig. 2.18.

Some special quantiles:

- deciles (tenths)—the ordered observations are divided into ten equal parts. $p = s/10, s = 1, \ldots, 9$—deciles: $x_{0.1}, x_{0.2}, \ldots, x_{0.9}$
- quintiles—the ordered observations are divided into five equal parts. $p = r/5, r = 1, 2, 3, 4$—quintiles: $x_{0.2}, x_{0.4}, x_{0.6}, x_{0.8}$
- quartiles—the ordered observations are divided into four equal parts. $p = q/4, q = 1, 2, 3$—quartiles: $x_{0.25}, x_{0.5}, x_{0.75}$

Median (Central Value)

The value which divides the ordered observations into two equal parts is called the median $x_z = x_{0.5}$. The median is much less sensitive to outlying or extreme observations than other measures such as the mean which we study below. The median x_z corresponds to the second quartile $x_{0.5}$.

Median for Ungrouped Data

- for n odd: $x_{0.5} = x_{(\frac{n+1}{2})}$
- for n even: $x_{0.5} = (x_{(n/2)} + x_{(n/2+1)})/2$. This is simply the mid-point of the two center-most observations.

Median for Grouped Variables

- The median for grouped data is defined as the mid-point of the class which contains the central portion of the data.

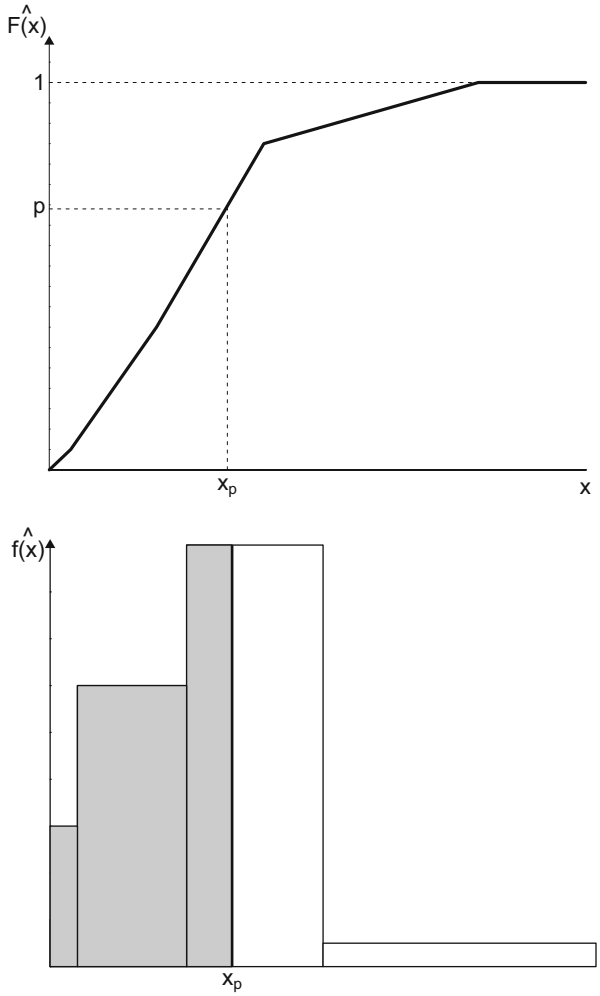


Fig. 2.18 Quantiles of grouped data

- Formally, let x_j^l and x_j^u be the lower and upper boundaries of the class for which $F(x_{j-1}^u) = F(x_j^l) \leq 0.5$ and $F(x_j^u) \geq 0.5$. Then,

$$x_{0.5} = x_j^l + \frac{0.5 - F(x_j^l)}{f(x_j)} \cdot (x_j^u - x_j^l)$$

- The median can be easily determined from the graph of the distribution function since $F(x_{0.5}) = 0.5$, see Fig. 2.19.

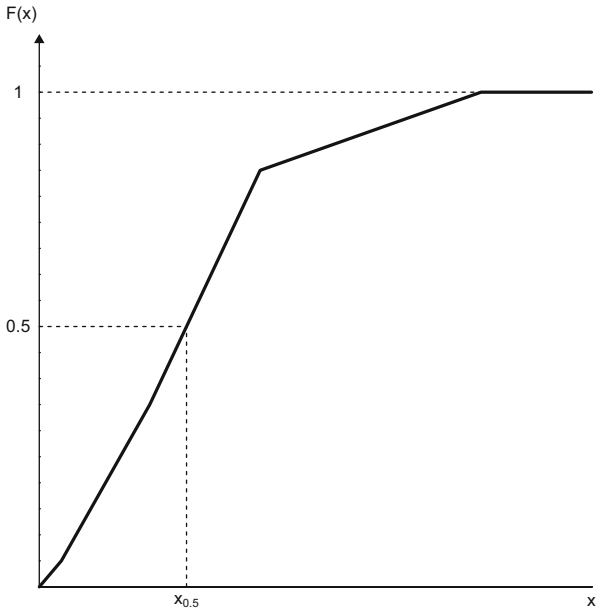


Fig. 2.19 Median for grouped continuous data

Properties of the Median (of Numerical Variables)

- optimality

$$\sum_{i=1}^{n} |x_i - x_{0.5}| = \sum_{j=1}^{k} |x_j - x_{0.5}| \cdot f(x_j) \rightarrow min.$$

  The median is optimal in the sense that it minimizes the sum of absolute deviations of the observations from a point that lies in the midst of the data (Fig. 2.19).
- linear transformation $y_i = a + bx_i \longrightarrow y_{0.5} = a + bx_{0.5}$

  If the data are transformed linearly, then the median is shifted by that same linear transformation.

Calculation of Quartiles The empirical distribution function (third column of the Table 2.12) implies that both the first quartile $x_{0.25}$, $p = 0.25$ and the second quartile $x_{0.5}$, $p = 0.50$ belong to third group (3000–5000 EUR). By interpolation we find

Table 2.12
Example—Monthly net income of households (up to 25000 EUR)

Income range (EUR)	Proportion of households: $f(x)$	Empirical distribution function: $F(x)$
1–800	0.044	0.044
800–1400	0.166	0.210
1400–3000	0.471	0.681
3000–5000	0.243	0.924
5000–25000	0.076	1.000

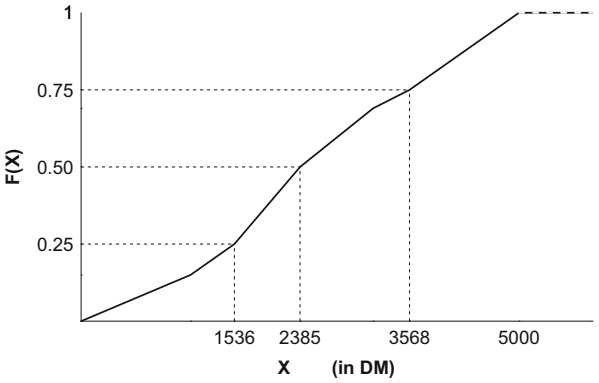


Fig. 2.20 Graph of the empirical distribution function and quartiles

the following (Fig. 2.20).

$$x_{0.25} = 1400 + 1600 \cdot \frac{0.25 - 0.21}{0.471} = 1535.88 \text{ EUR}$$

$$x_{0.50} = 1400 + 1600 \cdot \frac{0.50 - 0.21}{0.471} = 2385.14 \text{ EUR}$$

$$x_{0.75} = 3000 + 2000 \cdot \frac{0.75 - 0.681}{0.243} = 3567.90 \text{ EUR}$$

The Interpretation 25 % of the households has net monthly income not exceeding 1535.88 EUR and 75 % of the households has income higher than 1535.88 EUR (first quartile). 50 % of the households have income smaller than 2385.14 EUR and 50 % of the households have income higher than 2385.14 EUR (second quartile). 75 % of the households have income less than 3567.90 EUR and 25 % of the households have income exceeding 3567.90 EUR (third quartile).

The above also implies that 50 % of the households has net income between 1535.88 EUR and 3567.90 EUR.

Arithmetic Mean

The arithmetic mean or average, denoted $\bar{X}$, is obtained by summing all observations and dividing by n. The arithmetic mean is sensitive to outliers. In particular, an extreme value tends to "pull" the arithmetic mean in its direction.

The mean can be calculated in various ways, using the original data, using the frequency distribution and using the relative frequency distribution. For discrete data, each method yields a numerically identical answer.

Calculation using original data:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Calculation using the frequency and relative frequency distribution:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} x_j h(x_j) = \sum_{j=1}^{k} x_j f(x_j)$$

Properties of the Arithmetic Mean

- Center of gravity: The sum of the deviations of the data from the arithmetic mean is equal to zero.

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0 \quad \Leftrightarrow \quad \sum_{j=1}^{k}(x_j - \bar{x})h(x_j) = 0$$

- Minimum sum of squares: The sum of squares of the deviations of the data from the arithmetic mean is smaller than the sum of squares of deviations from any other value c.

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 < \sum_{i=1}^{n}(x_i - c)^2$$

$$\sum_{j=1}^{k}(x_j - \bar{x})^2 h(x_j) < \sum_{j=1}^{k}(x_j - c)^2 h(x_j)$$

- Pooled data: Assume that the observed data are in disjoint sets $D_1,\ D_2,\ \ldots,\ D_r$, and that the arithmetic mean $\bar{x}_p$ for each of the sets is known. Then the arithmetic mean of all observed values (considered as one set) can be calculated using the formula

$$\bar{x} = \frac{1}{n}\sum_{p=1}^{r} \bar{x}_p n_p \qquad n = \sum_{p=1}^{r} n_p$$

where n_p denotes the number of observations in p-th group ($p = 1, \ldots, r$).

Table 2.13
Example 1—Monthly income of households (MIH)

MIH (EUR)	Proportion of households $f(x)$	Cumulative distribution function $F(x)$
1–800	0.044	0.044
800–1400	0.166	0.210
1400–3000	0.471	0.681
3000–5000	0.243	0.924
5000–25000	0.076	1.000

Table 2.14
Example 2—Monthly income of 716 people

$\bar{x} = 1881.40$ EUR
$x_{0.25} = 1092.50$ EUR
$x_{0.50} = 1800.00$ EUR
$x_{0.75} = 2400.00$ EUR
'mode' = 2000.00 EUR

- Linear transformation:

$$y_i = a + bx_i \longrightarrow \bar{y} = a + b\bar{x}$$

- Sum:

$$z_i = x_i + y_i \longrightarrow \bar{z} = \bar{x} + \bar{y}$$

From the data of Example 1 given in Table 2.13 we can calculate the arithmetic mean using the mid-points of the groups:

$$\bar{x} = 400 \cdot 0.044 + 1100 \cdot 0.166 + 2200 \cdot 0.471 + 4000 \cdot 0.243 + 15000 \cdot 0.076$$
$$= 17.6 + 182.6 + 1036.2 + 972 + 1140 = 3348.4 \text{ EUR}.$$

The arithmetic mean 3348.4 EUR is higher than the median calculated above (2385.14 EUR). This can be explained by the fact that the arithmetic mean is more sensitive to the relatively small number of large incomes. The high values shift the arithmetic mean but do not influence the median (Table 2.14).

Explained: Average Prices of Cars

This dataset contains prices (in USD) of 74 cars. The distribution of prices is displayed using a dotplot below. The price variable is on the horizontal axis. The data are randomly scattered in the vertical direction for better visualization.

In Fig. 2.21, the median is displayed in red and the arithmetic mean in magenta. As can be seen, the two values almost coincide.

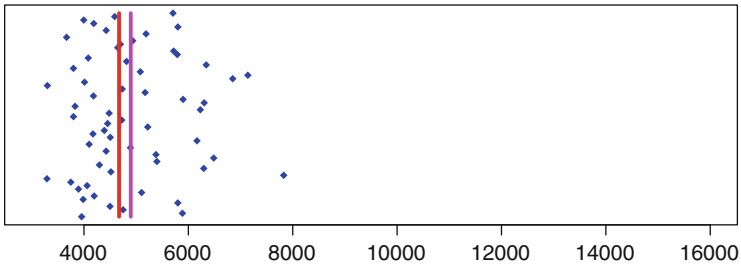


Fig. 2.21 Prices for 74 cars (USD)—arithmetic mean: 4896.417 (*magenta*) and median: 4672.000 (*red*)

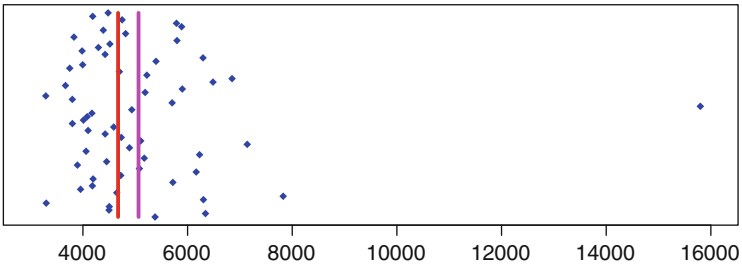


Fig. 2.22 Corrected prices for 74 cars (USD)—arithmetic mean: 5063.083 (*magenta*) and median: 4672.000 (*red*)

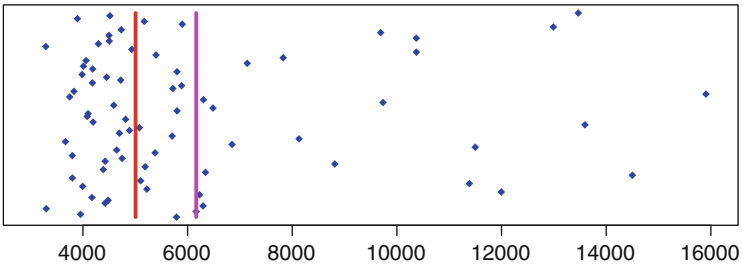


Fig. 2.23 Repeated measurements of car prices—arithmetic mean: 5063.083 (*magenta*) and median: 5006.500 (*red*)

For symmetric distributions, the median and arithmetic mean are identical. This is almost true for our example.

However, during a check of the data, it was discovered that one value had not been entered correctly. The value 15962 USD was incorrectly changed to 5962 USD. Figure 2.22 contains corrected values:

The median (because it is robust) did not change. On the other hand, the arithmetic mean has increased significantly, as it is sensitive to extreme values. The miscoded observation takes on a value well outside the main body of the data.

The measurements were repeated after some time with the results shown in Fig. 2.23.

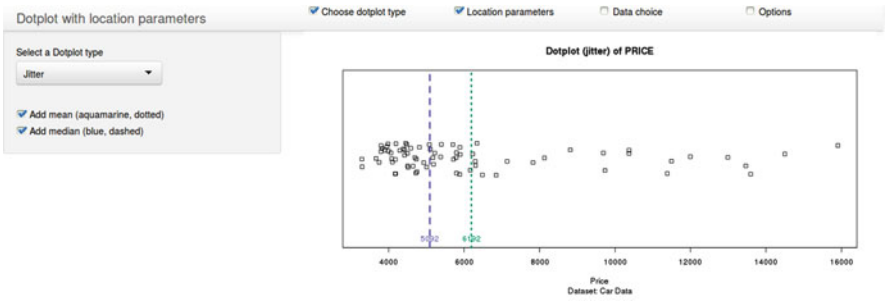


Fig. 2.24 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_dot1

Now, there are a number of relatively more expensive cars. The distribution of prices is now skewed to the right. These more extreme observations pull the mean to the right much more so than the median. Thus for right-skewed distributions, the arithmetic mean is larger than the median.

Interactive: Dotplot with Location Parameters

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please select

- a dotplot type, e.g., jitter
- if you like the mean and median to be included in the plot

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

Output

The interactive example allows us to display a one-dimensional frequency distribution in the form of a dotplot for a variety of variables. Possible values are displayed along the horizontal axis. For easier visualization, the observations may be randomly shifted (jitter) in the vertical direction. The median and the arithmetic mean can be displayed graphically and numerically (Fig. 2.24).

Interactive: Simple Histogram

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

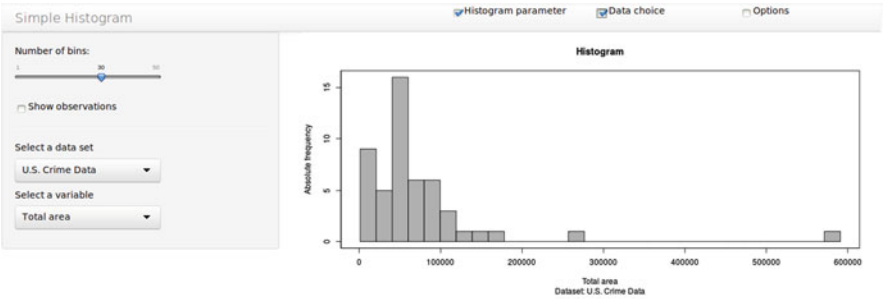


Fig. 2.25 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_hist

Please select

- the number of bins
- if you like the observations to be shown

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

Output

The graphic displays all observations of a variable summarized in a histogram (Fig. 2.25).

2.5 Location Parameters: Mean Values—Harmonic Mean, Geometric Mean

If the observed variables are ratios, then the arithmetic mean may not be appropriate.

Harmonic Average

The harmonic average, denoted $\bar{x}_H$, is useful for variables which are ratios. We assume that all data points are not equal to zero, i.e., $x_i \neq 0$. As a consequence the $x_j \neq 0$.

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

$$\bar{x}_H = \frac{\sum_{j=1}^{k} g_j}{\sum_{j=1}^{k} \frac{g_j}{x_j}}, \quad j = 1, \ldots, k$$

In the latter formula, g_j provides additional information which will become clear in the example below.

Example 1

Part of the road j	1	2	3	4
Distance g_j in km	2	4	3	8
Speed x_j in km/h	40	50	80	100

We would like to calculate the average speed of the car during the period of travel. It is inappropriate to simply average the speeds since they are measured over differing periods of time. In the table, g_j is the distance traveled in each segment. Using the above formula we calculate:

Total time: $\sum_{j=1}^{k} \frac{g_j}{x_j} = 0.2475\,\text{h}$

Total distance: $\sum_{j=1}^{k} g_j = 17\,\text{km}$

Average: $\bar{x}_H = \frac{17}{0.2475} = \frac{2+4+3+8}{\frac{2}{40}+\frac{4}{50}+\frac{3}{80}+\frac{8}{100}} = 68.687\,\text{km/h}$

The arithmetic mean would lead to an incorrect result 67.5 km/h, because it does not account for the varying lengths of the various parts of the road. Correct use of the arithmetic mean would involve calculating the time spent along each segment. In the above example these times are denoted by $h_j = g_j/x_j$ for each segment.

$$h_1 = g_1/x_1 = 0.05;\ h_2 = g_2/x_2 = 0.08;$$

$$h_3 = g_3/x_3 = 0.0375;\ h_4 = g_4/x_4 = 0.08;$$

$$\bar{x} = \frac{40 \cdot 0.05 + 50 \cdot 0.08 + 80 \cdot 0.0375 + 100 \cdot 0.08}{0.05 + 0.08 + 0.0375 + 0.08} = 68.687\,\text{km/h}$$

Thus, in order to calculate the average of ratios using additional information for the **numerator** (in our case x_j with the additional information g_j) we use the **harmonic average**. In order to calculate the average from ratios using additional information on the **denominator**, we choose the **arithmetic average**.

Example 2 Four students, who have part time jobs, have the hourly (respectively weekly) salaries given in Table 2.15.

We are supposed to find the average hourly salary. This calculation cannot be done using only the arithmetic average of the hourly salaries, because that would

Table 2.15 Hourly and weekly salary of four students

Student	Euro/h	Weekly salary in Euro
A	18	180
B	20	300
C	15	270
D	19	380

Table 2.16 Hourly salary and working hours of four students

Student	Euro/h	Working hours
A	18	10
B	20	15
C	15	18
D	19	20

not take into account the different times spent in the job. The variable of interest is a ratio (Euro/h) and the additional information (weekly salary in Euro) is related to the numerator of this ratio. Hence, we will use the harmonic average.

$$\bar{x}_H = \frac{\sum\limits_j g_j}{\sum\limits_j \frac{g_j}{x_j}} = \frac{180 + 300 + 270 + 380}{\frac{180}{18} + \frac{300}{20} + \frac{270}{15} + \frac{380}{19}} = \frac{1130}{63} = 17.94$$

These four students earn on average **17.94 Euro/h** (Table 2.15). The situation changes if we are given the number of hours worked per week (instead of the weekly salary).

Now, the additional information (weekly working hours) is related to the denominator of the ratio. Hence, we can use an arithmetic average, in this case the **weighted arithmetic average**.

$$\bar{x} = \frac{18 \cdot 10 + 20 \cdot 15 + 15 \cdot 18 + 19 \cdot 20}{10 + 15 + 18 + 20} = \frac{1130}{63} = 17.94$$

The average salary is again **17.94 Euro/h**.

Geometric Average

The geometric mean, denoted $\bar{x}_G$, is used to calculate the mean value of variables which are positive, are ratios (e.g., rate of growth) and are multiplicatively related.

$$\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n}$$

The logarithm of the geometric average is equal to the arithmetic average of the logarithms of the observations:

$$\log \bar{x}_G = \frac{1}{n} \sum_{i=1} n \log x_i$$

Mean Growth Rate and Forecast

Let $x_0, x_1, \ldots, x_n$ be the measurements ordered according to the time of observation from 0 to n. The growth rates can be calculated as

$$i_t = x_t / x_{t-1}$$
$$i_1 \cdot i_2 \cdot \cdots \cdot i_n = x_n / x_0$$

The product of all growth rates is equal to the total growth from time 0 to n. **The average growth rate** will be obtained as a geometric average of the growth rates in distinct time periods:

$$\bar{i}_g = \sqrt{n}\, i_1 \cdot i_2 \cdot \cdots \cdot i_n = \sqrt{n} \frac{x_n}{x_0}$$

Knowing the mean growth rate and the value in time n, we can **forecast** the value in time $n + T$.

$$x^{\star}_{n+T} = x_n \cdot (\bar{i}_G)^T$$

Solving this equation with respect to T, we obtain a formula for the time which is necessary to reach the given value:

$$T = \frac{\log(x_{n+T}) - \log(x_n)}{\log(\bar{i}_G)}$$

Example 1

Now we calculate:

- mean value (geometric average)
- forecast for 1990
- time (year), when GDP reaches the value 2500.

$$\bar{i}_G = \sqrt[8]{\frac{1971.8}{1733.8}} = 1.0162$$
$$x^{\star}_{1990} = 1971.8 \cdot 1.0162^2 = 2036.2\,\text{bn DM}$$
$$T = \frac{\log(2500) - \log(1971.8)}{\log(1.0162)} = 14.77\,\text{years}.$$

Table 2.17 Gross domestic product (GDP) for Germany in 1985 prices (bn DM)

Year	t	GDP x_t	i_t
1980	0	1733.8	–
1981	1	1735.7	1.0011
1982	2	1716.5	0.9889
1983	3	1748.4	1.0186
1984	4	1802.0	1.0307
1985	5	1834.5	1.0180
1986	6	1874.4	1.0217
1987	7	1902.3	1.0149
1988	8	1971.8	1.0365

Table 2.18 German stock index (DAX) during the period 1990–1997

Year	1989	1990	1991	1992	1993	1994	1995	1996	1997
DAX (end of the year)	1791	1399	1579	1546	2268	2107	2254	2889	4250
DAX (change)		−21.9 %	12.9 %	−2.1 %	46.7 %	−7.1 %	7.0 %	28.0 %	47.1 %

The value of GDP of 2500 is forecasted in year $1988 + 15 = 2003$ (Table 2.17).

Example 2

The German stock index (DAX) was changing during the period 1990–1997, as shown in Table 2.18.

We want to find the average yearly change in the DAX over the period. Use of the arithmetic average leads to an **incorrect result** as illustrated below.

- $\bar{x} = \frac{(-21.9)+(12.9)+(-2.1)+(46.7)+(-7.1)+(7.0)+(28.2)+(47.1)}{8} = \frac{110.80}{8} = 13.85\,\%$
- Starting in the year 1989 and using the "average change of DAX" to calculate the value of the DAX in 1997, one obtains:

  1990 $1791 \cdot 1.1385 = 2093$

  1991 $2093 \cdot 1.1385 = 2383$

  ... ...

  1997 $4440 \cdot 1.1385 = \mathbf{5055}$
- The result **5055** is much higher than the actual value of the DAX in 1997 which was **4250**.

The correct mean value is, in this case, the geometric mean, because it measure the growth during a certain period. The value of DAX in 1990 can be calculated from the value in 1989 and the relative change as follows:

$$\begin{aligned} \text{DAX}_{1990} &= (1 + (-0.219)) \cdot \text{DAX}_{1989} \\ &= (1 + (-0.219)) \cdot 1791 = 0.781 \cdot 1791 = 1399 \end{aligned}$$

Analogously, we can "forecast" the value for 1991 from the relative change and the value of DAX in 1990:

$$\begin{aligned}\mathrm{DAX}_{1991} &= (1 + 0.129) \cdot \mathrm{DAX}_{1990} \\ &= (1 + 0.129)) \cdot 1399 = 1.129 \cdot 1399 = 1579\end{aligned}$$

The values are multiplicatively related. The geometric mean yields the following:

$$\begin{aligned}X_G &= \sqrt[8]{0.781 \cdot 1.129 \cdot 0.979 \cdot 1.467 \cdot 0.929 \cdot 1.070 \cdot 1.282 \cdot 1.417} \\ &= 1.1141\end{aligned}$$

The average growth rate per year of the DAX over the period 1990–1997 was **11.41 %**. Using this geometric mean and the value of DAX in 1989 to predict the value of DAX in 1997, we obtain the correct result:

1990	$1791 \cdot 1.1141 = 1995$
1991	$1995 \cdot 1.1141 = 2223$
...	...
1997	$3815 \cdot 1.1141 = \mathbf{4250}$

The average growth rate of DAX in 1990–1997 can be used also to forecast the value of at the end of year 1999. We obtain the prediction:

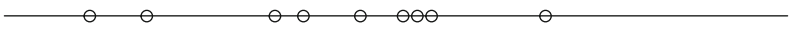
$$\mathrm{DAX}_{1999} = \mathrm{DAX}_{1997} \cdot 1.1141 \cdot 1.1141 = 4250 \cdot 1.1141^2 = 5275$$

2.6 Measures of Scale or Variation

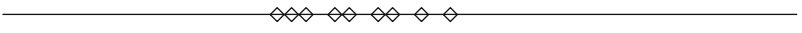
The various measures of location outlined in the previous sections are not sufficient for a good description of one-dimensional data. An illustration of this follows:

Monthly expenditures for free time and holidays (in EUR):

- data from 10 two person households: 210, 250, 340, 360, 400, 430, 440, 450, 530, 630 displayed on the axis:



- data from 10 four person households: 340, 350, 360, 380, 390, 410, 420, 440, 460, 490 displayed on the axis:



The arithmetic average $\bar{X}$ is in both cases is equal to 404 EUR, but the graphs show visible differences between the two distributions. For households with four people the values are more concentrated around the center (in this case the mean) than for households with two people, i.e., the spread or variation is smaller.

Measures of scale measure the variability of data. Together with measures of location (such as means, medians, and modes) they provide a reasonable description of one-dimensional data. Intuitively one would want measures of dispersion to have the property that if the same constant was added to each of the data-points, the measure would be unaffected. A second property is that if the data were spread further apart, for example through multiplication by a constant greater than one, the measure should increase.

Range

The range is the simplest measure of scale:

Range for Ungrouped Data

- The range, denoted R, is defined as the difference between the largest and the smallest observed value

$$R = x_{\max} - x_{\min} = x_{(n)} - x_{(1)}$$

  where $x_{(1)}, \ldots, x_{(n)}$ are the ordered data, i.e., the order statistics.

Range for Grouped Data

- For grouped data, the range R is defined as the difference between the upper bound of the last (highest) class x_k^u and the lower bound of the first (smallest) class x_1^l:

$$R = x_k^u - x_1^l$$

Properties

- For a linear transformation we have: $y_i = a + bx_i \longrightarrow R_y = |b|R_x$

  Note that addition of the constant a which merely shifts the data does not affect the measure of variability.

Interquartile Range

The interquartile range is the difference between the third quartile $x_{0,75}$ and the first quartile $x_{0,25}$:

$$QA = x_{0.75} - x_{0.25}$$

The interquartile range is the width of the central region which captures 50 % of the observed data. The interquartile range relative to the median is defined as $QA_r = QA/x_{0.5}$.

Properties

- Robust towards extreme values (outliers)
- Linear transformation: $y_i = a + bx_i \longrightarrow QA_y = |b|QA_x$
  Again addition of the constant a does not affect the measure of variability.

Mean Absolute Deviation

The mean of the absolute deviations of the observed values from a fixed point c is called the mean absolute deviation (MAD) and it is denoted by d. The fixed point c can be any value. Usually, it is chosen to be one of the measures of location; typically the mean $\bar{x}$ or median $x_{0.5}$.

As with the range and the interquartile range, adding the same constant to all the data. Multiplication by a constant rescales the measure by the absolute value of that same constant. Each of the formulas below may be used for ungrouped data. If the data have been grouped, then one would use the second formula where the x_j are mid-points of the classes, and $h(x_j)$ and $f(x_j)$ are the absolute and relative frequencies:

$$d = \frac{1}{n}\sum_{i=1}^{n}|x_i - c|$$

$$d = \frac{1}{n}\sum_{j=1}^{k}|x_j - c|h(x_j) = \sum_{j=1}^{k}|x_j - c|f(x_j)$$

Properties

- The optimality property of the median implies that the median is the value which minimizes the mean absolute deviation. Thus any other value substituted for c above would yield a larger value of this measure.
- For a linear transformation of the data: $y_i = a + bx_i \longrightarrow d_y = |b|d_x$

Example

- Observed values: 2, 5, 9, 20, 22, 23, 29
  $x_{0.5} = 20,\ d(x_{0.5}) = 8,29$
  $\bar{x} = 15.71,\ d(\bar{x}) = 8.90$

The Variance and the Standard Deviation

The mean of the squared deviations of the observed values from a certain fixed point c is called the mean squared error (MSE) or the mean squared deviation. The point c can be chosen ad libitum.

$$MQ(c) = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$$

$$MQ(c) = \frac{1}{n}\sum_{j=1}^{k}(x_j - c)^2 h(x_j) = \sum_{j=1}^{k}(x_j - c)^2 f(x_j)$$

The Variance If we choose the point c to be the mean $\bar{x}$, then the MSE is called the variance. The variance of the observed values will be denoted as s^2 and may be computed as follows.

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

$$s^2 = \frac{1}{n}\sum_{j=1}^{k}(x_j - \bar{x})^2 h(x_j) = \sum_{j=1}^{k}(x_j - \bar{x})^2 f(x_j)$$

Standard Deviation The standard deviation (s) is defined as the square root of the variance.

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

$$s = \sqrt{\frac{1}{n}\sum_{j=1}^{k}(x_j - \bar{x})^2 h(x_j)} = \sqrt{\sum_{j=1}^{k}(x_j - \bar{x})^2 f(x_j)}$$

The variance s^2 (and therefore also the standard deviation s) is always greater than or equal to 0. Zero variance implies that the observed data are all identical and consequently do not have any spread.

Properties

- The mean squared error with respect to $\bar{x}$ (the variance) is smaller than the mean square error with respect to any other point c. This result can be proved as follows:

$$\begin{aligned} MSE(c) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - c)^2 \\ &= \frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 + 2(\bar{x} - c)\sum_{i=1}^{n}(x_i - \bar{x}) + n(\bar{x} - c)^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + (\bar{x} - c)^2 = s^2 + (\bar{x} - c)^2 \end{aligned}$$

  The middle term of the middle line vanishes since $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. These formulas imply that the mean square error $MSE(c)$ is always greater than or equal to the variance. Obviously equality holds only if $c = \bar{x}$.
- For linear transformations we have: $y_i = a + bx_i \longrightarrow s_y^2 = b^2 s_x^2,\ s_y = |b| s_x$
- Standardization: by subtracting the mean and dividing by the standard deviation one creates a new dataset for which the mean is zero and the variance is one. Let: $z_i = a + bx_i$, where $a = -\bar{x}/s_x,\ b = 1/s_x$, then

$$z_i = \frac{x_j - \bar{x}}{s_x}$$
$$\Rightarrow \bar{z} = 0, \quad s_z^2 = 1$$

Example

- Observed values: 2, 5, 9, 20, 22, 23, 29
- $x_{0.5} = 20 \qquad MSE(x_{0.5}) = 109.14$
- $\bar{x} = 15.71 \qquad MSE(\bar{x}) = Variance = 90.78$

Theorem (pooling) *Let us assume that the observed values (data) are divided into r groups with n_i $i = 1, .., r$ observations. Assume also that the means and variances in these groups are known. To obtain the variance s^2 of the pooled data we may use:*

$$s^2 = \sum_{i=1}^{r}\frac{n_i}{n}s_i^2 + \sum_{i=1}^{r}\frac{n_i}{n}(\bar{x_i} - \bar{x})^2$$

$\bar{x_1}, \ldots, \bar{x_r}$ are the arithmetic averages in the groups
$s_1^2, \ldots, s_r^2$ are the variances in the groups
$n_1, \ldots, n_r$ are numbers of observations in the groups, $n = n_1 + \cdots + n_r$

Variance Decomposition The above formula illustrates that the variance can be decomposed into two parts:

Total variance = variance *within* the groups + variance *between* the groups.

Coefficient of Variation In order to compare the standard deviations for different distributions, we introduce a relative measure of scale (relative to the mean), the so-called coefficient of variation. The coefficient of variation expresses variation as a percentage of the mean:

$$v = s/\bar{x} \quad \bar{x} > 0$$

Example The mean values and the standard deviations of two sets of observations are:

$\bar{x}_1 = 250 \quad s_1 = 10$
$\bar{x}_2 = 750 \quad s_2 = 30$

By comparing the standard deviations, we conclude that the variation in the second dataset is three times higher than the variation in the first. But, in this case it would be more appropriate to compare the coefficients of variation since the data have very different means:

$v_1 = 10/250 = 0.04$
$v_2 = 30/750 = 0.04$

The relative spread of both datasets is the same.

Explained: Variations of Pizza Prices

The price (in EUR) of Dr. Oetker pizza was collected in 20 supermarkets in Berlin (Fig. 2.26):

3.99; 4.50; 4.99; 4.79; 5.29; 5.00; 4.19; 4.90; 4.99; 4.79; 4.90; 4.69; 4.89; 4.49; 5.09; 4.89; 4.99; 4.29; 4.49; 4.19

- The average price for a pizza in these 20 supermarkets is **4.27 Euro (= mean)**
- The median price is **4.84 Euro (= median)**
- The difference between the highest and smallest price is **1.30 Euro (= range)**
- If the MAD is calculated around the mean it is **0.29 Euro (= MAD)** if calculated around the median it is **0.28 Euro (= MAD)**.
- 50 % of all prices lie in the interval between **4.49 Euro (quartile $x_{0.25}$) and 4.99 Euro (quartile $x_{0.75}$)**, this interval is of width **0.50 Euro (= interquartile range)**.2
- Mean square error around the mean is **0.12241 Euro^2 (= variance)**, the square root of the variance is **0.34987 Euro (= standard deviation)**.

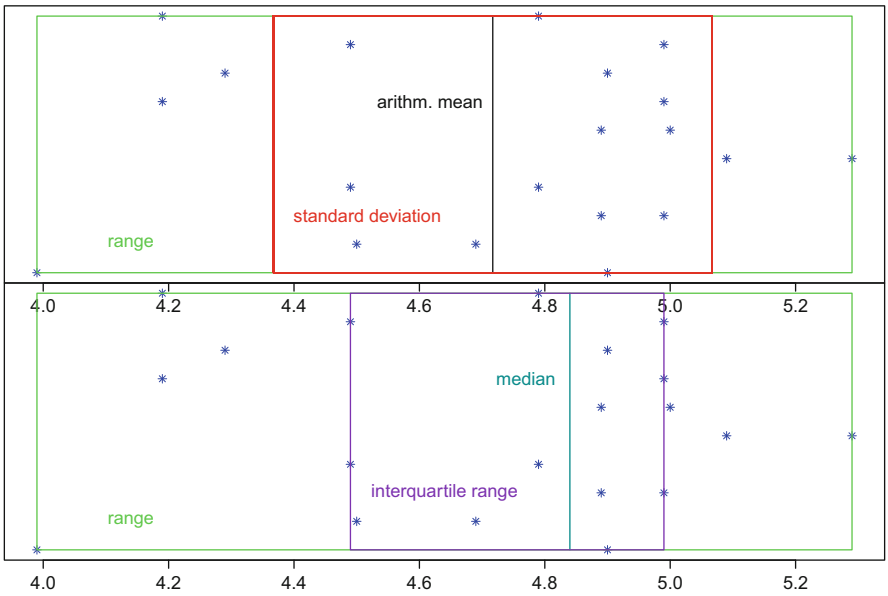


Fig. 2.26 Prices for pizza in 20 supermarkets—parameters of scale

Enhanced: Parameters of Scale for Cars

The price of 74 types of cars in USD was collected in 1985. The data are displayed in Fig. 2.27. The upper panel displays the range (green), arithmetic average (black), and the standard deviation (red). The lower panel displays the range (green), median (mint green), and the interquartile range (magenta).

Arithmetic average:	4896.417
Median:	4672
Range	4536
Interquartile range	1554.75
Standard deviation	991.2394

During a check of the data, it was discovered that there was an input error. The correct value of 15962 USD was incorrectly recorded as 5962 USD. Figure 2.28 contains the corrected results.

Arithmetic average:	5063.083
Median:	4672
Range	12508
Interquartile range	1554.75
Standard deviation	1719.064

It is clear that the range increased, because it is a function of the extreme values. The value of interquartile range did not change since no prices within this range were altered. The standard deviation increased significantly. The reason is that

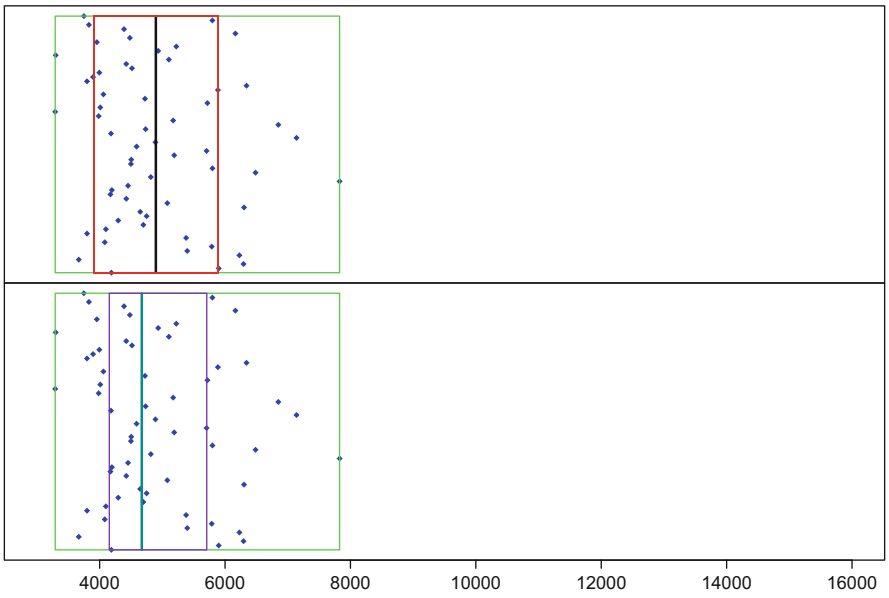


Fig. 2.27 Prices of 74 cars in USD—upper panel: range (*green*), arithmetic average (*black*), and the standard deviation (*red*); lower panel: range (*green*), median (*mint green*), and the interquartile range (*magenta*)

standard deviation is calculated from all observed prices and involves the squares of deviations which causes it to be particularly sensitive to extreme values (outliers).

The investigation was repeated after some time. The results are presented in Fig. 2.29.

Arithmetic average:	6165.257
Median:	5006.5
Range	12615
Interquartile range	2112
Standard deviation	2949.496

Now, there are a number of expensive vehicles whose prices are substantially different from the lower priced cars. Thus the price are skewed to the right. For skewed distributions, the standard deviation is typically higher than the interquartile range. This feature is demonstrated in the above example.

Interactive: Dotplot with Scale Parameters

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

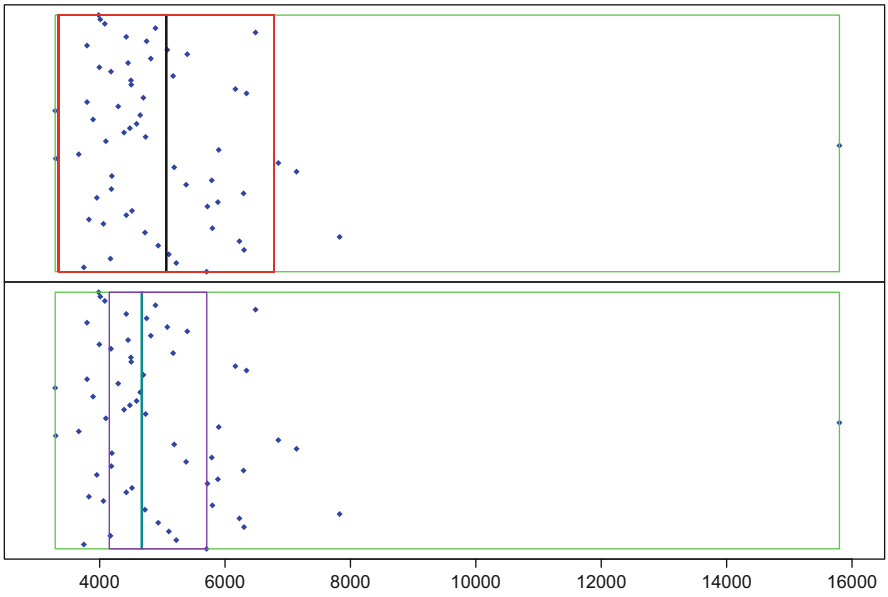


Fig. 2.28 Corrected prices of 74 cars in USD—upper panel: range (*green*), arithmetic average (*black*), and the standard deviation (*red*); lower panel: range (*green*), median (*mint green*), and the interquartile range (*magenta*)

Please select

- a dotplot type, e.g., jitter
- if you like the mean, median, range, or interquartile range to be included in the plot

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

Output

The interactive example in Fig. 2.30 allows us to display a one-dimensional frequency distribution in the form of a dotplot for a variety of variables. Possible values are displayed along the horizontal axis. For easier visualization, the observations may be randomly shifted (jitter) in the vertical direction. Furthermore, the median, the arithmetic mean, range, and interquartile range can be included.

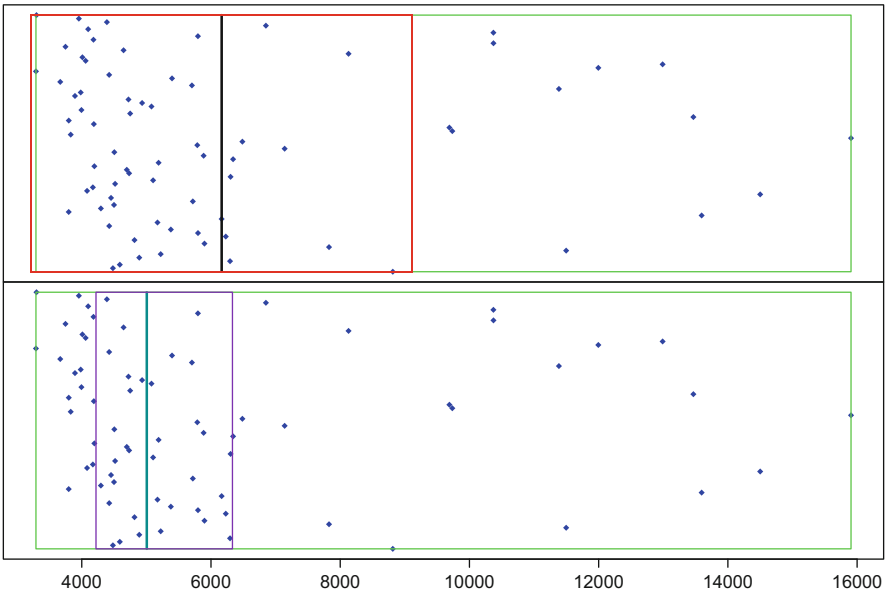


Fig. 2.29 Repeated investigation of prices of 74 cars in USD—upper panel: range (*green*), arithmetic average (*black*), and the standard deviation (*red*); lower panel: range (*green*), median (*mint green*), and the interquartile range (*magenta*)

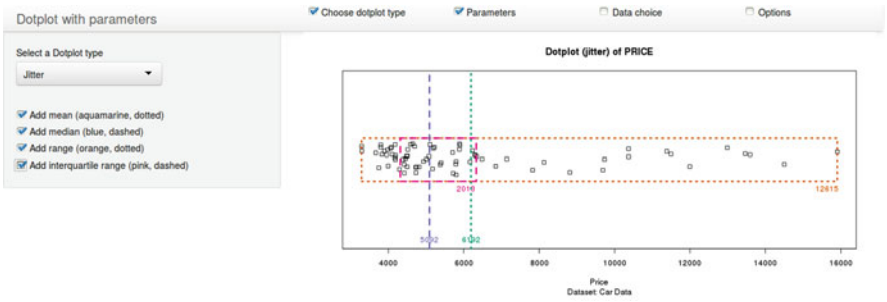


Fig. 2.30 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_dot2

2.7 Graphical Display of the Location and Scale Parameters

Boxplot (Box-Whisker-Plot)

Unlike the stem-and-leaf diagram, the boxplot does not contain information about all observed values. It displays only the most important information about the frequency distribution. Specifically, the boxplot contains the smallest and the largest

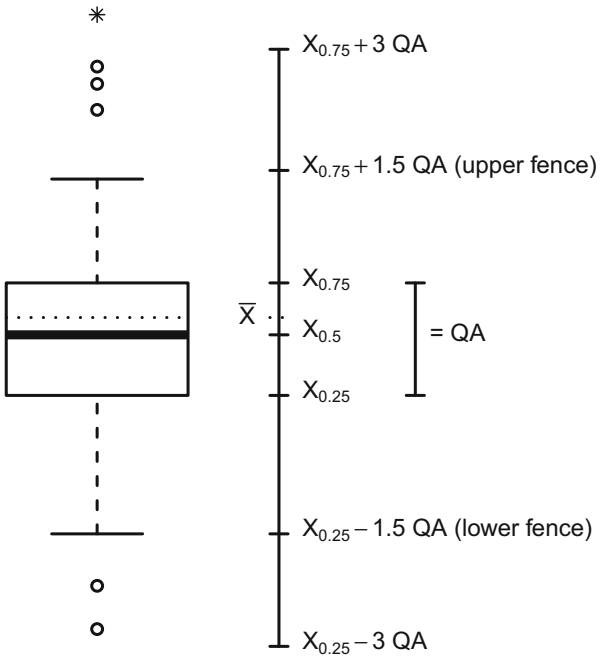
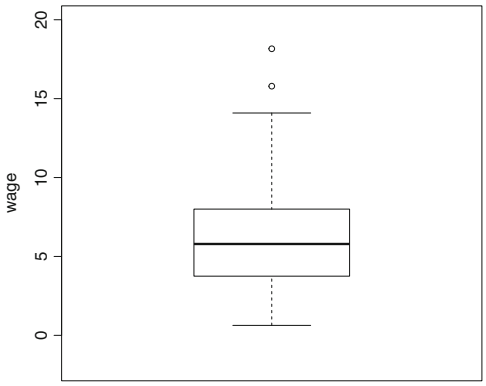


Fig. 2.31 The structure of a boxplot

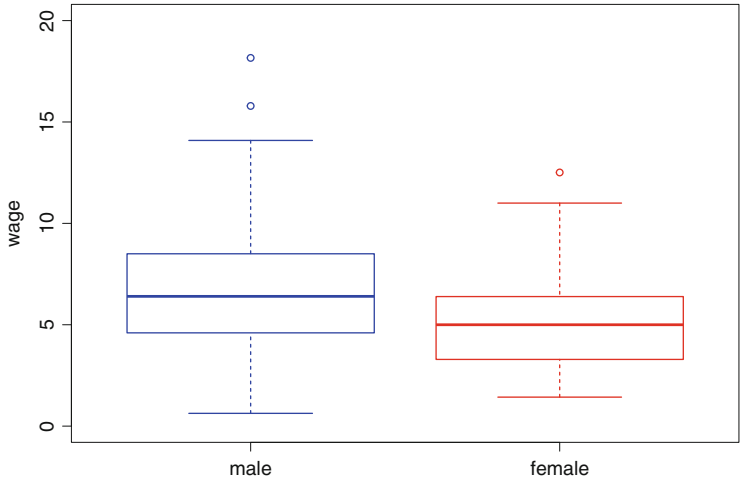
observed values $x_{(1)}$ and $x_{(n)}$ and three quartiles $x_{0.25}, x_{0.5}ax_{0.75}$. The second quartile $x_{0.5}$ is of course the median (Fig. 2.31).

The quartiles are denoted by a line and the first and third quartile are connected so that we obtain a box. The line inside this box denotes the median. The height of this box is the interquartile range which is the difference between the third and the first quartile: $x_{0.75}$ and $x_{0.25}$. Inside this box, one finds the central 50 % of all observed values.

The whiskers show the smallest and largest values within a 1.5 multiple of the interquartile range calculated from the boundary of the box. The bounds $x_{0.25} - 1.5 \cdot QA$ and $x_{0.75} + 1.5 \cdot QA$ are called the lower and upper fence, respectively. The values lying outside the fences are marked as outliers with a different symbol. Usually, the boxplot also displays the mean as a dashed line. The boxplot provides quick insight into the location, scale, shape, and structure of the data.



Example—boxplot of student salaries in USD



Example—boxplot of student salaries in USD; males and females separated

Explained: Boxplot of Car Prices

The prices of 74 types of cars were obtained in 1983. The results are displayed in Fig. 2.32.

The upper panels of the graphs contain dotplots. The lower panels show boxplots. The values lying outside a 1.5 multiple (resp. 3 multiple) of the interquartile range are denoted as extreme (outlying) observations. These outlying observations produce a large difference between the median (solid line) and the mean (dashed line).

Table 2.19
Example—Student salaries in USD

Total	Men	Women
$x_{\min} = 1$	$x_{\min} = 1$	$x_{\min} = 1.74997$
$x_{\max} = 44.5005$	$x_{\max} = 26.2903$	$x_{\max} = 44.5005$
$R = 43.5005$	$R = 25.2903$	$R = 42.7505$
$x_{0.25} = 5.24985$	$x_{0.25} = 6.00024$	$x_{0.25} = 4.74979$
$x_{0.5} = 7.77801$	$x_{0.5} = 8.92985$	$x_{0.5} = 6.79985$
$x_{0.75} = 11.2504$	$x_{0.75} = 12.9994$	$x_{0.75} = 10.0001$
$QA = 6.00065$	$QA = 9.99916$	$QA = 5.25031$
$\bar{x} = 9.02395$	$\bar{x} = 9.99479$	$\bar{x} = 7.87874$
$s^2 = 26.408$	$s^2 = 27.9377$	$s^2 = 22.2774$
$s = 5.13887$	$s = 5.28562$	$s = 4.7199$
$v = 0.57$	$v = 0.53$	$v = 0.60$

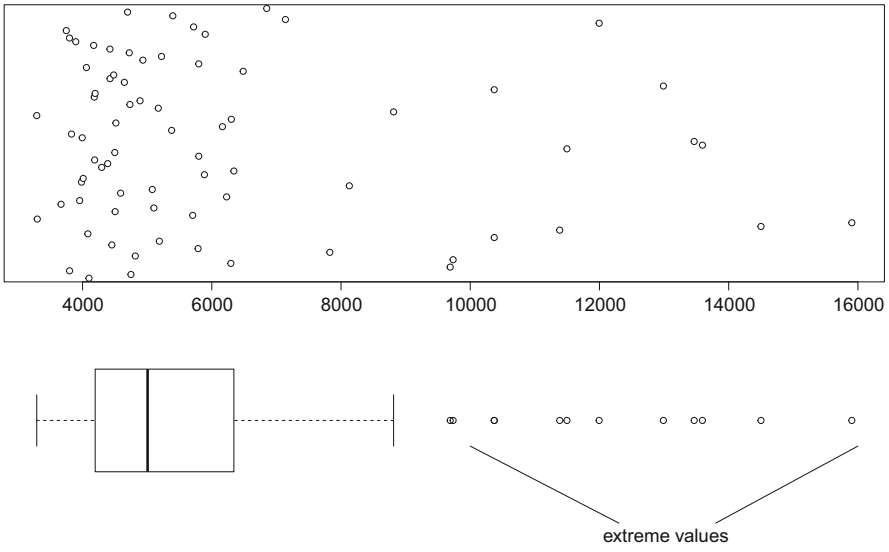


Fig. 2.32 Boxplot of prices of 74 cars

Interactive: Visualization of One-Dimensional Distributions

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please choose

- a dotplot type, e.g., jitter
- the number of bins for the histogram
- if you like the mean and median to be included in the plots

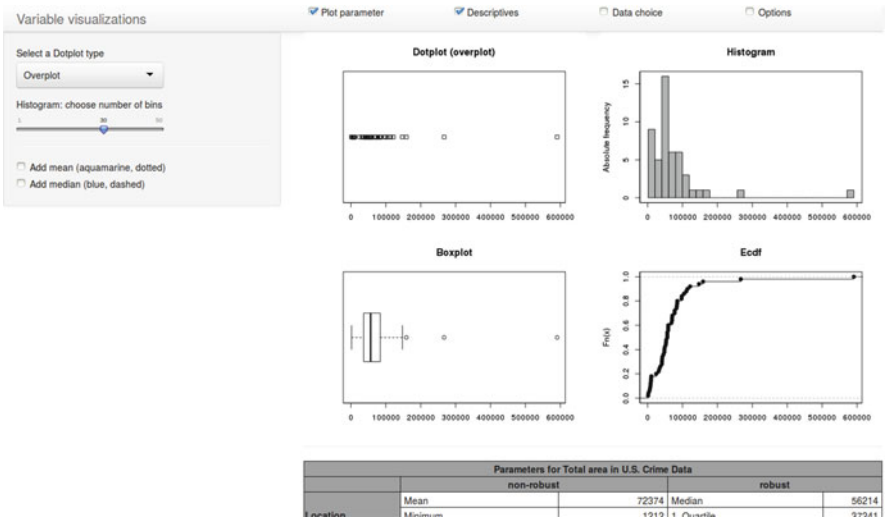


Fig. 2.33 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_vis

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

Output

The interactive example in Fig. 2.33 allows us to display a one-dimensional frequency distribution in the form of a dotplot, a histogram, a boxplot, and cumulative distribution function for a variety of variables. Possible values are displayed along the horizontal axis. For easier visualization, the observations may be randomly shifted (jitter) in the vertical direction. Furthermore, the median and the arithmetic mean can be included. You also receive a table showing the numerical values of certain parameters.

Chapter 3
Probability Theory

3.1 The Sample Space, Events, and Probabilities

Probability theory is concerned with the outcomes of random experiments. These can be either real world processes or thought experiments. In both cases,

- the experiment has to be infinitely repeatable and
- there has to be a well-defined set of outcomes.

The set of all possible outcomes of an experiment is called the sample space which we will denote by S.

Consider the process of rolling a die. The set of possible outcomes is the set $S = \{1, 2, 3, 4, 5, 6\}$. Each element of S is a basic outcome. However, one might be interested in whether the number thrown is even, or whether it is greater than 3, and so on. Thus we need to be able to speak of various combinations of basic outcomes, that is subsets of S.

An event is defined to be a subset of the set of possible outcomes S. We will denote an event using the symbol E. Events which consist of only one element, such as a two was thrown, are called *simple events* or *elementary events*. Simple events are by definition not divisible into more basic events, as each of them includes one and only one possible outcome.

Example Rolling a single die once results in the occurrence of one of the simple events {1}, {2}, {3}, {4}, {5}, {6}. As we have indicated, the sample space S is {1,2,3,4,5,6}.

Example For tossing a coin twice we have the following
sample space: $S = \{TT, TH, HT, HH\}$ and
simple events: $\{TT\}$,$\{TH\}$,$\{HT\}$,$\{HH\}$, $T \equiv$Tail, $H \equiv$Head.
This specification also holds if two coins are tossed once.

It will be convenient to be able to combine events in various ways, in order to make statements such as "one of these two events happened" or "both events occurred." For example, one might want to say that "either a 2 or 4 was thrown,"

W.K. Härdle et al., *Introduction to Statistics*, DOI 10.1007/978-3-319-17704-5_3

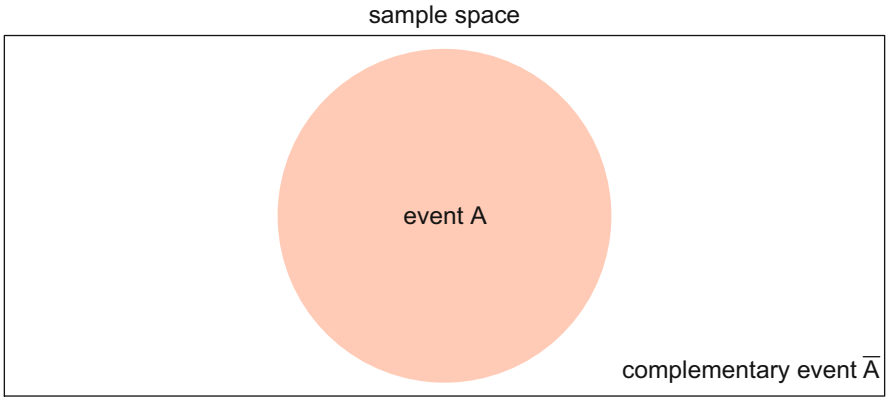


Fig. 3.1 Most simple Venn diagram

or "an even number larger than 3 was thrown." Since events are sets (in particular subsets of the set S), we may draw upon the conventional tools of set theory.

Venn Diagram

A common graphical representation of events as subsets of the sample space is the Venn diagram (Fig. 3.1). It can be used to visualize various combinations of events such as intersections and unions.

3.2 Event Relations and Operations

In the last section, we have defined events as subsets of the sample space S. In interpreting events as sets, we can apply the same operations and relations to events that we know from basic set theory. We shall now recapitulate some of the most important concepts of set theory.

Subsets and Complements

A is subset of B is denoted by $A \subset B$. Thus if event A occurs, B occurs as well (Fig. 3.2).

A and B are equivalent events if and only if (abbreviated as "iff") $A \subset B$ and $B \subset A$. Any event A is obviously a subset of S, $A \subset S$. We define the complement of A, denoted by $\overline{A}$, to be the set of points in S that are not in A.

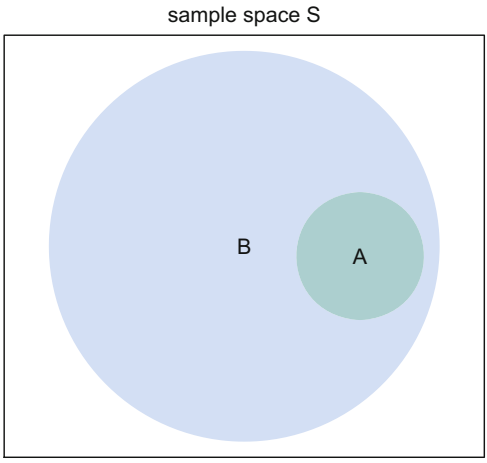


Fig. 3.2 Venn diagram for event relation A is subset of $B, A \subset B$

Union of Sets

The set of points belonging to either the set A or the set B is called the union of sets A and B, and is denoted by $A \cup B$. Thus if the event "A or B" has occurred, then a basic outcome in the set $A \cup B$ has taken place (Fig. 3.3).

Set unions can be extended to n sets and hence n events $A_1, A_2, \ldots, A_n$: in which case we have $A_1 \cup A_2 \cup \ldots \cup A_n = \cup_{i=1}^{n} A_i$

Example Rolling a die once

Define $A = \{1, 2\}$ and $B = \{2, 4, 6\}$.

Then, $A \cup B = \{1, 2, 4, 6\}$.

General Results

- $A \cup A = A$
- $A \cup S = S$ where S is the sample space.
- $A \cup \emptyset = A$ where $\emptyset$ is the null set, the set with no elements in it.
- $A \cup \overline{A} = S$

Intersection of Sets

The set of points common to the sets A *and* B is known as intersection of A and B, $A \cap B$. Thus if the event "A and B" has occurred, then a basic outcome in the set $A \cap B$ has taken place (Fig. 3.4).

Set intersections can be extended to n sets and hence to n events $A_1, A_2, \ldots, A_n$: $A_1 \cap A_2 \cap \ldots \cap A_n = \cap_{i=1}^{n} A_i$

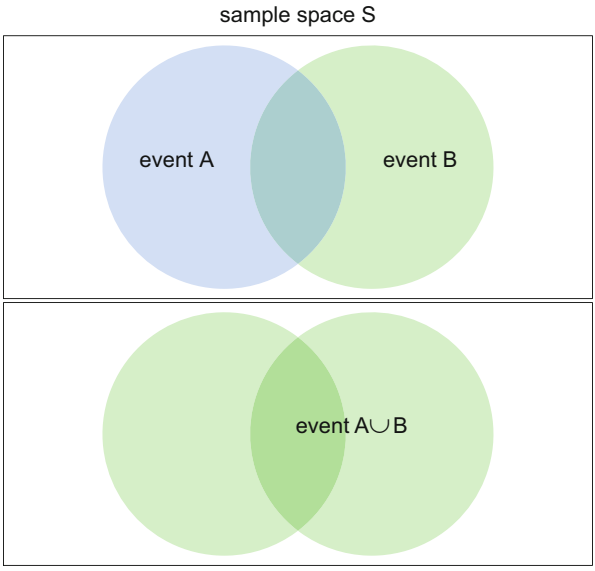


Fig. 3.3 Venn diagram for the union of two sets, $A \cup B$

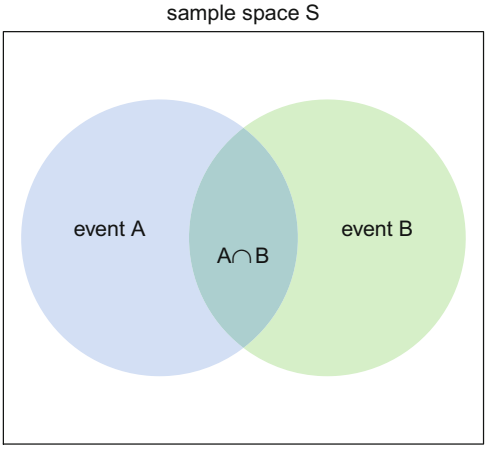


Fig. 3.4 Venn diagram for the intersection of two sets, $A \cap B$

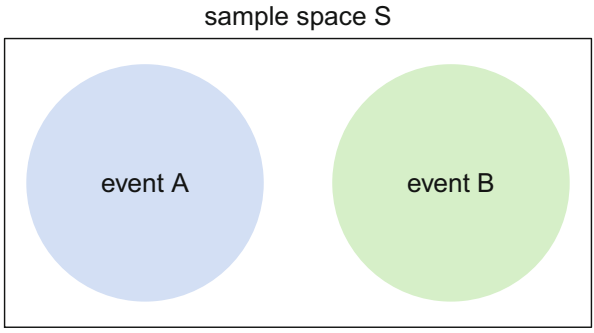


Fig. 3.5 Venn diagram of disjoint events

Example Rolling a die once
Define $A = \{1, 2\}$ and $B = \{2, 4, 6\}$
Then $A \cap B = \{2\}$

General Results

- $A \cap A = A$
- $A \cap S = A$
- $A \cap \emptyset = \emptyset$
- $A \cap \overline{A} = \emptyset$
- $\emptyset \cap S = \emptyset$

Disjoint Events Two sets or events are said to be disjoint (or mutually exclusive) if their intersection is the empty set: $A \cap B = \emptyset$. Interpretation: events A and B cannot occur simultaneously (Fig. 3.5).

By definition, A and $\overline{A}$ are mutually exclusive. The reverse doesn't hold, i.e., disjoint events are not necessarily complements of each other.

Example Rolling a die once
Define $A = \{1, 3, 5\}$ and $B = \{2, 4, 6\}$.
Then, $B = \overline{A}$ and $A = \overline{B}$.
$\Rightarrow A \cap B = A \cap \overline{A} = \emptyset$
Interpretation: events A and B are disjoint and complementary.
Define $C = \{1, 3\}$ and $B = \{2, 4\}$.
$\Rightarrow C \cap D = \emptyset$
Interpretation: events C and D are disjoint but not complementary.

Logical Difference of Sets or Events

The set or event C is the logical difference of events A and B if it represents the event: "A has occurred but B has not occurred," i.e., it is the outcomes in A, that are not in B: $A \backslash B = C \equiv A \cap \overline{B}$ (Fig. 3.6).

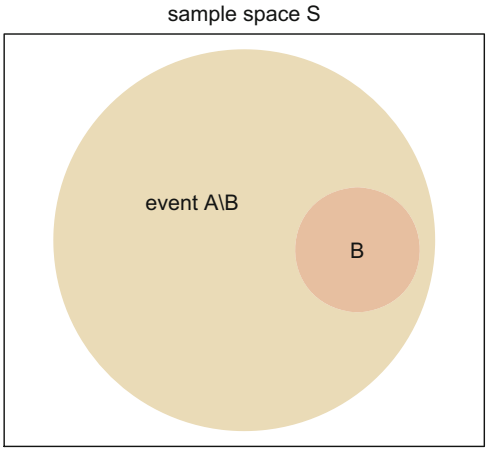


Fig. 3.6 Venn diagram for logical difference of two events, $A \backslash B$

Example Rolling a six-sided die once
Define $A = \{1, 2, 3\}$ and $B = \{3, 4\}$.
Then, $A \backslash B = C = \{1, 2\}$ and $B \backslash A = \{4\}$.

Disjoint Decomposition of the Sample Space

A set of events $A_1, A_2, \ldots, A_n$ is called disjoint decomposition of S, if the following conditions hold:

- $A_i \neq \emptyset \quad (i = 1, 2, \ldots, n)$
- $A_i \cap A_k = \emptyset \quad (i \neq k; i, k = 1, 2, \ldots, n)$
- $A_1 \cup A_2 \cup \ldots \cup A_n = S$

One can think of such a decomposition as a partition of the sample space where each basic outcome falls into exactly one set or event. Sharing a birthday cake results in a disjoint decomposition or partition of the cake.

Example Rolling a six-sided dice

- Sample space: $S = \{1, 2, 3, 4, 5, 6\}$.
- Define $A_1 = \{1\}, A_2 = \{3, 4\}, A_3 = \{1, 3, 4\}, A_4 = \{5, 6\}, A_5 = \{2, 5\}, A_6 = \{6\}$.
- Claim: one possible disjoint decomposition is given by A_1, A_2, A_5, A_6.
- Proof: $A_1 \cap A_2 = \emptyset, A_1 \cap A_5 = \emptyset, A_1 \cap A_6 = \emptyset, A_2 \cap A_5 = \emptyset, A_2 \cap A_6 = \emptyset, A_5 \cap A_6 = \emptyset,$
  $A_1 \cup A_2 \cup A_5 \cup A_6 = S.$

Table 3.1 Summary of event relations

Verbal	Technical	Algebraic
If A occurs, then B occurs also	B is subset of A	$A \subset B$
B and A always occur together	A and B are equivalent events	$A \equiv B$
A and B cannot occur together	A and B are disjoint events	$A \cap B = \emptyset$
A occurs if and only if B does not occur	A and B are complementary events	$B = \overline{A}$
A occurs if and only if at least one A_i occurs	A is union of A_i	$A = \cup_i A_i$
A occurs if and only if all A_i occur	A is intersection of all A_i	$A = \cap_i A_i$

Some Set Theoretic Laws

- De Morgan's laws
  $\overline{A \cap B} = \overline{A} \cup \overline{B}$
  $\overline{A \cup B} = \overline{A} \cap \overline{B}$
- Associative laws
  $(A \cap B) \cap C = A \cap (B \cap C)$
  $(A \cup B) \cup C = A \cup (B \cup C)$
- Commutative laws
  $A \cap B = B \cap A$
  $A \cup B = B \cup A$
- Distributive laws
  $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
  $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

3.3 Probability Concepts

Probability is a measure $P(\bullet)$ which quantifies the degree of (un)certainty associated with an event. We will discuss three common approaches to probability.

Classical Probability

Laplace's classical definition of probability is based on equally likely outcomes. He postulates the following properties of events:

- the sample space is composed of a finite number of basic outcomes
- the random process generates exactly basic outcome and hence one elementary event
- the elementary events are equally likely, i.e., occur with the same probability

Accepting these assumptions, the probability of any event A (subset of the sample space) can be computed as

$$P(A) = \frac{\#\,(\text{basic outcomes in } A)}{\#\,(\text{basic outcomes in } S)} = \frac{\#\,(\text{elementary events comprising } A)}{\#\,(\text{elementary events comprising } S)}$$

Properties

- $0 \leq P(A) \leq 1$
- $P(\emptyset) = 0$
- $P(S) = 1$

Example Rolling a six-sided die
Sample space: $S = \{1, 2, 3, 4, 5, 6\}$.
Define, event $A =$ "any even number"
Elementary events in A: {2},{4},{6}
$P(A) = \frac{3}{6} = 0.5$

Statistical Probability

Richard von Mises originated the relative frequency approach to probability: The probability $P(A)$ for an event A is defined as the limit of the relative frequency of A, i.e., the value the relative frequency will converge to if the experiment is repeated an infinite number of times. It is assumed that replications are independent of each other.

Let $h_n(A)$ denote the absolute frequency of A occurring in n repetitions. The relative frequency of A is then defined as

$$f_n(A) = \frac{h_n(A)}{n}$$

According to the statistical concept of probability we have

$$P(A) = \lim_{n \to \infty} f_n(A)$$

Since $0 \leq f_n(A) \leq 1$ it follows that $0 \leq P(A) \leq 1$.

Example Flipping a coin

Denote by A the event "head appears." Absolute and relative frequencies of A after n trials are listed in Table 3.2. This particular sample displays a non-monotonic convergence to 0.5, the theoretical probability of a head occurring in repeated flips of a "fair" coin.

Visualizing the sequence of relative frequencies $f_n(A)$ as a function of sample size, as done in Fig. 3.7, provides some intuition into the character of the convergence.

A central objective of statistics is to estimate or approximate probabilities of events using observed data. These estimates can then be used to make probabilistic statements about the process generating the data (e.g., confidence intervals which we

Table 3.2 Flipping of a coin

n	$h_n(A)$	$f_n(A)$
10	7	0.700
20	11	0.550
40	17	0.425
60	24	0.400
80	34	0.425
100	47	0.470
200	92	0.460
400	204	0.510
600	348	0.580
800	404	0.505
1000	492	0.492
2000	1010	0.505
3000	1530	0.510
4000	2032	0.508
5000	2515	0.503

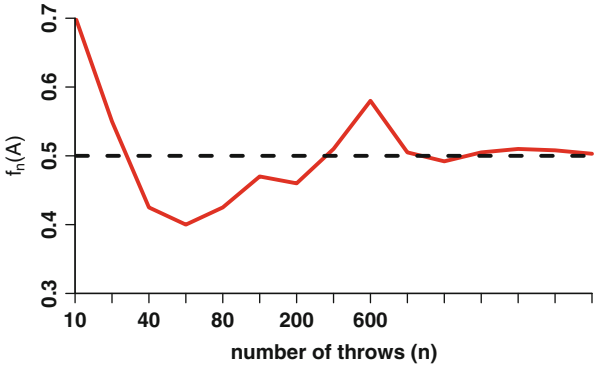


Fig. 3.7 Relative frequencies of A="head appears" as a function of sample size n

will study later), to test propositions about the process and to predict the likelihood of future events.

Axiomatic Foundation of Probability

P is a probability measure. It is a function which assigns a number $P(A)$ to each event A of the sample space S.

- Axiom 1
  $P(A)$ is real-valued with $P(A) \geq 0$.
- Axiom 2
  $P(S) = 1$.
- Axiom 3
  If two events A and B are mutually exclusive ($A \cap B = \emptyset$), then
  $P(A \cup B) = P(A) + P(B)$

Properties

Let $A, B, A_1, A_2, \ldots \subset S$ be events and $P(\bullet)$ a probability measure. Then the following properties follow from the above three axioms:

1. $P(A) \leq 1$
2. $P(\overline{A}) = 1 - P(A)$
3. $P(\emptyset) = 1 - P(S) = 0$
4. $(A \cap B = \emptyset) \Rightarrow P(A \cap B) = P(\emptyset) = 0$
5. If $A \subset B$, then $P(A) \leq P(B)$
6. If $A_i \cap A_j = \emptyset$ for $i \neq j$, then $P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$
7. $P(A \backslash B) = P(A) - P(A \cap B)$

Addition Rule of Probability

Let A and B be any two events (Fig. 3.8). Then,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Extension to three events A, B, C:

$$\begin{aligned} P(A \cup B \cup C) = {} & P(A) + P(B) + P(C) \\ & - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \end{aligned}$$

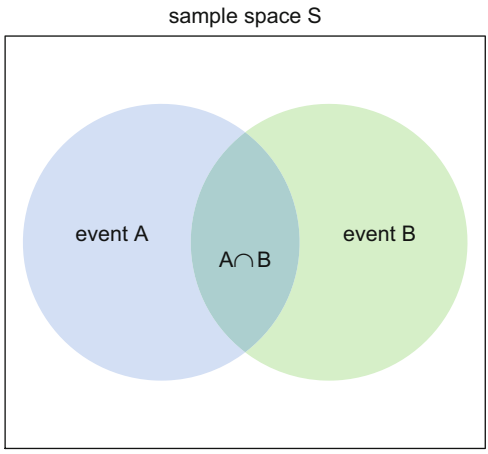


Fig. 3.8 Addition rule of probability

Extension to n events, $A_1, A_2, \cdots, A_n$:

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n}\left((-1)^{k-1} P\left(\bigcap_{I\in\{1,\cdots,n\};|I|=k}^{i\in I} A_i\right)\right)$$

More Information: Derivation of the Addition Rule

1) The event B can be rewritten as a union of two disjoint sets $A \cap B$ and $\bar{A} \cap B$ as follows:

$$B = (A \cap B) \cup (\bar{A} \cap B)$$

as illustrated in the Venn diagram in Fig. 3.9.

The probability $P(B)$ is, according to axiom 3,

$$P(B) = P[(A \cap B) \cup (\bar{A} \cap B)] = P(A \cap B) + P(\bar{A} \cap B)$$

which implies

$$P(\bar{A} \cap B) = P(B) - P(A \cap B)$$

2) We rewrite the event $A \cup B$ as a union of two disjoint sets A and $\bar{A} \cap B$ so that

$$A \cup B = A \cup (\bar{A} \cap B)$$

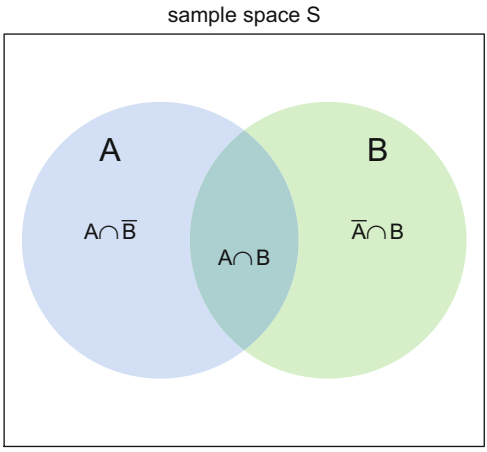


Fig. 3.9 Rewriting a set as the union of two disjoint sets, $B = (A \cap B) \cup (\bar{A} \cap B)$

The probability $P(A \cup B)$ follows from axiom 3

$$P(A \cup B) = P[A \cup (\bar{A} \cap B)] = P(A) + P(\bar{A} \cap B)$$

Now we obtain the desired result by calculating $P(\bar{A} \cap B)$ using the formula given in part one:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

More Information: Implications of the Probability Axioms

Proof of Property 5 Let us show that for $A \subset B$ it follows that $P(A) \leq P(B)$. Then, the event B can be rewritten as $B = A \cup (B \setminus A)$, where A and $B \setminus A$ are disjoint sets. According to axiom 3, we have the following: $P(B) = P(A) + P(B \setminus A)$. Nonnegativity of the probability $P(B \setminus A) \geq 0$ implies that $P(B) \geq P(A)$. This rule can be illustrated using the Venn diagram in Fig. 3.10.

Proof of Property 7 Let us prove that $P(A \setminus B) = P(A) - P(A \cap B)$.

We have $A \setminus B = A \cap \bar{B}$ and $A = (A \cap B) \cup (A \cap \bar{B})$, where $(A \cap B)$ and $(A \cap \bar{B})$ are clearly disjoint.

Using axiom 3 the probability of A can be calculated as

$$P(A) = P[(A \cap B) \cup (A \cap \bar{B})] = P(A \cap B) + P(A \cap \bar{B}) = P(A \cap B) + P(A \setminus B)$$

This result is displayed in Fig. 3.11.

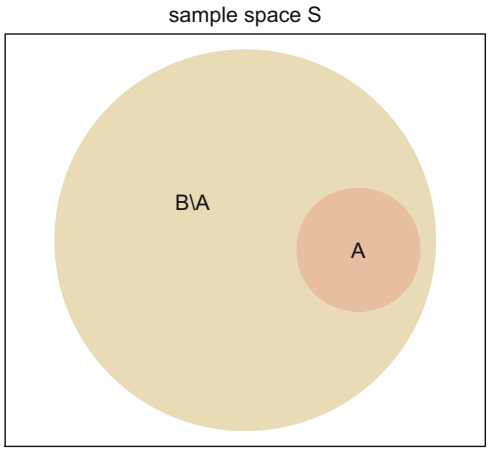


Fig. 3.10 Illustration of set relation $A \subset B$ which implies $P(A) \leq P(B)$

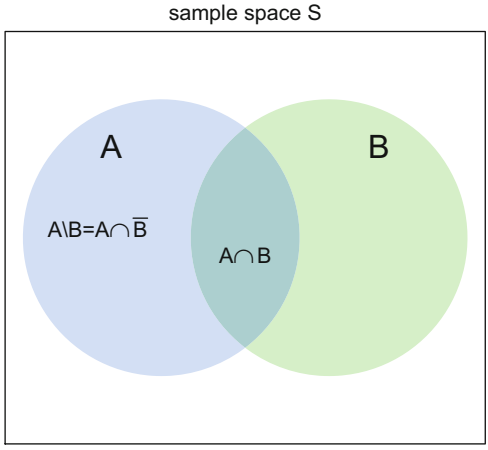


Fig. 3.11 Illustration of $P(A \setminus B) = P(A) - P(A \cap B)$

Explained: A Deck of Cards

Assume you have shuffled a standard deck of 52 playing cards. You are interested in the probability of a randomly drawn card being a queen or a "heart." We are thus interested in the probability of the event ($\{\text{Queen}\} \cup \{\text{Heart}\}$). Following Laplace's notion of probability, we proceed as follows: There are 4 queens and 13 hearts in the deck. Hence,

- $P(\{\text{Queen}\}) = \frac{4}{52}$
- $P(\{\text{Heart}\}) = \frac{13}{52}$

But there is also one card which is both a queen *and* a heart. As this card is included in both counts, we would overstate the probability of either queen or heart appearing if we simply added both probabilities. In fact, the addition rule of

probability requires one to deduct the probability of this joint event:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Here,

$$P(A \cap B) = P(\{\text{Queen}\} \cap \{\text{Heart}\}) = \frac{1}{52}$$

Thus,

$$\begin{aligned} P(\{\text{Queen}\} \cup \{\text{Heart}\}) &= P(\{\text{Queen}\}) + P(\{\text{Heart}\}) \\ &\quad -P(\{\text{Queen}\} \cap \{\text{Heart}\}) \\ &= \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} \end{aligned}$$

The probability of drawing queen's face and/or heart suit is $16/52$.

3.4 Conditional Probability and Independent Events

Conditional Probability

Let A and B be two events defined on the sample space S. The conditional probability of A given B is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ for } P(B) > 0$$

The conditional probability assumes that B has occurred and asks what is the probability that A has occurred. By assuming that B has occurred, we have defined a new sample space $S = B$ and a new probability measure $P(A|B)$.

If $B = A_2 \cap A_3$, then we may write

$$P(A_1|A_2 \cap A_3) = \frac{P(A_1 \cap A_2 \cap A_3)}{P(A_2 \cap A_3)}, \text{ for } P(A_2 \cap A_3) > 0$$

We may also define the conditional probability of B given A:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \text{ for } P(A) > 0$$

Multiplication Rule

By rearranging the definition of conditional probability we can extract a formula for the probability of both A and B occurring:

$$P(A \cap B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$$

and, in analogous fashion,

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2)$$

Generalization for events $A_2, A_2, \ldots A_n$:

$$P(A_1 \cap \ldots \cap A_n = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdot \ldots \cdot P(A_n|A_1 \cdot \ldots \cdot \cdots A_{n-1})$$

Independent Events

The notion underlying the concept of conditional probability is that a priori information concerning the occurrence of events does in general influence probabilities of other events. (For example, if one knows that someone is a smoker, then one would assign a higher probability to that individual contracting lung cancer.) In general, one would expect: $P(A) \neq P(A|B)$.

The case $P(A) = P(A|B)$ has an important interpretation. If the probability of A occurring remains the same, whether or not B has occurred, we say that the two events are statistically (or stochastically) independent. (For example, knowing whether an individual is tall or short does not affect one's assessment of that individual developing lung cancer.)

We define stochastic independence of two events A and B by the condition $P(A \cap B) = P(A) \cdot P(B)$ which implies that the following conditions hold:

$$\begin{aligned} P(A) &= P(A|B) \\ P(B) &= P(B|A) \\ P(A|B) &= P(A|\overline{B}) \\ P(B|A) &= P(B|\overline{A}) \end{aligned}$$

The multiplication condition defining stochastic independence of two events also holds for n independent events:

$$P(A_1 \cap \ldots \cap A_n) = P(A_1) \cdot \ldots \cdot P(A_n)$$

To establish statistical independence of n events, one must ensure that the multiplication rule holds for any subset of the events. That is

$$P\left(A_{i_1} \cap \ldots \cap A_{i_m}\right) = P\left(A_{i_1}\right) \cdot \ldots \cdot P\left(A_{i_m}\right),$$
$$\text{for } i_1, \ldots, i_m \text{ distinct integers} < n$$

It is important not to confuse stochastic independence with mutual exclusivity. For example, if two events A and B with $P(A) > 0$ and $P(B) > 0$ are mutually exclusive, then $P(A \cap B) = 0$, as $P(\emptyset) = 0$ and $A \cap B = \emptyset$. In which case $P(A \cap B) \neq P(A) \cdot P(B)$.

A small example should clarify the difference between independence and mutual exclusivity (rowing Cambridge versus Oxford): click on the symbol of the loudspeaker.

Two-Way Cross-Tabulation

In many applications the researcher is interested in associations between two categorical variables. The simplest case is if one observes two binary variables, i.e., there are two variables, each with two possible outcomes. For example, suppose that for a randomly selected individual we observe whether or not they smoke and whether or not they have emphysema. Let A be the outcome that the individual smokes and B be the outcome that they have emphysema. We can construct separate sample spaces $\left\{A, \overline{A}\right\}$ and $\left\{B, \overline{B}\right\}$.for each of the two variables. Alternatively we can construct the sample space of ordered pairs:

$$S = \left\{(A, B), \left(A, \overline{B}\right), \left(\overline{A}, B\right), \left(\overline{A}, \overline{B}\right)\right\}$$

In tabulating data of this type, we would simply count the number of individuals corresponding to each of the four basic outcomes. No information is lost regarding the two variables individually because we can always obtain frequencies for both categories of either variable by summing over the two categories of the other variable. For example, to calculate the number of individuals who have emphysema, we add up all those who smoke and have emphysema (i.e., (A, B)) and all those who do not smoke and have emphysema (i.e., $(A, \overline{B})$). Relative frequencies for categories of the individual variables are called marginal relative frequencies.

Relative frequencies arising from bivariate categorical data are usually displayed by cross-tabulating the categories of the two variables. Marginal frequencies are included as sums of the columns/rows representing the categories of each of the variables. The resulting matrix is called an $(r \times c)$-contingency table, where r and c denote the number of categories observed for each variable. In our example with two categories for each variable, we have a (2×2)-contingency table.

We may summarize the probabilities associated with each basic outcome in a similar way shown in Table 3.3.

Table 3.3 (2×2)-table of events $A, \overline{A}, B$ and $\overline{B}$

	B	$\overline{B}$	Sum
A	$P(A \cap B)$	$P(A \cap \overline{B})$	$P(A)$
$\overline{A}$	$P(\overline{A} \cap B)$	$P(\overline{A} \cap \overline{B})$	$P(\overline{A})$
Sum	$P(B)$	$P(\overline{B})$	$P(S) = 1$

The structure of this table is particularly helpful in checking for independence between events. Recall that the joint probability of two independent events can be calculated as the product of the probabilities of the two individual events. In this case, we want to verify whether the joint probabilities in the main body of the table are equal to the products of the marginal probabilities. If they are not, then the events are not independent. For example, under independence, we would have $P(A)\,P(B) = P(A \cap B)$.

If one replaces the probabilities in Table 3.3 with their sample frequencies, then independence implies that the estimated joint probabilities should be approximately equal to the products of the estimated marginal probabilities. Formal procedures for testing independence will be discussed later.

More Information: Derivation of Rules for Independent Events

We want to prove the following proposition: For any pair of independent events A and B we have $P(A) = P(A|B)$.

Assume that the events A and B are independent. Then we have

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

Similarly, we can show that $P(B|A) = P(B)$. Next, suppose that $P(A) = P(A|B)$ we want to show that this implies the multiplicative rule, i.e., that A and B are independent:

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

Indeed stochastic independence can be defined equivalently in a number of ways.

Explained: Two-Way Cross-Tabulation

Joint probabilities of two binary variables are arranged in the contingency table below. Are the variables represented by the events $\{A, \overline{A}\}$ respectively $\{B, \overline{B}\}$ (mutually) independent?

For the multiplication condition of independence to be satisfied, the inner cells of the contingency table must equal the product of their corresponding marginal probabilities. This is true for all four cells shown in Table 3.4.

In this very special example with two binary variables it is, however, not necessary to verify the validity of the multiplication rule for each of the four cells. As we have already seen, stochastic independence of two events implies stochastic independence of the complementary. Consequently, if the multiplication condition holds for one of the four cells, it must hold for the other three. This is only true because the only two events to be considered for each variable are complements.

Explained: Screws

A master and his apprentice produce hand-made screws. The data were collected over the course of the year 1998 and are provided in Table 3.5.

What is the probability, that a randomly selected screw is not faulty given that it was produced by the master? In order to calculate this probability, we will use this notation:

$$A = \{\text{screw is good}\}$$
$$B = \{\text{screw produced by master}\}$$
$$C = \{\text{screw produced by apprentice}\}$$

Table 3.4 (2×2)-table of joint probabilities of two independent binary variables

	B	$\overline{B}$	Sum
A	$1/3$	$1/6$	$1/2$
$\overline{A}$	$1/3$	$1/6$	$1/2$
Sum	$2/3$	$1/3$	1

	B	$\overline{B}$	Sum
A	$1/3 = 1/2 \cdot 2/3$	$1/6 = 1/2 \cdot 1/3$	$1/2$
$\overline{A}$	$1/3 = 1/2 \cdot 2/3$	$1/6 = 1/2 \cdot 1/3$	$1/2$
Sum	$2/3$	$1/3$	1

Table 3.5 Production of hand-made screws by a master and his apprentice

Total production	2000	screws
Group 1 (the master)	1400	Screws
	1162	Good screws
	238	Faulty screws
Group 2 (the apprentice)	600	Screws
	378	Good screws
	222	Faulty screws

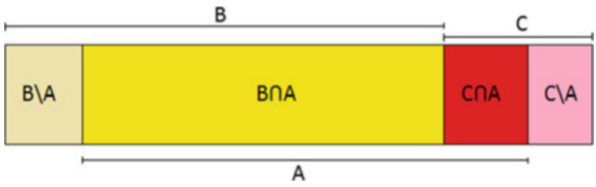


Fig. 3.12 Production of hand-made screws by a master and his apprentice

The situation is displayed in the Venn diagram in Fig. 3.12.

We would like to calculate $P(A|B)$. This conditional probability is defined as $P(A|B) = P(A \cap B)/P(B)$. The event $A \cap B$ corresponds to selection of a good screw produced by the master. In order to calculate $P(A \cap B)$, we divide the number of screws with this property by the total number of screws: $P(A \cap B) = 1162/2000$.

The probability $P(B)$ can be calculated as a ratio of the number of screws produced by the master and total production: $P(B) = 1400/2000$. Thus, we obtain:

$$P(A|B) = 1162/1400 = 0.83\,.$$

3.5 Theorem of Total Probabilities and Bayes' Rule

Recall the disjoint decomposition we have introduced earlier in this chapter as a set of events $A_1, A_2, \ldots, A_n$ satisfying

- $A_i \neq \emptyset \quad (i = 1, 2, \ldots, n)$
- $A_i \cap A_k = \emptyset \quad (i \neq k; i, k = 1, 2, \ldots, n)$
- $A_1 \cup A_2 \cup \ldots \cup A_n = S$

Theorem of Total Probabilities

$A_1, A_2, \ldots, A_n$ be a disjoint decomposition. Then, for any event $B \subset S$ with $P(B) > 0$:

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

We have applied the multiplication rule of probability $P(A \cap B) = P(B|A_i)\ P(A_i)$.

Bayes' Rule

Let $A_1, A_2, \ldots, A_n$ be a disjoint decomposition. Then, for any event $B \subset S$ with $P(B) > 0$ and given conditional probabilities $P(B|A_1)$, $P(B|A_2)$, …, $P(B|A_n)$:

$$P\left(A_j|B\right) = \frac{P\left(B|A_j\right) P\left(A_j\right)}{\sum_{i=1}^{n} P\left(B|A_i\right) P\left(A_i\right)} \quad \forall j = 1, \ldots, n$$

The *Bayesian approach* to statistics interprets the $P\left(A_j|B\right)$ as *posterior probabilities* and $P\left(A_i\right)$ as *prior probabilities*. This conceptual approach to statistics accounts for prior information in the form of subjective belief rather than defining probabilities as limits of relative frequencies.

Explained: The Wine Cellar

In this example we will apply both the theorem of total probabilities and Bayes' rule.

Wolfram has a wine cellar. Having invited guests for a dinner party, he considers showing off in the most economical fashion. He knows that his guests usually buy their wine at the local supermarket. So he decides to provide above average food and not to spend too much time choosing the accompanying wine. His stock currently consists of Qualitätswein, Kabinett, and Spätlese in the proportions $5 : 3 : 2$. The proportion of white wine in these classes is $1/5$, $1/3$, and $1/4$, respectively.

Being a technocrat not only in pedantically monitoring his stock, he wants to compute the probability for producing a bottle of white wine when randomly picking one. He estimates probabilities by their relative proportions in the stock population:

$A_1 \equiv \{\text{Qualitätswein}\}$ $P(A_1) = 0.5$
$A_2 \equiv \{\text{Kabinett}\}$ $P(A_2) = 0.3$
$A_3 \equiv \{\text{Spätlese}\}$ $P(A_3) = 0.2$

This classification establishes a disjoint decomposition of Wolfram's wine stock:

$$A_1 \cup A_2 \cup A_3 = S$$
$$A_1 \cap A_2 = \emptyset,\ A_1 \cap A_3 = \emptyset,\ A_2 \cap A_3 = \emptyset.$$

Let B represent the event of picking a bottle of white wine. Then we know:

$$P(B|A_1) = 1/5$$
$$P(B|A_2) = 1/3$$
$$P(B|A_3) = 1/4$$

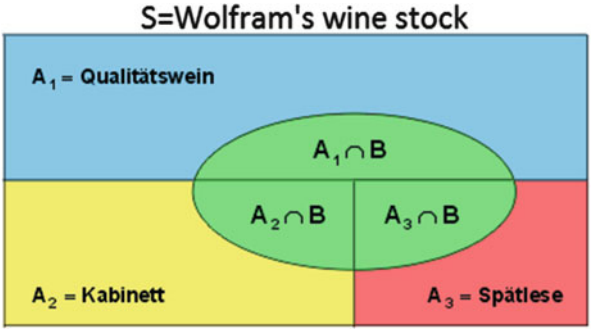


Fig. 3.13 Wolfram's wine cellar containing quantities of three different wines

Being short of time, Wolfram decides to have the food delivered from a gourmet deli. Now he has spare time to draw a Venn diagram as shown in Fig. 3.13.

As A_1, A_2, and A_3 establish a disjoint decomposition, $A_1 \cap B$, $A_2 \cap B$, and $A_3 \cap B$ must be disjoint as well. Thus, for $B = (A_1 \cap B) \cup (A_2 \cap B) \cup (A_3 \cap B)$

$$\begin{aligned} P(B) &= P[(A_1 \cap B) \cup (A_2 \cap B) \cup (A_3 \cap B)] \\ &= P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B) \end{aligned}$$

As he doesn't know the probabilities for the union sets on the right-hand side, Wolfram applies the multiplication rule, substituting $P(B|A_i)\,P(A_i)$ for $P(A_i \cap B)$:

$$\begin{aligned} P(B) &= P(B|A_1)\,P(A_1) + P(B|A_2)\,P(A_2) + P(B|A_3)\,P(A_3) \\ &= 1/5 \cdot 0.5 + 1/3 \cdot 0.3 + 1/4 \cdot 0.2 = 0.25 \end{aligned}$$

Thus randomly selecting a bottle will result in a white wine with a 25 % probability.

Given that Wolfram has selected a bottle of white wine, what is the probability that it is Qualitätswein, that is, what is $P(A_1|B)$?

Wolfram wants to apply the definition for conditional probability,

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)}$$

He has already calculated $P(B)$ using the theorem of total probability. But what about the numerator on the right-hand side? Wolfram chooses to rearrange the definition for the conditional probability of B, given A_1 to yield a multiplication rule he can substitute into the numerator:

$$\begin{aligned} P(B|A_1) &= \frac{P(A_1 \cap B)}{P(A_1)} \\ \Leftrightarrow P(A_1 \cap B) &= P(B|A_1)\,P(A_1) \end{aligned}$$

This yields

$$\begin{aligned} P(A_1|B) &= \frac{P(A_1 \cap B)}{P(B)} \\ &= \frac{P(B|A_1)\,P(A_1)}{P(B|A_1)\,P(A_1) + P(B|A_2)\,P(A_2) + P(B|A_3)\,P(A_3)} \\ &= \frac{P(B|A_1)\,P(A_1)}{\sum_{i=1}^{3} P(B|A_i)\,P(A_i)} \\ &= \frac{0.2 \cdot 0.5}{0.25} = 0.4 \end{aligned}$$

Enhanced: Virus Test

Assume 0.5 % of the population is infected with a particular virus that leads to acute disease only after a long period of time.

A clinical study shows that 99 % of the individuals suffering from the symptoms that confirm an infection with the virus test positive. On the other hand, 2 % of people not developing the symptoms test positive as well.

What is the probability that a person testing positive has the infection?

Let us first formalize the problem. Instead of using the set theoretic notation we will now define indicator variables for the two binary variables corresponding to the infection (I) and the test (T):

$$I = \begin{cases} 1 \text{ if a person is infected} \\ 0 \text{ if a person is not infected} \end{cases}$$

$$T = \begin{cases} 1 \text{ if the test is positive} \\ 0 \text{ if the test is not positive} \end{cases}$$

Using the above we know the following probabilities.

$$\begin{aligned} P(I = 1) &= 0.005 \\ P(T = 1|I = 1) &= 0.99 \\ P(T = 1|I = 0) &= 0.02 \end{aligned}$$

We would like to calculate $P(I = 1|T = 1)$. The definition of conditional probability contains probabilities which not readily available:

$$P(I = 1|T = 1) = \frac{P[(I = 1) \cap (T = 1)]}{P(T = 1)}, \text{ for } P(T = 1) > 0$$

To replace the numerator by a known quantity we rearrange

$$P(T=1|I=1) = \frac{P[(I=1)\cap(T=1)]}{P(I=1)}, \text{ for } P(I=1) > 0$$

to yield

$$P[(I=1)\cap(T=1)] = P(T=1|I=1)\,P(I=1)$$

The denominator can be calculated using the theorem of total probabilities:

$$P(T=1) = P(I=1|T=1)\,P(I=1) + P(T=1|I=1)\,P(I=0)\,.$$

We thus get

$$P(I=1|T=1) = \frac{P(T=1|I=1)\,P(I=1)}{P(I=1|T=1)\,P(I=1) + P(T=1|I=1)\,P(I=0)}.$$

Performing the calculation we obtain a somewhat surprising result:

$$P(I=1|T=1) = \frac{0.99\cdot 0.005}{0.99\cdot 0.005 + 0.02\cdot 0.995} = 0.199.$$

Thus a randomly selected person who tests positive has an 80 % chance of not being infected. But don't forget about one crucial assumption we have made: the proportion of infected people has to be the same in the population and the sample of tested persons. This may be true for large scale clinical tests. But in practice, there is usually a reason for testing a person, e.g., him/her having been exposed to an infected person.

Interactive: Monty Hall Problem

The Monty Hall problem (named after Monty Hall, television host of the show "Let's make a deal") is based on the following situation:

Monty Hall shows his guest three doors A, B, and C. The main prize is hidden behind one of them, other doors conceal smaller prizes. For now, let us assume that the main prize is behind the door B.

Monty Hall asks the player to choose one door. After the player chooses (let us say door A), one of the doors which does not contain the main prize is opened (let us say door C). The player can now decide whether to continue with his original choice (door A) or if he wants to choose the other closed door (door B). What is the probability that the main prize is behind the originally selected door (A) or behind

the other (unopened and not selected) door (B)? To answer the question, let's have a look at the interactive example.

The Interactive Example

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please select

- the door the guest points to
- if the guest decides to keep or change the door

Use

- "Make a deal" to manually play the game with "virtual Monty"
- the slider to cause an automated playing of the game

Output

The resulting graphic in Fig. 3.14 allows you to study the relative frequency of winning the game depending on your strategy. The statistical definition of probability ensures that your question will be answered after a sufficient number of games.

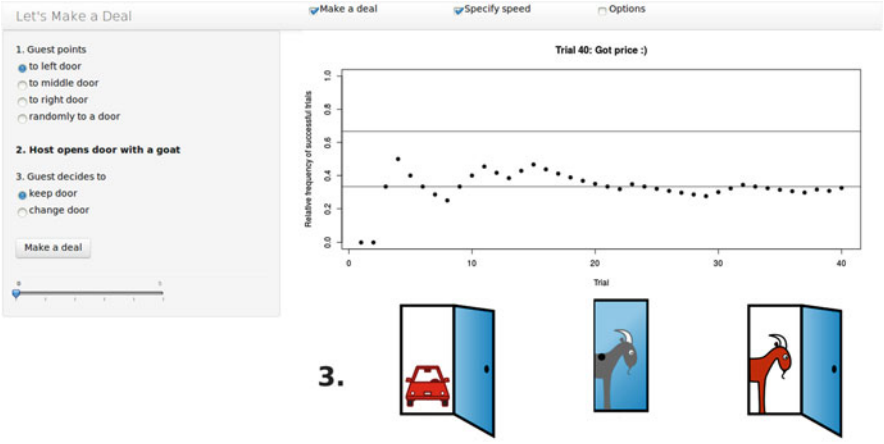


Fig. 3.14 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_hall

Analytical Solution

Let us define the events

$$A: \text{ Main is price behind door A}$$
$$B: \text{ Main is price behind door B}$$
$$C: \text{ Main is price behind door C}$$
$$a: \text{ Monty opens door A}$$
$$b: \text{ Monty opens door B}$$
$$c: \text{ Monty opens door C}$$

Initially, the probability is $1/3$ that you have selected the winning door.

$$P(A) = P(B) = P(C) = 1/3$$

These probabilities are valid **before** Monty opens a door; we can denote them as the a priori probabilities. Let us say that you choose door A. Monty now opens one of the other doors which does not contain the main price. We distinguish two situations:

- Situation 1
  If the prize is behind your door (A), then Monty can open either of the remaining two doors (door B or C). Let us assume that his decision is random—this means that both door have probability 1/2.
- Situation 2
  If the prices **is not** behind your door, then it has to be behind door B or C and Monty has to open (i.e., he will open with probability 1) the other one.

Let us assume that Monty opens door B. Mathematically, this means

$$\text{Situation 1: } P(b|A) = \frac{1}{2}$$
$$\text{Situation 2: } P(b|C) = 1$$

As a player, you do not know which situation has occurred.

When Monty opens the door, you can stick to your original decision or you can change it and open door C. Which decision is better, i.e., which of the doors A or C are more likely to conceal the main prize, if we know that Monty has opened door B?

We would like to calculate the probabilities $P(A|b)$ and $P(C|b)$. The a priori probabilities were $P(A) = P(C) = \frac{1}{3}$. When Monty opens door B, we can calculate the a posteriori probabilities by applying the Bayes rule and the Total Probabilities Theorem:

$$P(A|b) = \frac{P(b|A) \cdot P(A)}{P(b)} = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}$$

$$P(C|b) = \frac{P(b|C) \cdot P(C)}{P(b)} = \frac{1 \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

Changing your decision pays off!

Interactive: Die Rolling Sisters

Three siblings are playing dice. The youngest one (a boy) gave one die to each of his sisters. They roll the die n times and the one who obtains six the most frequently wins.

The sisters remember that one of the dice is "loaded." The probability of obtaining six with this die is $1/3$, the probability of other numbers is uniform at $2/15$.

The first sister rolled the die n times and she has X sixes. The other sister wants to calculate the probability that her die is loaded. This can be done easily.

Let us look at the actual number of sixes which can be $0, 1, 2, \ldots,$ or n. For simplicity, suppose $n = 3$. For a fair die we will write $W = 0$, for a loaded die, $W = 1$. All throws are mutually independent and therefore we obtain:

$$\begin{aligned}
P(X = 0|W = 0) &= P(\text{no 6 in the three throws}) \\
&= 5/6 \cdot 5/6 \cdot 5/6 = 0.5787 \\
P(X = 1|W = 0) &= P(\text{just 1 six in the three throws}) \\
&= 1/6 \cdot 5/6 \cdot 5/6 \cdot 3 = 0.3472 \\
P(X = 2|W = 0) &= P(\text{exactly 2 sixes in the three throws}) \\
&= 1/6 \cdot 1/6 \cdot 5/6 \cdot 3 = 0.0694
\end{aligned}$$

$$P(X = 3|W = 0) = P(\text{all three throws give 6})$$
$$= 1/6 \cdot 1/6 \cdot 1/6 = 0.0046$$

For the same experiment with the loaded die ($W = 1$) we obtain:

$$P(X = 0|W = 1) = 2/3 \cdot 2/3 \cdot 2/3 = 0.2963$$
$$P(X = 1|W = 1) = 1/3 \cdot 2/3 \cdot 2/3 \cdot 3 = 0.4444$$
$$P(X = 2|W = 1) = 1/3 \cdot 1/3 \cdot 2/3 \cdot 3 = 0.2222$$
$$P(X = 3|W = 1) = 1/3 \cdot 1/3 \cdot 1/3 = 0.0370$$

Let us say that the first sister obtains two sixes from her three throws ($X = 2$). What is the probability that she played with the loaded die?

We want to calculate the probability $P(W = 1|X = 2)$. According to the Bayes rule we have

$$P(W = 1|X = 2) = \frac{P(X = 2|W = 1)P(W = 1)}{P(X = 2|W = 0)P(W = 0) + P(X = 2|W = 1)P(W = 1)}$$

Using $P(W = 1) = P(W = 0) = 1/2$ leads in the numerator to $0.2222 \cdot 0.5 = 0.1111$ and in the denominator $0.0694 \cdot 0.5 + 0.2222 \cdot 0.1458$ so that the probability $P(W = 1|X = 2) = 0.1111/0.1458 = 0.762$.

The Interactive Example

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please select

- the number of throws n
- the number of sixes X
- the probability of a six when the die is loaded

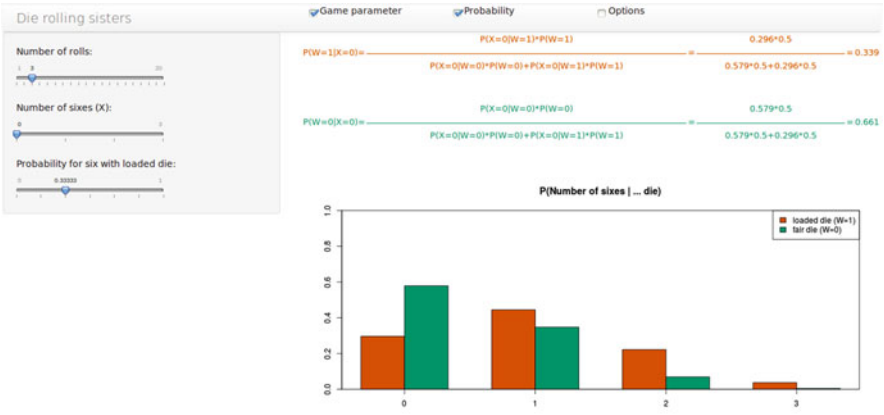


Fig. 3.15 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_die

Output

You receive a graphic as in Fig. 3.15 which displays the probabilities $P(W = 1|X)$ and $P(W = 0|X)$. By changing the parameters you may study the influence of your changes on the resulting probabilities.

Chapter 4
Combinatorics

4.1 Introduction

Combinatorial theory investigates possible patterns of orderings of finitely many elements, composed groups (sets) of such orderings, and the number of these orderings and groups.

Different Ways of Grouping and Ordering

Groups of elements can differ in several ways: they can contain either all elements just once, or some elements several times and others not at all; moreover, two groups that contain the same set of elements and differ from each other just by the ordering of the elements can be considered to be the same or not.

Examples with three elements a, b, and c:

- A group that contains every element exactly once: *b c a*
- A group that contains some elements more times and other elements not at all: *b b*
- Two groups that differ from each other just by the orderings of their elements: *a b* and *b a*

As you can see on this simple example, groupings of elements can form three basic types:

- **Permutation**
- **Variation**
- **Combination**

W.K. Härdle et al., *Introduction to Statistics*, DOI 10.1007/978-3-319-17704-5_4

Use of Combinatorial Theory

Combinatorial theory mainly helps to answer questions such as:

- How many different ways can 5 different digits be ordered?
- How many ways exists for a choice of 10 words out of 30?
- How many possibilities are available for filling a lottery coupon?

Answers to these questions make possible to determine, for example, the probability of winning a lottery prize. Therefore, the use of combinatorial theory is most relevant in probability theory, which, on the other hand, really use the results of combinatorial theory.

4.2 Permutation

Every group of certain n elements that contains all n elements is called permutation of these elements. It follows that different permutations of the same set of elements differ from each other just by orderings of the elements.

There are three kinds of permutations:

Permutations Without Repetition

Permutations without repetition are such permutations in which every element is contained just once, and thus, all n elements are different. The number of permutations without repetition, which is denoted from now on by $P(n)$, is:

$$P(n) = 1 \cdot 2 \cdot 3 \cdot \cdots \cdot n = n\,!$$

Examples

- For two different elements (a and b) the number of possible permutations is $P(2) = 1 \cdot 2 = 2\,! = 2$. Quite naturally, the two possible permutations are: $\{a, b\}$ and $\{b, a\}$.
- For three different elements (a, b, and c) the number of possible permutations is $P(3) = 1 \cdot 2 \cdot 3 = 3\,! = 6$. All six possible permutations are as follows: $\{a, b, c\}$; $\{a, c, b\}$; $\{b, a, c\}$; $\{b, c, a\}$; $\{c, a, b\}$; and $\{c, b, a\}$.

Permutations with Repetition

This kind of a permutation allows for ordered groups of elements in which some elements are the same. Let us assume that there are g identical elements among all n elements in a permutation. The number of possible permutations with repetition of n elements is denoted $P(n; g)$ and it can be determined by the following formula:

$$P(n; g) = \frac{n\,!}{g\,!} \qquad g \leq n\,,$$

where g is the number of the identical elements (the size of their group).

Examples

- First, consider once again the case of two elements: For two different elements (a and b) the number of possible permutations is $P(2) = 2\,!/1\,! = 2$ (which is the same as for permutations without repetition): $\{a, b\}$ and $\{b, a\}$.

  For two identical elements (a and a) the number of possible permutations is $P(2; 2) = 2\,!/2\,! = 1$. The only possible permutation is of course: $\{a, a\}$.
- For a group of three elements, it is possible to have $g = 1$ (the same as for permutations without repetition), $g = 2$ (two elements are the same, while the third is different), or $g = 3$ (all three elements are the same):

  For $g = 1$ the group is formed by (a, b, and c), so the number of possible permutations is $P(3) = 3\,!/1\,! = 6.$, as before. All six permutations are: $\{a, b, c\}$; $\{a, c, b\}$; $\{b, a, c\}$; $\{b, c, a\}$; $\{c, a, b\}$; and $\{c, b, a\}$.

  For $g = 2$ (a,a, and b), the number of possible permutations is $P(3; 2) = 3\,!/2\,! = 3$. Three are: $\{a, a, b\}$; $\{a, b, a\}$; and $\{b, a, a\}$.

  For $g = 3$ (a, a, and a), the number of possible permutations is $P(3; 3) = 3\,!/3\,! = 1$ and the only possible permutations is: $\{a, a, a\}$.

Apparently, permutations without repetition are a special case of permutations with repetition. Permutations with repetition are then a special case of permutations with more groups of identical elements.

Permutations with More Groups of Identical Elements

For permutations of this kind, it is possible that there are more (different) groups of identical elements among all n elements of a permutation. For r such groups, the number of permutations is

$$P(n; g_1, \ldots, g_r) = \frac{n\,!}{g_1\,! \cdot g_2\,! \cdot \cdots \cdot g_r\,!}$$

where g_i represents the size of ith group and it holds that $g_1 + g_2 + g_3 + \cdots + g_r \leq n$.

Explained: Beauty Competition

There are 14 competitors in a beauty competition. Every juror in the jury has to create his own private ranking of these 14 competitors. How many jurors are needed in order to get all possible rankings (every one different from all others) of all 14 competitors if we assume that every juror has different tastes?

To create a ranking, it is just necessary to order all n elements (14 competitors); thus, every juror has to create a **permutation**.

Now, one has to decide whether we deal with permutations with or without repetitions. As every competitor can be in a ranking from a juror included just once, we consider permutations without repetition.

$$P(n) = 1 \cdot 2 \cdot 3 \cdot \cdots \cdot n = n!$$

$$P(14) = 1 \cdot 2 \cdot 3 \cdot \cdots \cdot 14 = 14! = 87,178,291,200$$

For all the possible rankings, more than 87 billion of jurors is needed. Clearly, it would not be easy to find them since the total population of the Earth is currently about 6 billion.

4.3 Variations

Every group of k elements chosen from a set of n elements in which the **ordering of elements matters** is called a variation of the kth order of n elements.

Variations with Repetition

A variation with repetition is a variation in which every element can be present more than once. The number of possible variations of the kth order of n elements with repetition is denoted by $V^W(n;k)$ and it can be computed as

$$V^W(n;k) = n^k$$

Examples with elements a, b, and c ($n = 3$):

- For $k = 1$ we have $V^W(3;1) = 3^1 = 3$. The three possible variations are: $\{a\}$; $\{b\}$, and $\{c\}$.
- For $k = 2$ we obtain $V^W(3;2) = 3^2 = 9$ and the variations are: $\{a,a\}$; $\{b,b\}$; $\{c,c\}$; $\{a,b\}$; $\{b,a\}$; $\{a,c\}$; $\{c,a\}$; $\{b,c\}$; and $\{c,b\}$.

- For $k = 3$ we have $V^W(3;3) = 3^3 = 27$. The variations are: $\{a,a,a\}$; $\{b,b,b\}$; $\{c,c,c\}$; $\{a,a,b\}$; $\{a,b,a\}$; $\{b,a,a\}$; $\{a,a,c\}$; $\{a,c,a\}$; $\{c,a,a\}$; $\{b,b,a\}$; $\{b,a,b\}$; $\{a,b,b\}$; $\{b,b,c\}$; $\{b,c,b\}$; $\{c,b,b\}$; $\{c,c,a\}$; $\{c,a,c\}$; $\{a,c,c\}$; $\{c,c,b\}$; $\{c,b,c\}$; $\{b,c,c\}$; $\{a,b,c\}$; $\{a,c,b\}$; $\{b,a,c\}$; $\{b,c,a\}$; $\{c,a,b\}$; and $\{c,b,a\}$.

Variations Without Repetition

In this type of variations, every element (from the set of all n elements) can be included at most once. The number of possible variations of the kth order of n elements without repetition is denoted $V(n;k)$ and equals

$$V(n;k) = n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot (n-k+2) \cdot (n-k+1) = \frac{n!}{(n-k)!}.$$

Examples with elements a, b, and c ($n = 3$)

- For $k = 1$ we have $V(3;1) = 3!/2! = 3$. The three possible variations are obviously: $\{a\}$; $\{b\}$; and $\{c\}$.
- For $k = 2$ we obtain $V(3;2) = 3!/1! = 6$ and the variations are: $\{a,b\}$; $\{b,a\}$; $\{a,c\}$; $\{c,a\}$; $\{b,c\}$; and $\{c,b\}$.
- For $k = 3$ we obtain $V(3;3) = 3!/0! = 6$. The six possible variations are: $\{a,b,c\}$; $\{a,c,b\}$; $\{b,a,c\}$; $\{b,c,a\}$; $\{c,a,b\}$; and $\{c,b,a\}$.

Explained: Lock Picking

Everybody knows it—**a briefcase with a several-digit-code lock**. Imagine you changed that code some time ago and forgot it a few days later. Now, the question is how many numbers (i.e., series of digits) you have to try to open the briefcase in the worst case scenario.

In most cases, such a lock has three digits to choose and every digit can be any number between 0 and 9. Therefore, it is a choice of 3 digits (k elements) from 10 (all n elements). It is clear that ordering of digits matters here—the sequence 462 will have a different effect (it can even open the briefcase) than the sequences 264 or 426. These two pieces of information (k elements out of total n elements and importance of ordering) indicate that **variations** are the right concept to use for this problem. (Please be aware of the fact that in this connection often used term "numerical combination" is factually wrong, it is numerical variation!)

Also, it is necessary to decide whether we deal with variations with or without repetition: every digit can equal to any value between 0 and 9, so, for example, the sequence 666 is possible. The correct concept is therefore **variations with repetition**.

$$k = 3 \qquad n = 10 \text{ (0 to 9)}$$

$$V^W = (n, k) = n^k$$

$$V^W = (10, 3) = 10^3 = 10 \cdot 10 \cdot 10 = 1000$$

Three-digit numerical lock (with digits 0–9) makes 1000 variations possible. If one tries to open the lock and it takes 2 s on average, in the worst case scenario, it will take 33.33 mins to open the briefcase.

4.4 Combinations

Every group of k elements chosen from a set of n elements in which the **ordering of the chosen elements is unimportant** is called a combination of the kth order of n elements.

Combinations Without Repetition

Ordering of elements does not play any role when the number of combinations is to be determined (i.e., groups *ab* and *ba* are equivalent combinations). That is why the number of combinations of the kth order is lower than the number of variations of the kth order from the same set of n elements. The number of variations, which differ from each other just by ordering of their elements, is given by $P(k)$. Hence, the number of combinations of the kth order of n elements without repetition (it is denoted here by $K(n; k)$) is:

$$K(n; k) = \frac{V(n; k)}{P(k)} = \frac{n\,!}{k\,! \cdot (n-k)\,!} = \binom{n}{k}$$

Examples with elements a, b, and c ($n = 3$)

- For $k = 1$ we have $K(3; 1) = 3$ and these three possibilities are: $\{a\}$; $\{b\}$; and $\{c\}$.
- For $k = 2$ we have $K(3; 2) = V(3; 2)/P(2) = 6/2 = 3$: $\{a, b\}$; $\{a, c\}$; and $\{c, b\}$.
- For $k = 3$ we have $K(3; 3) = V(3; 3)/P(3) = 3/3 = 1$, so there is just one combination: $\{a, b, c\}$.

Combinations with Repetition

Combinations with repetition can include one element many times; hence, the maximal possible number of combinations of the kth order of n elements with repetition (denoted by $K^W(n;k)$) is

$$K^W(n;k) = \binom{n+k-1}{k}$$

Examples with elements a, b, and c ($n = 3$)

- For $k = 1$ we have $K^W(3;1) = 3$ and the three possibilities are: $\{a\}$; $\{b\}$; and $\{c\}$.
- For $k = 2$ is $K^W(3;2) = 6$: $\{a,b\}$; $\{a,c\}$; $\{c,b\}$; $\{a,a\}$; $\{b,b\}$; and $\{c,c\}$.

Explained: German Lotto

Millions of Germans try every Saturday their luck in the lottery called Lotto. They choose 6 numbers from 49 and hope that, thanks to these 6 numbers, they will get rich. They base the choice often on various almost "mystical" numbers—numbers such as the date of somebody's birthday, the birthday of their dog, numbers hinted by a horoscope, and so on. How many possibilities for a choice of 6 numbers out of 49 actually exists?

From 49 numbers (elements), exactly 6 is chosen. The order in which the numbers are chosen in unimportant—it does not matter whether one crosses first 4 and then 23 or vice versa. That means that ordering of elements is not taken into consideration. Therefore, permutations (simple reordering of n elements) and variations as well (ordering of elements matters) are not the right choice. The right concept is a **combination**.

Nevertheless, there are still two possibilities—combinations with or without repetition. Since every number of the lottery ticket can be crossed just once, repetition of numbers (elements) is not possible and we use **combinations without repetition**.

$$n = 49 \qquad k = 6$$

$$K(n,k) = \binom{n}{k} = \frac{V(n,k)}{P(k)} = \frac{n\,!}{k\,! \cdot (n-k)\,!}$$

$$K(n,k) = \frac{49\,!}{6\,! \cdot (49-6)\,!} = 13\,983\,816$$

There is 13983816 possible combinations of 6 numbers from 49.

4.5 Properties of Euler's Numbers (Combination Numbers)

Euler's symbol $\binom{n}{k}$, read n over k, is used in combinatorial theory very often. So it is useful to know several important properties of these so-called combination numbers.

Symmetry

$$\binom{n}{k} = \binom{n}{n-k}$$

Proof of symmetry

$$\frac{n!}{k!(n-k)!} = \frac{n!}{(n-k)!(n-(n-k))!}$$

Specific Cases

$$\binom{n}{0} = \frac{n!}{0!(n-0)!} = 1$$

$$\binom{n}{1} = \frac{n!}{1!(n-1)!} = n$$

$$\binom{0}{0} = \binom{n}{n} = 1$$

$$\binom{n}{k} = 0 \text{ for } k > n \geq 0$$

Sum of Two Euler's Numbers

$$\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}$$

Derivation of the Property

$$
\begin{aligned}
&\frac{n\,!}{k\,!(n-k)\,!} + \frac{n\,!}{(k+1)\,!(n-(k+1))\,!} \\
&= \frac{(k+1)n\,!}{(k+1)k\,!(n-k)\,!} + \frac{(n-k)n\,!}{(k+1)\,!(n-(k+1))\,!(n-k)} \\
&= \frac{n\,!((k+1)+(n-k))}{(k+1)\,!(n-k)\,!} \\
&= \frac{n\,!(n+1)}{((n+1)-(k+1))\,!(k+1)\,!} \\
&= \frac{(n+1)\,!}{((n+1)-(k+1))\,!(k+1)\,!} \\
&= \binom{n+1}{k+1}
\end{aligned}
$$

Euler's Numbers and Binomial Coefficients

Table 4.1 contains in the left column an expression of the form $(a+b)^n$ and in the right column summands obtained by expansion of the expression in the left column.

Pascal's Triangle

In the Pascal triangle, one can find all the coefficients from Table 4.1, please note the additive dependence between the two rows of the triangle.

Table 4.1 Binomial coefficients

$(a+b)^0$	1
$(a+b)^1$	$1a+1b$
$(a+b)^2$	$1a^2+2ab+1b^2$
$(a+b)^3$	$1a^3+3a^2b+3ab^2+1b^3$
$(a+b)^4$	$1a^4+4a^3b+6a^2b^2+4ab^3+1b^4$
$(a+b)^5$	$1a^5+5a^4b+10a^3b^2+10a^2b^3+5ab^4+1b^5$
$\cdots$	$\cdots$

$$
\begin{array}{ccccccccccccc}
 & & & & & & 1 & & & & & & \\
 & & & & & 1 & & 1 & & & & & \\
 & & & & 1 & & 2 & & 1 & & & & \\
 & & & 1 & & 3 & & 3 & & 1 & & & \\
 & & 1 & & 4 & & 6 & & 4 & & 1 & & \\
 & 1 & & 5 & & 10 & & 10 & & 5 & & 1 & \\
1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1
\end{array}
$$

$$(a+b)^6 = a^6 + 6a^5b + 15a^4b^2 + 20a^3b^3 + 15a^2b^4 + 6ab^5 + b^6$$

Binomial Theorem

Binomial theorem documents the mentioned dependence between Euler's numbers and combination numbers.

$$
\begin{aligned}
(a+b)^n &= \binom{n}{0} a^n + \binom{n}{1} a^{n-1}b + \binom{n}{2} a^{n-2}b^2 + \ldots \\
&\quad \cdots + \binom{n}{n-1} ab^{n-1} + \binom{n}{n} b^n \\
&= \sum_{k=0}^{n} \binom{n}{k} a^{n-k}b^k
\end{aligned}
$$

Chapter 5
Random Variables

5.1 The Definition

Definition A random variable is a function that assigns (real) numbers to the results of an experiment. Each possible outcome of the experiment (i.e., value of the corresponding random variable) occurs with a certain probability.

- X: random variable
- $x_i,\ (i = 1, \ldots, n)$: results of n experiments—the values of the random variable X

A random variable is created by assigning a real number to each event E_j (an outcome of an experiment). The event E_j is an element of the set S of all possible outcomes of an experiment. The random variable is then defined by a function that maps the elements of the set S with numbers on the real line.

$$X : E_j \rightarrow X(E_J) = x_j$$

More Information

A random variable is a function that assigns real numbers to the outcomes of an experiment. Random variables (i.e., the functions) are usually denoted by capital letters. The value of a random variable (i.e., a realization) is not known *before* we conduct the experiment.

A realization of a random variable is obtained only *after* observing the outcome of the experiment. The realization of random variables are usually denoted by small letters. This notation allows us to distinguish the random variable from its realization.

In practice, we usually only have the realizations of the random variables. The goal of statistics is to use these values to obtain the properties of the (unknown) random variable that generates these observations.

W.K. Härdle et al., *Introduction to Statistics*, DOI 10.1007/978-3-319-17704-5_5

Table 5.1 The number of tails in three tosses of a coin

S	Number of tails
$\{hhh\}$	0
$\{hht, hth, thh\}$	3
$\{htt, tht, tth\}$	3
$\{ttt\}$	1

Explained: The Experiment

Two outcomes are possible if you toss a coin: heads (h) or tails (t). Let us consider three tosses ($k = 3$). Our experiment will examine the number (n) of tails obtained in three tosses of a coin. There are 8 possible ($V^W(n;k) = n^k \rightarrow V^W(2;3) = 2^3 = 8$) outcomes of this experiment

$$S = \{hhh, hht, hth, thh, htt, tht, tth, ttt\}$$

The random variable for this experiment assigns a real number $(0, 1, 2, 3)$ to each element of S based on the number of tails appearing in the tosses. For example, tails appears once ($n = 1$): $\{(hho) \cup (hoh) \cup (ohh)\}$. This random variable "works" as shown in Table 5.1.

The corresponding random variable, denoted by the capital letter X, is defined as

$$X = \{\,\text{Number (n) of tails in three tosses of the coin}\,\}.$$

This definition implies that the value of the random variable X has to be one of the following 4 numbers: $x_1 = 0; x_2 = 1; x_3 = 2; x_4 = 3$.

Enhanced: Household Size I

The government carried out a socioeconomic study that examined the relationship between the size of a household and its lifestyle choices.

Let us assume that the government has obtained the following results:

$$E_1 = \{\,\text{households with one person}\,\}$$
$$E_2 = \{\,\text{households with two people}\,\}$$
$$E_3 = \{\,\text{households with three people}\,\}$$
$$E_4 = \{\,\text{households with four and more people}\,\}$$

The set of the possible outcomes from the experiment consists of the following events: $S = \{E_1, E_2, E_3, E_4\}$. We assign a real number to each event $E_i \in S$:

$$\begin{array}{ll} S & R \\ & \\ E_1 & \rightarrow \quad 1 \\ E_2 & \rightarrow \quad 2 \\ E_3 & \rightarrow \quad 3 \\ E_4 & \rightarrow \quad 4 \end{array}$$

The resulting random variable X is defined as the size of the household. The set of possible values of this random variable is $(1, 2, 3, 4)$, this means that the possible results of this random variable are $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4$.

5.2 One-Dimensional Discrete Random Variables

A random variable is one-dimensional if the experiment only considers *one* outcome.

Discrete Random Variable

Definition A random variable is called discrete if the set of all possible outcomes $x_1, x_2, \ldots$ is finite or countable.

Density Function

Definition The density function f gives the probability that the random variable X is equal to x_i. The probability of x_i is $f(x_i)$.

$$P(X = x_i) = f(x_i) \qquad i = 1, 2, \ldots$$
$$f(x_i) \geq 0, \qquad \sum_i f(x_i) = 1$$

The density function can be plotted using a histogram.

Distribution Function

Definition The distribution function F of a random variable X evaluated at a realization x is defined as the probability that the value of the random variable X is not greater than x.

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$$

The distribution function of a discrete random variable is a step function that only increases only at increments of x_i. The distribution function increases in increments of $f(x_i)$. The distribution function is also constant between the points x_i and x_{i+1}.

The distribution function allows us to compute the probabilities of other events involving X:

$$P(a < X \leq b) = F(b) - F(a), \text{ or } P(X > a) = 1 - F(a).$$

Explained: One-Dimensional Discrete Random Variable

We count the number of tails (t) in three tosses of a coin. We define the random variable X:

$$X = \{ \text{The number of tails in three tosses of a coin} \}$$

with the following four outcomes $x_1 = 0; x_2 = 1; x_3 = 2; x_4 = 3$.

The calculation of the probabilities $P(E_j)$ is based on the Multiplication Theorem for independent random events (Table 5.2, Fig. 5.1).

Table 5.2 Probabilities for the number of tails in three tosses of a coin

Event E_j	Probability $P(E_j)$	Number of tails (t) x_j	Probability function $P(X = x_j) = f(x_j)$
$E_1 = \{hhh\}$	$P(E_1) = 0.125$	$x_1 = 0$	$f(x_1) = 0.125$
$E_2 = \{hho\}$	$P(E_2) = 0.125$		
$E_3 = \{hoh\}$	$P(E_3) = 0.125$	$x_2 = 1$	$f(x_2) = 0.375$
$E_4 = \{ohh\}$	$P(E_4) = 0.125$		
$E_5 = \{hoo\}$	$P(E_5) = 0.125$		
$E_6 = \{oho\}$	$P(E_6) = 0.125$	$x_3 = 2$	$f(x_3) = 0.375$
$E_7 = \{ooh\}$	$P(E_7) = 0.125$		
$E_8 = \{ooo\}$	$P(E_8) = 0.125$	$x_4 = 3$	$f(x_4) = 0.125$

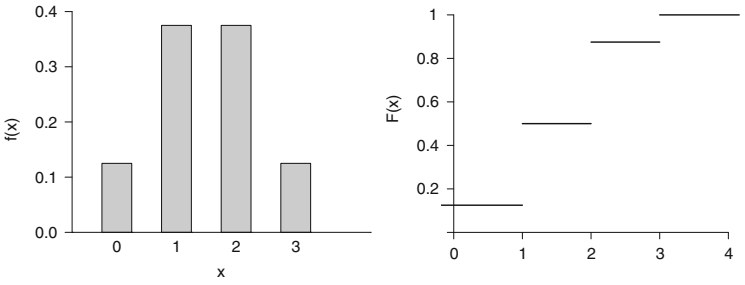


Fig. 5.1 Probabilities and distribution function for the number of tails in three tosses of a coin

The (cumulative) distribution function is obtained by summing the probabilities of the different values of the random variable X. For instance

$$F(1) = f(0) + f(1) = 0.125 + 0.375 = 0.5$$

Formula for the (cumulative) distribution function:

$$F(x) = \begin{cases} 0.000 \text{ for } x < 0 \\ 0.125 \text{ for } 0 \leq x < 1 \\ 0.500 \text{ for } 1 \leq x < 2 \\ 0.875 \text{ for } 2 \leq x < 3 \\ 1.000 \text{ for } 3 \leq x \end{cases}$$

Enhanced: Household Size II

The household sizes in Berlin in April 1998 are provided on page 64 in "Statistisches Jahrbuch" published by "Statistisches Landesamt Berlin," Kulturbuch-Verlag Berlin (Table 5.3).

Let X denote the size of a randomly chosen household from Berlin in April 1998. We can observe the following outcomes:

$x_1 = 1$ household with one person
$x_2 = 2$ household with two persons
$x_3 = 3$ household with three persons
$x_4 = 4$ household with four or more persons

Before we choose the household, we cannot say anything about its size. The value of the random variable can take any from the four possible outcomes. We let $X =$ household size denote the random variable in this experiment. X is discrete, because the set of all possible outcomes is finite—the outcome must take a value of 1, 2, 3, or 4 (and more).

The probabilities are given by the frequency distribution of the households in Berlin (Fig. 5.2). This density function provides an overview of all possible outcomes together with their probabilities (Table 5.4).

Table 5.3 Household sizes in Berlin in April 1998

Household size	Number of households (1000)
1	820.7
2	564.7
3	222.9
4 and more	195.8
Sum	1804.1

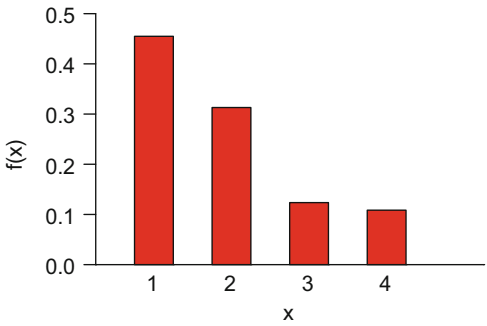


Fig. 5.2 Probabilities of household sizes in Berlin in April 1998

Table 5.4 Probabilities of household sizes in Berlin in April 1998

Household size x_j	$f(x_j)$
1	0.4549
2	0.3130
3	0.1236
4 and more	0.1085
Sum	1.0000

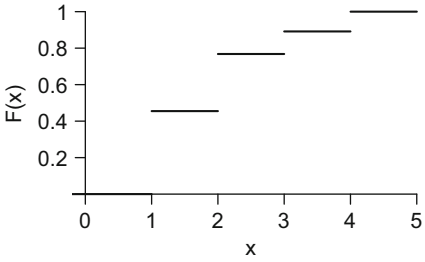


Fig. 5.3 Cumulative probabilities of household sizes in Berlin in April 1998

Table 5.5 Cumulative probabilities of household sizes in Berlin in April 1998

Household size x_j	$F(x)$
1	0.4549
2	0.7679
3	0.8915
4 and more	1.0000

The probability that a household (from Berlin in April 1998) contains two persons ($X = 2$) is equal to 0.313 (Fig. 5.2, Table 5.4). The distribution function $F(x) = P(X \leq x)$ is (Fig. 5.4, Table 5.5):

Similarly, the distribution function provides the probability that a household has at most two members ($X \leq 2$) is equal to 0.7679. The distribution function also allows us to compute the probabilities of other outcomes, e.g.

- probability that a household has more than two members ($X > 2$) is

$$P(X > 2) = 1 - F(2) = 1 - 0.7679 = 0.2321$$

or

$$P(X > 2) = f(3) + f(4) = 0.1236 + 0.1085 = 0.2321.$$

- probability that a household has more than one member but less than four members is equal to

$$P(1 < X \leq 3) = F(3) - F(1) = 0.8915 - 0.4549 = 0.4366$$

or

$$P(1 < X \leq 3) = f(2) + f(3) = 0.3130 + 0.1236 = 0.4366.$$

5.3 One-Dimensional Continuous Random Variables

Definition A continuous random variable takes values on the real line from either a finite or infinite interval.

Density Function

Definition If a function $f(x)$ has the following properties:

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

$$f(x) \geq 0$$

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1$$

The function $f(x)$ is the density of the continuous random variable X.

Distribution Function

Definition The distribution function can be obtained from the density:

$$F(x) = P(-\infty < X \leq x)$$
$$= \int_{-\infty}^{x} f(t)\,dt.$$

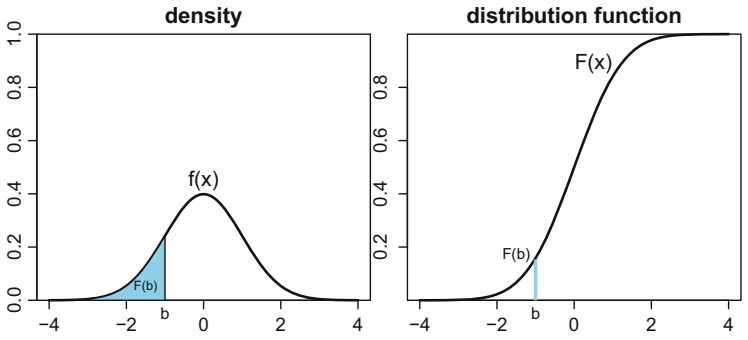


Fig. 5.4 Density and distribution function of a continuous random variable

The distribution function $F(x)$ is equal to the area under the density $f(u)$ for $-\infty < u \leq x$ (Fig. 5.4).

The density function, if it exists, can be computed as the first derivative of the distribution function:

$$\frac{\partial F(x)}{\partial x} = F'(x) = f(x).$$

More Information: Continuous Random Variable, Density, and Distribution Function

The density function of a continuous random variable has the following properties:

- it cannot be negative
- the area under the curve is equal to one
- probability that the random variable X lies between a and b is equal to the area between the density and the x-axis on the interval $[a, b]$

The density function $f(x)$ computes the probability that a random variable lies in the interval $[x, x + dx]$.

The probability that a continuous random variable will be equal to a specific real number is always equal to zero, since the area under a specific point is equal to zero:

$$\int_x^x f(t)\, dt = F(x) - F(x) = 0\,.$$

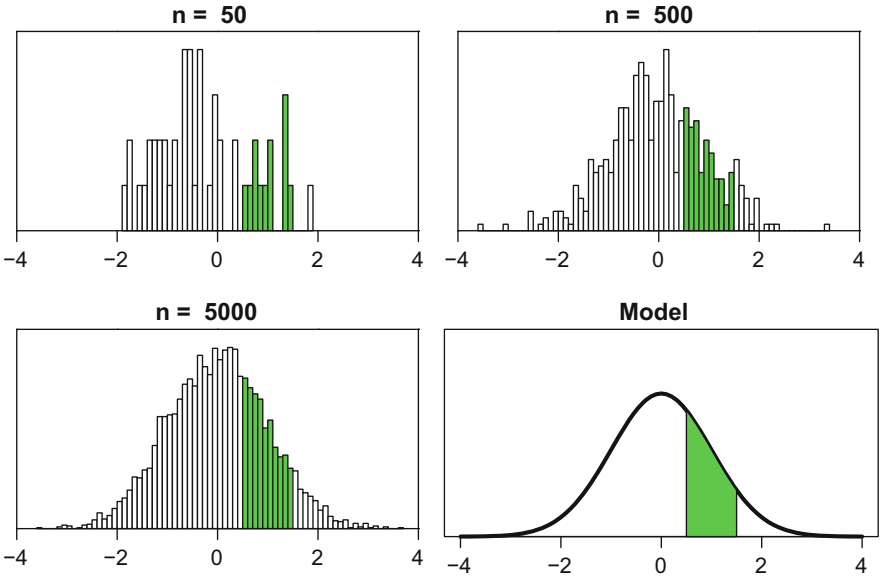


Fig. 5.5 Smoothing histograms by increasing the number of observations

This implies as a corollary: the probability that continuous random variable X falls into an interval does not depend on the closeness or openness of the interval.

$$P(a \leq X \leq b) = P(a < X < b) \text{ because } P(a) = P(b) = 0\,.$$

The diagram in Fig. 5.5 illustrates that a histogram can be smoothed by increasing the number of observations. In the limit (i.e., as n$\rightarrow\ \infty$) the histogram can be approximated by a continuous function.

The area between the points a and b corresponds to the probability that a random variable X will fall in the interval $[a, b]$. This probability can be computed using integrals.

A distribution function, $F(x)$, is the probability that the random variable X is less than or equal to x. Its properties follow:

- $F(x)$ is nondecreasing, i.e., $x_1 < x_2$ implies that $F(x_1) \leq F(x_2)$
- $F(x)$ is continuous
- $0 \leq F(x) \leq 1$
- $\lim_{x\rightarrow-\infty} F(x) = 0$
- $\lim_{x\rightarrow+\infty} F(x) = 1$

A distribution function cannot be decreasing because this would imply negative probabilities. In general, the distribution function is defined for all real numbers. Limits on the sample space are necessary for the complete description of the distribution function.

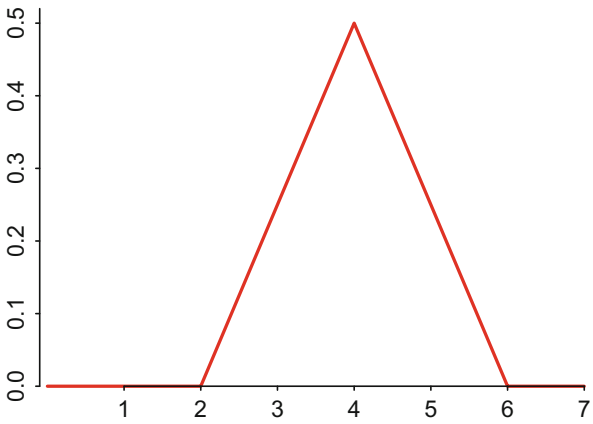


Fig. 5.6 Density of the triangular distribution

Explained: Continuous Random Variable

Let us consider the function

$$f(x) = \begin{cases} 0.25x - 0.5 & \text{for } 2 < x \leq 4 \\ -0.25x + 1.5 & \text{for } 4 < x \leq 6 \\ 0 & \text{otherwise.} \end{cases}$$

Is this function a density? We need to verify whether $\int_{-\infty}^{\infty} f(x)\,dx = 1$:

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)\,dx &= \int_2^4 (0.25x - 0.5)\,dx + \int_4^6 (-0.25x + 1.25)\,dx \\ &= \left[0.25\frac{1}{2}x^2 - 0.5x\right]_2^4 + \left[-0.25\frac{1}{2}x^2 + 1.5x\right]_4^6 = 1 \end{aligned}$$

This means that $f(x)$ is a density. In particular, it is the density of the triangular distribution (named after the shape of the density shown in Fig. 5.6).

Enhanced: Waiting Times of Supermarket Costumers

The waiting times (in minutes) of supermarket customers were collected, which resulted in the frequency distribution shown in Table 5.6. The relative frequencies are used to construct the histogram and the frequency polygon shown in Figs. 5.7 and 5.8.

The continuous random variable $X = \{\text{waiting time}\}$ defines the groups (bins) with constant bin width of 0.5 min. The probabilities are approximated by relative frequencies (statistical definition of the probability).

Table 5.6 Waiting times of supermarket customers in minutes

Waiting time	Relative frequency	Cumulative relative frequency
8.0–8.5	0.002	0.002
8.5–9.0	0.004	0.006
9.0–9.5	0.009	0.015
9.5–10.0	0.013	0.028
10.0–10.5	0.020	0.048
10.5–11.0	0.043	0.091
11.0–11.5	0.094	0.185
11.5–12.0	0.135	0.320
12.0–12.5	0.169	0.489
12.5–13.0	0.158	0.647
13.0–13.5	0.139	0.786
13.5–14.0	0.078	0.864
14.0–14.5	0.065	0.929
14.5–15.0	0.030	0.959
15.0–15.5	0.010	0.969
15.5–16.0	0.014	0.983
16.0–16.5	0.006	0.989
16.5–17.0	0.004	0.993
16.0–17.5	0.003	0.996
17.5–18.0	0.004	1.000

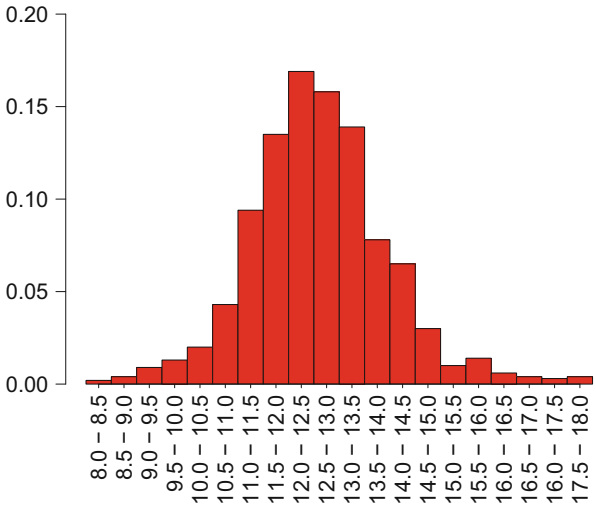


Fig. 5.7 Histogram of the waiting times of supermarket customers

Note In Fig. 5.7, the probabilities are given as the height of the boxes (and not the areas of the boxes). This implies that the sum of the areas of all of the boxes is equal to 0.5 (and not to 1). Similarly, the polygon on Fig. 5.8 cannot be a density because

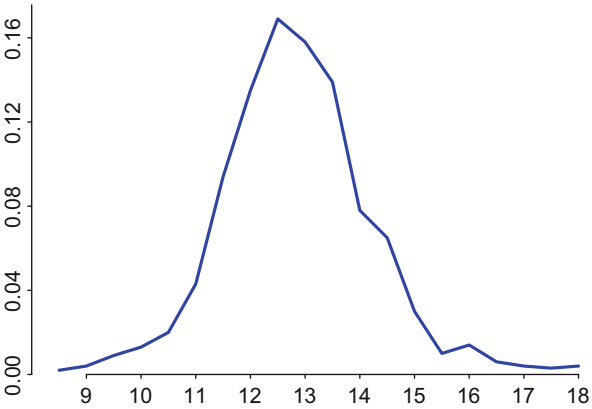


Fig. 5.8 Polygon of the waiting times of supermarket customers

Table 5.7 Relative frequency density of waiting times of supermarket costumers

Waiting time	Relative frequency density
8.0–8.5	0.004
8.5–9.0	0.008
9.0–9.5	0.018
9.5–10.0	0.026
10.0–10.5	0.040
10.5–11.0	0.086
11.0–11.5	0.188
11.5–12.0	0.270
12.0–12.5	0.338
12.5–13.0	0.316
13.0–13.5	0.278
13.5–14.0	0.156
14.0–14.5	0.130
14.5–15.0	0.060
15.0–15.5	0.020
15.5–16.0	0.028
16.0–16.5	0.012
16.5–17.0	0.008
16.0–17.5	0.006
17.5–18.0	0.008

it does not satisfy the condition

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1\,.$$

In order to obtain the density of X, we need to compute the relative frequency density, which is obtained as the ratio of the relative frequencies and the widths of the corresponding groups summarized in Table 5.7.

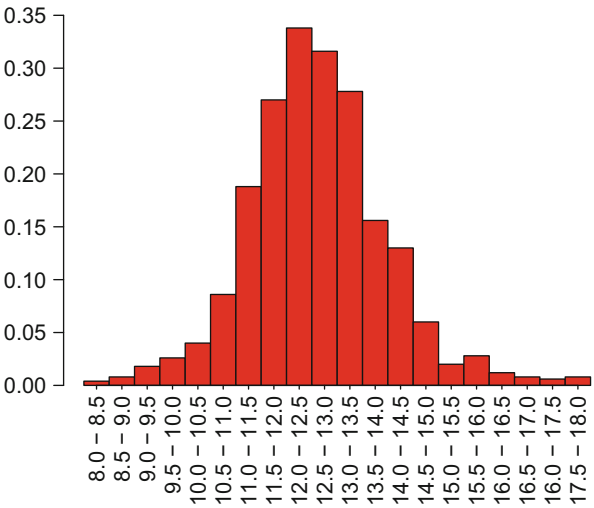


Fig. 5.9 Histogram of the waiting times of supermarket customers using relative frequency density

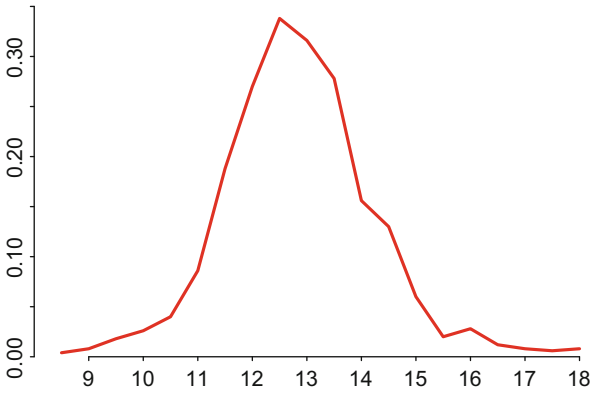


Fig. 5.10 Density of the waiting times of supermarket customers

Using this relative frequency density we obtain another histogram and a smoothed density function.

In Fig. 5.9 the probabilities of the groups are given by the area. This implies that the sum of these areas is equal to one. The density in Fig. 5.10 is (an approximate) density function of the (continuous) random variable $X = \{\text{waiting time of the customer}\}$. The corresponding distribution function $F(x)$ is given in Fig. 5.11.

5.4 Parameters

A random variable is completely described by its density and distribution function. However, some important aspects of the probability distribution can be

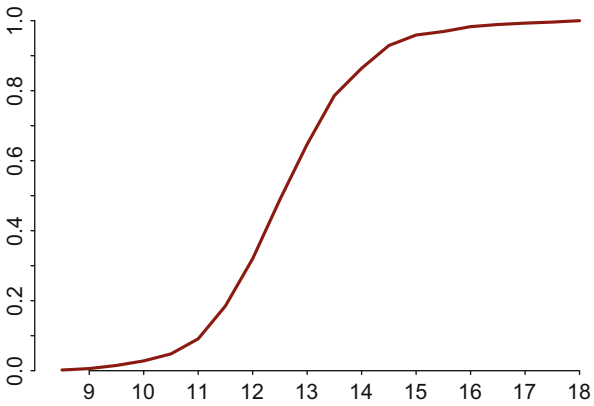


Fig. 5.11 Distribution function of the waiting times of supermarket customers

characterized by a small number of parameters. The most important of which are the location and scale parameters of a random variable.

Expected Value

The expected value of a random variable X, denoted by $E(X)$ or μ, corresponds to the arithmetic mean of an empirical frequency distribution. The expected value is the value that we, on average, expect to obtain as an outcome of the experiment. By repeating the experiment many times, the expected value $E(X)$ is the number that will be obtained as an average of all the outcomes of an experiment.

Definition Let us consider the discrete random variable X with outcomes x_i and the corresponding probabilities $f(x_i)$. Then, the expression

$$E(X) = \mu = \sum_i x_i f(x_i)$$

defines the expected value of the random variable X.

For a continuous random variable X, with density $f(x)$, we define the expected value as

$$E(X) = \mu = \int_{-\infty}^{+\infty} x \cdot f(x)\, dx$$

Properties of the Expected Value Let X and Y be two random variables with the expected values $E(X)$ and $E(Y)$. Then:

- for $Y = a + bX$ with any a, b
  $E(Y) = E(a + bX) = a + bE(X)$

- for $Z = X + Y$
  $E(Z) = E(X + Y) = E(X) + E(Y)$
- for X, Y independent random variables
  $E(XY) = E(X)E(Y)$

Variance

Definition The variance, which is usually denoted by $Var(X)$ or σ^2, is defined as expected value of the squared difference between a random variable and its expected value:

$$Var(X) = E[(X - E(X))^2] = E(X^2) = [E(X)]^2$$

For discrete random variable we obtain

$$Var(X) = \sigma^2 = \sum_i [x_i - E(X)]^2 \cdot f(x_i) = \sum_i x_i^2 f(x_i) - [E(X)]^2$$

and for a continuous random variable the variance is defined as

$$Var(X) = \sigma^2 = \int_{-\infty}^{+\infty} [x - E(X)]^2 \cdot f(x)\, dx = \int_{-\infty}^{+\infty} x^2 f(x)\, dx - [E(X)]^2$$

The Properties of the Variance Assume that X and Y are two random variables with the variances $Var(X)$ and $Var(Y)$. Then:

- for $Y = a + bX$, where a and b are constants
  $Var(Y) = Var(a + bX) = b^2 Var(X)$
- for X, Y independent random variables and $Z = X + Y$
  $Var(Z) = Var(X) + Var(Y)$

$$\sigma_Z = \sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$$

Standard Deviation

The standard deviation σ is defined as the square root of the variance, which summarizes the spread of the distribution. Large values of the standard deviation mean that the random variable X is likely to vary in a large neighborhood around the expected value. Smaller values of the standard deviation indicate that the values of X will be concentrated around the expected value.

Standardization

Sometimes, it is useful to transform a random variable in order to obtain a distribution that does not depend on any (unknown) parameters. It is easy to show that the standardized random variable

$$Z = \frac{X - E(X)}{\sigma_X}$$

has expected value $E(Z) = 0$ and variance $Var(Z) = 1$.

Chebyshev's Inequality

Chebyschev's inequality provides a *bound* on the probability that a random variable falls within some interval around its expected value. This inequality only requires us to know the expected value and the variance of the distribution; we do not have to know the distribution itself. The inequality is based on the interval $[\mu - k \cdot \sigma; \mu + k \cdot \sigma]$ which is centered around μ.

Definition Consider the random variable X with expected value μ and variance σ. Then, for any $k > 0$, we have

$$P(\mu - k \cdot \sigma \leq X \leq \mu + k \cdot \sigma) \geq 1 - \frac{1}{k^2}$$

Denoting $k \cdot \sigma = a$, we obtain

$$P(\mu - a \leq X \leq \mu + a) \geq 1 - \frac{\sigma^2}{k^2}$$

We can use the inequality to also obtain a bound for the complementary event that the random variable X falls outside the interval, i.e., $\{|X - \mu| > k \cdot \sigma\}$

$$P(|X - \mu| > k \cdot \sigma) < 1/k^2$$

and for $k \cdot \sigma = a$

$$P(|X - \mu| > a) < \sigma^2/a^2.$$

Note that the exact probabilities $\{|X - \mu| < k \cdot \sigma\}$ and $\{|X - \mu| \leq k \cdot \sigma\}$ depend on the specific distribution X.

Explained: Continuous Random Variable

Let X be continuous random variable with the density

$$f(x) = \begin{cases} 0.25x - 0.5 & \text{for } 2 < x \leq 4 \\ -0.25x + 1.5 & \text{for } 4 < x \leq 6 \\ 0 & \text{otherwise} \end{cases}$$

We calculate the expected value of X:

$$\begin{aligned} E(X) = \mu &= \int_{-\infty}^{\infty} xf(x)\,dx \\ &= \int_2^4 x(0,25x - 0,5)\,dx + \int_4^6 x(-0,25x + 1,5)\,dx \\ &= \int_2^4 (0,25x^2 - 0,5x)\,dx + \int_4^6 (-0,25x^2 + 1,5x)\,dx \\ &= \left[0,25\frac{1}{3}x^3 - 0,5\frac{1}{2}x^2\right]_2^4 + \left[-0,25\frac{1}{3}x^3 + 1,5\frac{1}{2}x^2\right]_4^6 \\ &= 4 \end{aligned}$$

Now we calculate the variance:

$$\begin{aligned} Var(X) = \sigma^2 &= \int_{-\infty}^{\infty} x^2 f(x)\,dx - [E(X)]^2 \\ &= \int_2^4 x^2(0,25x - 0,5)\,dx + \int_4^6 x^2(-0,25x + 1,5)\,dx - 4^2 \\ &= \int_2^4 (0,25x^3 - 0,5x^2)\,dx + \int_4^6 (-0,25x^3 + 1,5x^2)\,dx - 4^2 \\ &= \left[0,25\frac{1}{4}x^4 - 0,5\frac{1}{3}x^3\right]_2^4 + \left[-0,25\frac{1}{4}x^4 + 1,5\frac{1}{3}x^3\right]_4^6 - 16 \\ &= 0,\ 6667\,. \end{aligned}$$

The standard deviation is equal to $\sigma = 0.8165$.

For this continuous random variable the distribution has an expected value 4 and a standard deviation 0.8165.

Table 5.8 Frequency distribution of the number of traffic accidents occurring at an intersection during a week

x_i	0	1	2	3	4	5
$f(x_i)$	0.008	0.18	0.32	0.22	0.14	0.06

Explained: Traffic Accidents

Let the random variable X denote the number of traffic accidents occurring at an intersection during a week. From long-term records, we know the frequency distribution of X given in Table 5.8.

The expected value of X, i.e., the expected number of crashes, can be computed as follows:

x_i	0	1	2	3	4	5
$f(x_i)$	0.08	0.18	0.32	0.22	0.14	0.06
$x_i f(x_i)$	0.00	0.18	0.64	0.66	0.56	0.30

This gives

$$E(X) = \mu = \sum x_i f(x_i) = 2.34\,.$$

This number of traffic accidents is, of course, not possible, since we cannot have 2.34 accidents during a week. The value $E(X) = 2.34$ just shows the center of the probability function of the random variable X.

Now we calculate the standard deviation:

x_i^2	0	1	4	9	16	25
$x_i^2 f(x_i)$	0.00	0.18	1.28	1.98	2.24	1.50

$$Var(X) = \sigma^2 = \sum x_i^2 f(x_i) - \mu^2 = 7.18 - 2.34^2 = 1.7044 \;\Rightarrow \sigma = 1.306\,.$$

We can expect that the distribution function for accidents at this intersection has a mean of 2.34 and a standard deviation of 1.306.

5.5 Two-Dimensional Random Variables

Consider two random variables X and Y. The joint probability distribution function of two random discrete variables X and Y is defined as the probability that X is equal to x_i at the same time that Y is equal to y_j.

$$P(\{X = x_i\} \cap \{Y = y_j\}) = P(X = x_i, Y = y_j) = f(x_i, y_j) \quad i, j = 1, 2, \ldots$$

Table 5.9 Structure of a contingency table

X/Y	y_1	...	y_j	...
x_1	$f(x_1, y_1)$	...	$f(x_1, y_j)$	...
:	:	...	:	...
x_i	$f(x_i, y_1)$	...	$f(x_i, y_j)$	...
:	:	...	:	...

if the following two conditions hold:

$$f(x_i, y_j) > 0 \quad i, j = 1, 2, \ldots \qquad \text{and} \sum_i \sum_j f(x_i, y_j) = 1.$$

These two-dimensional probability density functions, for discrete random variables, can be represented in the form of a contingency table (cross-tabulation) (Table 5.9).

For the density function of a pair of continuous random variables we have:

$$P(x < X \leq x + \triangle x; y < Y \leq y + \triangle y) = f(x, y)$$

$$F(x, y) \geq 0, \quad \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)\, dx\, dy = 1$$

The (cumulative) distribution function $F(x, y)$ is equal to the probability that the random variable X is not greater than x and, at the same time, the variable Y is not greater than y.

The distribution function for a pair of discrete random variables can be written as:

$$F(x, y) = P(X \leq x,\ Y \leq y) = \sum_{x_i \leq x} \sum_{y_j \leq y} f(x_i, y_j)$$

The distribution function for a pair of continuous random variables:

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v)\, du\, dv$$

Marginal Distribution

The marginal distribution, $f(x_i)$, of a discrete random variable X provides the probability that the variable X is equal to x_i without considering the variable Y.

Table 5.10 Contingency table with marginal distributions

X / Y	y_1	...	y_j	...	MD X
x_1	$f(x_1, y_1)$	...	$f(x_1, y_j)$	...	$f(x_1)$
:	:	...	:	...	:
x_i	$f(x_i, y_1)$	...	$f(x_i, y_j)$	...	$f(x_i)$
:	:	...	:	...	:
MD Y	$f(y_1)$	...	$f(y_j)$	...	1.00

The marginal distribution for the random variable Y, $f(y_j)$, is defined analogously.

$$P(X = x_i) = f(x_i) = \sum_j f(x_i, y_j)$$

$$P(Y = y_j) = f(y_j) = \sum_i f(x_i, y_j)$$

The resulting marginal distributions are one-dimensional (Table 5.10).

Similarly, we obtain the marginal densities for a pair of continuous random variables X and Y:

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

$$f(y) = \int_{-\infty}^{+\infty} f(x, y)\, dx$$

The Conditional Marginal Distribution Function

The conditional marginal distribution function $F_y(x)$ of the random variable X denotes the distribution function of the random variable X conditional on the value of Y. It is defined as:

$$P(X \leq x|Y) = F_y(x) = \begin{cases} \sum_{j=-\infty}^{+\infty} \sum_{i=-\infty}^{x} f(x_i, y_j) & \text{for } X \text{ discrete} \\ \int_{-\infty}^{+\infty} \int_{-\infty}^{x} f(u, v)\, du\, dv & \text{for } X \text{ continuous} \end{cases}$$

The conditional marginal distribution function $F_x(y)$ of the random variable Y denotes the distribution function of the random variable Y conditional on the value

Table 5.11 Joint probability distribution of X="Voted" and Y="Interest in politics"

Voted X	Interest in politics Y					
	Very int. (y_1)	Int. (y_2)	Medium int. (y_3)	Low int. (y_4)	No int. (y_5)	MD X
Yes (x_1)	0.107	0.196	0.398	0.152	0.042	0.895
No (x_2)	0.006	0.011	0.036	0.031	0.021	0.105
MD Y	0.113	0.207	0.434	0.183	0.063	1.000

of X, defined as:

$$P(Y \leq y|X) = F_x(y) = \begin{cases} \sum_{j=-\infty}^{y} \sum_{i=-\infty}^{+\infty} f(x_i, y_j) & \text{for } Y \text{ discrete} \\ \int_{-\infty}^{y} \int_{-\infty}^{+\infty} f(u, v)\, du\, dv & \text{for } Y \text{ continuous} \end{cases}$$

Explained: Two-Dimensional Random Variable

Example of Two Discrete Random Variables

The inhabitants of a district were asked

- whether they voted in the last election (random variable X with possible values $x_1 =$ yes, and $x_2 =$ no).
- whether they are interested in politics (random variable Y with possible values $y_1 =$ very intensively, $y_2 =$ intensively, $y_3 =$ medium interest, $y_4 =$ low interest, and $y_5 =$ no interest).

The joint probability distribution of these random variables is presented in the contingency table given in Table 5.11.

Each entry in Table 5.11 contains the probability that the random variable X will take the value x_i at the same time Y equals y_j, and vice versa. The entry of the $(1, 2)$ element provides the probability that a person who is very interested in politics voted in the last election. It is 0.196. The marginal distribution (MD) of X provides the probability distribution of the random variable "Voted." For example, 0.105 is the probability that a (randomly chosen) citizen participated in the most recent parliamentary elections. The marginal distribution (MD) of Y is the probability distribution of the random variable "interest in politics." For example, 0.183 is the probability that a (randomly chosen) citizen has low interest in politics.

Figure 5.12 presents the joint probability distribution function for voting behavior and interest in politics.

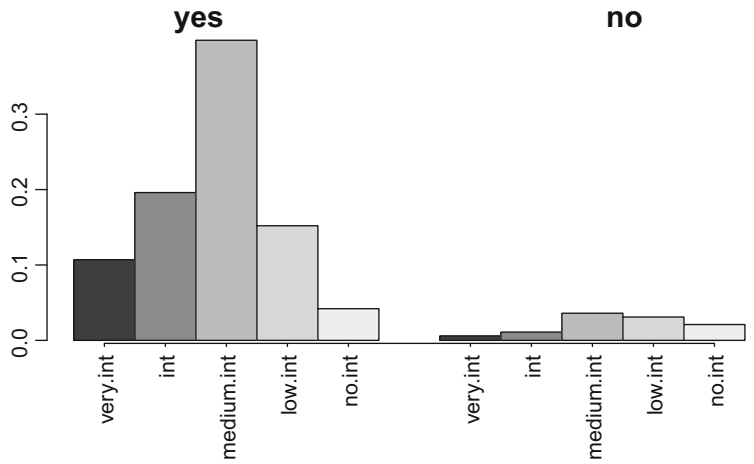


Fig. 5.12 Joint probability distribution of X="Voted" (yes or no) and Y="Interest in politics"

Example of Two Continuous Random Variables

Let us consider two continuous random variables X and Y with the joint density

$$f(x,y) = \begin{cases} \frac{x+3y}{2} & \text{for } 0 < x < 1, \text{ and } 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

For this density, we have the following:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = \int_0^1\int_0^1 \frac{x+3y}{2}\,dx\,dy = \int_0^1 \left[\frac{x^2}{4} + \frac{3xy}{2}\right]_0^1 dy$$
$$= \int_0^1 \left(\frac{1}{4} + \frac{3y}{2}\right) dy = \left[\frac{y}{4} + \frac{3y^2}{2}\right]_0^1 = 1$$

Figure 5.13 contains the graphical display of the two-dimensional probability distribution function of the variables X and Y. We obtain the following marginal distributions:

$$f(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_0^1 \frac{x+3y}{2}\,dy = \left[\frac{xy}{2} + \frac{3y^2}{4}\right]_0^1$$
$$f(x) = \begin{cases} \frac{x}{2} + \frac{3}{4} & \text{for } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

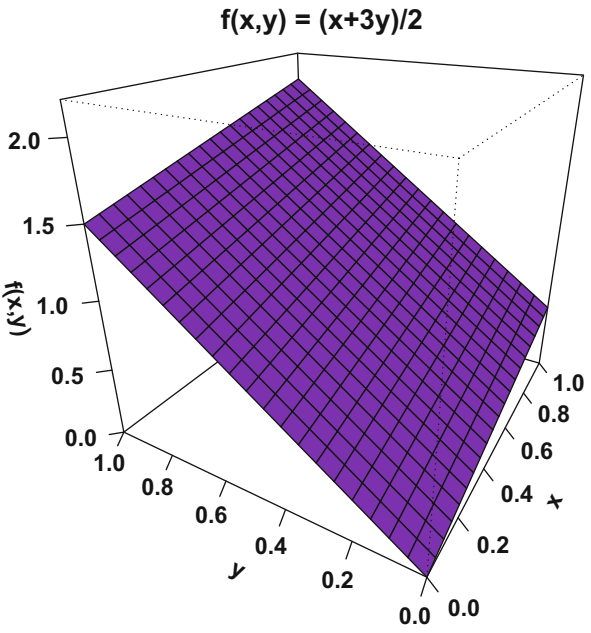


Fig. 5.13 Two-dimensional probability distribution function

and

$$f(y) = \int_{-\infty}^{\infty} f(x,y)\,dx = \int_0^1 \frac{x+3y}{2}\,dx = \left[\frac{x^2}{4} + \frac{3xy}{2}\right]_0^1$$

$$f(y) = \begin{cases} \frac{3y}{2} + \frac{1}{4} & \text{for } 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Enhanced: Link Between Circulatory Diseases and Patient Age

A cardiologist believes there may be a link between certain circulatory diseases and the age of a patient. Therefore, he collected the values of these two random variables for 100 patients. Let X denote the age of the patients and let Y be an indicator of disease status, which takes values $y_1 = 0$ (patient is healthy) and $y_2 = 1$ patient is sick.

The first step of the analysis, which allows us to assess the validity of the cardiologist's hypothesis, is to describe the joint distribution of the two random variables in the form of a contingency table.

In order to simplify the presentation of the results, we group the ages into intervals. The width of the interval for most of these age groups was 5 years; however for very young and very old patients we used 10 years:

$$20-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59 \text{ and } 60-69\,.$$

Table 5.12 Joint probability distribution of X="Age" and Y="Circulatory Disease"

	Circulatory disease Y		
Age X	$y_1 = 0$ (no)	$y_2 = 1$ (yes)	MD X
20–29	0.09	0.01	0.10
30–34	0.13	0.02	0.15
35–39	0.09	0.03	0.12
40–44	0.10	0.05	0.15
45–49	0.07	0.06	0.13
50–54	0.03	0.05	0.08
55–59	0.04	0.13	0.17
60–69	0.02	0.08	0.10
MD Y	0.57	0.43	1.00

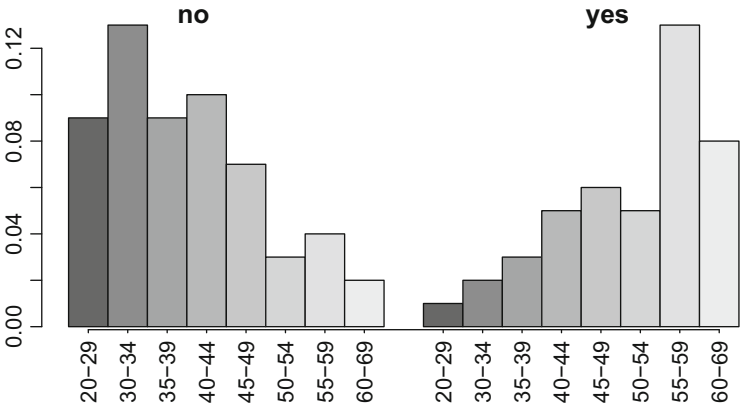


Fig. 5.14 Joint probability distribution of X="Age" and Y="Circulatory Disease" (yes or no)

Now, the joint probability distribution of the two random variables is shown in Table 5.12. The entries of this contingency table contain the probabilities that the random variable X falls within the group x_i at the same time Y equals y_j.

For example, the entry in the second row and first column, i.e., field (2,1), of the table provides the probability that a patient between 30 and 34 years of age does not suffer from a circulatory disease (0.13).

The marginal distribution (MD) of X is the probability distribution of the variable "Age." Using this marginal distribution, we can compute the probability that a patient has a certain age, e.g., the probability that a patient is between 30 and 34 years is 0.15.

The marginal distribution (MD) of Y provides the probability of the disease independently of the age of the patient. For example, using this marginal distribution the probability that a patient suffers from a circulatory disease is 0.43. Figure 5.14 contains the joint probability distribution functions for age and morbidity.

The cardiologist knows from experience that older persons are more likely to suffer from circulatory diseases than other age groups. Therefore, he changed the

Table 5.13 Joint probability distribution of X="Age" and Y="Circulatory Disease," using a simplified grouping

	Circulatory disease Y		
Age X	$y_1 = 0$ (no)	$y_2 = 1$ (yes)	MD X
Less than 40	0.32	0.07	0.39
41–54	0.19	0.15	0.34
More than 55	0.06	0.21	0.27
MD Y	0.57	0.43	1.00

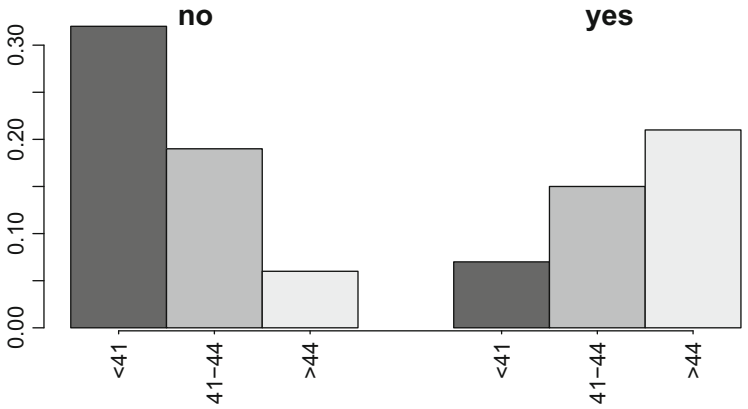


Fig. 5.15 Joint probability distribution of X="Age" and Y="Circulatory Disease" (yes or no), using a simplified grouping

age groupings to: younger than 40, 41–54 years, older than 55 years. Using this simplified grouping, we obtain the joint distribution shown in Table 5.13. The plot of this two-dimensional probability function for the simplified grouping is provided in Fig. 5.15.

Conclusion Grouping is necessary for discrete random variables with a large number of possible outcomes. The information obtained from a contingency table will depend on the groupings used. Therefore, it is recommended to perform detailed statistical analysis for different groupings.

5.6 Independence

(Stochastic) independence of two random variables X and Y is given by the Multiplication Theorem for independent random event.

If two events, A and B, are independent, then the probability that these two events occur at the same time equals the product of their probabilities.

$$P(A \cap B) = P(A) \cdot P(B)$$

Let us consider the events $A = \{X = x_i\}$ and $B = \{Y = y_j\}$. We can now define the independence of these two (discrete) random variables: We say that the random variables X and Y are stochastically independent if

$$P(X = x_i, Y = y_j) = P(X = x_i) \cdot p(Y = y_j)$$

or equivalently

$$f(x_i, y_j) = f(x_i) \cdot f(y_j)$$

for all pairs (x_i, y_j) of the possible outcomes of the random variables X and Y.

The random variables are dependent if there exists at least one pair of points (x_i, y_j) for which the joint distribution does not factor.

We define independence for two continuous random variables in a similar manner:

Two continuous random variables X and Y are stochastically independent if their densities, $f(x)$ and $f(y)$, are such that

$$f(x, y) = f(x) \cdot f(y)$$

for all values (x, y) on the real line.

Conditional Distribution

Let us denote by $P(X = x_i|Y = y_j)$ the probability that a discrete random variable X is equal to x_i conditional on Y equaling y_j. Similarly, we denote by $P(Y = y_j|X = x_i)$ the probability that Y is equal to y_j conditional on $X = x_i$.

Simple probability theory suggests

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{and}$$

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

With discrete random variables, for $A = \{X = x_i\}$ and $B = \{Y = y_j\}$, we obtain

$$\begin{aligned} P(X = x_i|Y = y_j) &= \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} \\ &= \frac{f(x_i, y_j)}{f(y_j)} = f(x_i|y_j) \end{aligned}$$

$$P(Y = y_j|X = x_i) = \frac{P(X = x_i, Y = y_j)}{P(X = x_i)}$$
$$= \frac{f(x_i, y_j)}{f(x_i)} = f(y_j|x_i)$$

Similarly, for continuous random variables:

$$f(x|y) = \frac{f(x, y)}{f(y)}$$
$$f(y|x) = \frac{f(x, y)}{f(x)}$$

For conditional distribution functions we have the following:

$$F(x|y) = \frac{F(x, y)}{F(y)} = \begin{cases} \dfrac{\sum_{i=-\infty}^{x} \sum_{j=-\infty}^{y} f(x_i, y_j)}{\sum_{i=-\infty}^{+\infty} \sum_{j=-\infty}^{y} f(x_i, y_j)} & \text{for } X \text{ and } Y \text{ discrete} \\ \dfrac{\int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v)\, dv\, du}{\int_{-\infty}^{+\infty} \int_{-\infty}^{y} f(u,v)\, dv\, du} & \text{for } X \text{ and } Y \text{ continuous} \end{cases}$$

$$F(y|x) = \frac{F(x, y)}{F(x)} = \begin{cases} \dfrac{\sum_{i=-\infty}^{x} \sum_{j=-\infty}^{y} f(x_i, y_j)}{\sum_{i=-\infty}^{x} \sum_{j=-\infty}^{+\infty} f(x_i, y_j)} & \text{for } X \text{ and } Y \text{ discrete} \\ \dfrac{\int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v)\, dv\, du}{\int_{-\infty}^{x} \int_{-\infty}^{+\infty} f(u,v)\, dv\, du} & \text{for } X \text{ and } Y \text{ continuous} \end{cases}$$

More Information

The independence of random variables is defined using the Multiplication Theorem for random events. To assess whether two random variables are stochastically dependent (or independent), we have to examine whether the product of the marginal distributions equals the joint distribution. If the product of the marginal distributions is equal to the joint distribution of the random variables, for all values x_i; y_j, then they are independent

We are often also interested in the one-dimensional distribution of one of these random variables. This marginal distribution does not depend on the value of the

other variable. This is why we calculate the row and column sums in the two-dimensional contingency table.

In addition, we often examine the distribution of one variable conditional on the value of the other variable. For example, the conditional distribution of X given Y, or Y given X. The conditional distribution is computed by dividing the values of the joint distribution by the values of the marginal distribution.

All formulas used for the discrete variables can be rewritten for continuous random variables.

Explained: Stochastic Independence

Example for Two Discrete Random Variables

The citizens of a certain town were asked

- whether they voted in the parliamentary elections (random variable X with possible outcomes $x_1 =$ yes, and $x_2 =$ no.
- whether they were interested in politics (random variable Y with possible outcomes $y_1 =$ very interested, $y_2 =$ interested, $y_3 =$ medium interest, $y_4 =$ low interest, and $y_5 =$ no interest.

The joint probability distribution of these random variables is provided in Table 5.14. From this joint distribution we can obtain the conditional distributions given in Table 5.15 and 5.16.

The probability that a randomly chosen citizen is very interested in politics who voted the last election ($X =$ yes) is 0.219. On the other hand, the probability for a randomly chosen citizen who is interested in politics but did not vote in the elections ($X =$ no) is only 0.105.

A person with low interest in politics ($Y =$ low interest) voted in the last election with 0.831 ($X =$ yes).

Comparing the conditional distributions $f(y_j|x_i)$ and $f(x_i|y_j)$ indicates that these random variables are not independent, since the conditional distributions differ. The dependence of these variables can be verified by computing $f(x_i, y_j) = f(x_i)f(y_j)$ for all i and j, and comparing it with the observed values of $f(x_i, y_j)$. For example,

Table 5.14 Joint probability distribution of X="Voted" and Y="Interest in politics"

	Interest in politics					
Voted	Very int. (y_1)	Int. (y_2)	Medium int. (y_3)	Low int. (y_4)	No int. (y_5)	MD X
Yes (x_1)	0.107	0.196	0.398	0.152	0.042	0.895
No (x_2)	0.006	0.011	0.036	0.031	0.021	0.105
MR Y	0.113	0.207	0.434	0.183	0.063	1.000

Table 5.15 Conditional distribution $f(y_j|x_i)$

	Interest in politics					
Voted	Very int. (y_1)	Int. (y_2)	Medium int. (y_3)	Low int. (y_4)	No int. (y_5)	
Yes (x_1)	0.120	0.219	0.444	0.170	0.047	1.00
No (x_2)	0.057	0.105	0.343	0.295	0.200	1.00

Table 5.16 Conditional distribution $f(x_i|y_j)$

	Interest in politics				
Voted	Very int. (y_1)	Int. (y_2)	Medium int. (y_3)	Low int. (y_4)	No int. (y_5)
Yes (x_1)	0.947	0.947	0.917	0.831	0.667
No (x_2)	0.053	0.053	0.083	0.169	0.333
	1.000	1.000	1.000	1.000	1.000

$f(x_1)f(y_2) = 0.895 \cdot 0.207 = 0.185$ but this is not equal to joint probability $f(x_1, y_2) = 0.196$ (see Table 5.14), which means that these random variables are not independent.

Example of Two Continuous Random Variables

The continuous random variables X and Y have the following joint density

$$f(x,y) = \begin{cases} \frac{x+3y}{2} & \text{for } 0 < x < 1, \text{ and } 0 < y < 1 \\ 0 & \text{otherwise} \end{cases},$$

with marginal distributions

$$f(x) = \begin{cases} \frac{x}{2} + \frac{3}{4} & \text{for } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$f(y) = \begin{cases} \frac{3y}{2} + \frac{1}{4} & \text{for } 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

In order to show the independence of continuous random variables, we have to verify that $f(x,y) = f(x)f(y)$:

$$f(x)f(y) = \left(\frac{x}{2} + \frac{3}{4}\right)\left(\frac{3y}{2} + \frac{1}{4}\right) = \frac{3}{4}xy + \frac{9}{8}y + \frac{1}{8}x + \frac{3}{16} \neq \frac{x+3y}{2} = f(x,y)$$

Since this equality does not hold these random variables are not independent (Figs. 5.16 and 5.17).

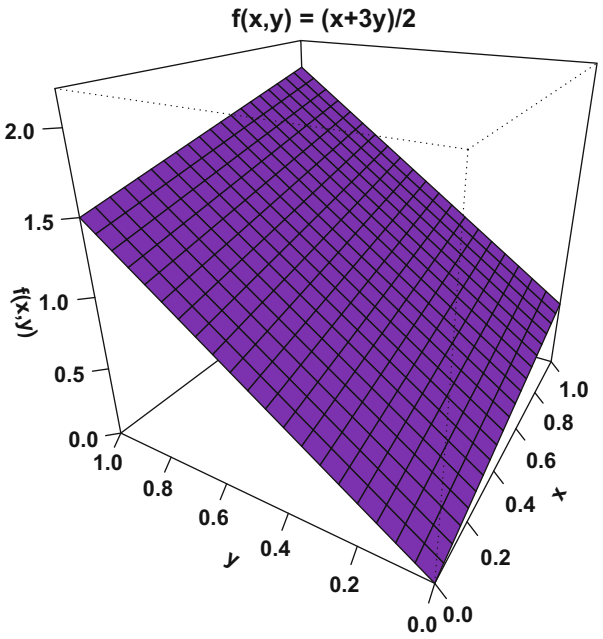


Fig. 5.16 The joint density $f(x; y)$

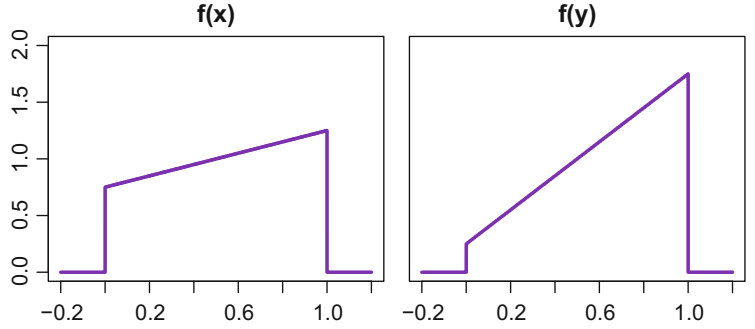


Fig. 5.17 The marginal distributions of X and Y: $f(x)$ and $f(y)$

Enhanced: Economic Conditions in Germany

In 1991, three thousand Germans were asked to express their opinion about current economic conditions in Germany. The responses to this question could take the values:

1–very good, 2—good, 3–reasonable, 4–bad, 5–very bad.

We define the random variable X as "current economic situation," which takes the values listed above. In addition to the reply to this question, the investigators also

Table 5.17 The current economic situation (X) and the residence of the respondent (Y) in 1991

Economic situation X		Residence Y West	Residence Y East	MD X
Very good	Observed	0.072	0.056	0.128
	Expected	0.063	0.065	
Good	Observed	0.257	0.204	0.461
	Expected	0.228	0.233	
Reasonable	Observed	0.151	0.227	0.378
	Expected	0.187	0.191	
Bad	Observed	0.012	0.014	0.026
	Expected	0.013	0.013	
Very bad	Observed	0.002	0.005	0.007
	Expected	0.003	0.004	
MD Y		0.494	0.506	1.000

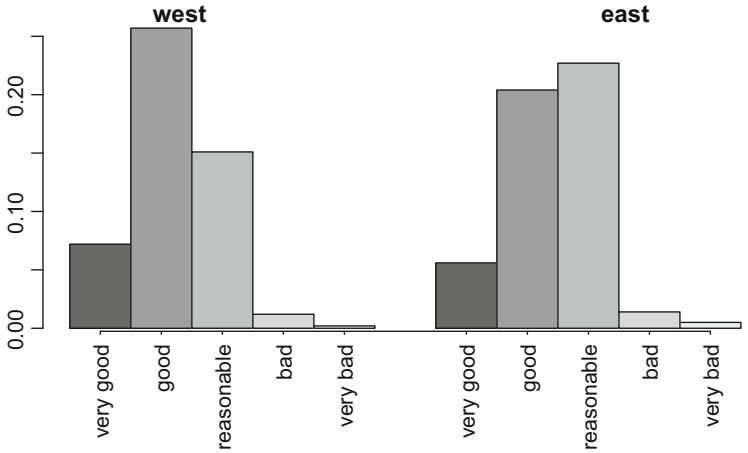


Fig. 5.18 The current economic situation (X) and the residence of the respondent (Y) in 1991

recorded the respondent's place of residence—the possible values of this variable were East (including the former East Berlin) and West (including the former West Berlin). This variable will be denoted as Y "residence" with possible values y_1—West, y_2—East.

The frequency distribution of these two random variables is provided in Table 5.17.

An interesting question to examine in this example is whether the assessment of the economic situation depends on the respondent's place of residence. Therefore, Table 5.17 contains also the probabilities calculated from the marginal distributions under the assumption of independence (i.e., $f(x_i, y_j) = f(x_i) \cdot f(y_j)$), these are denoted as "expected."

Figure 5.18 plots the joint distribution function of these random variables.

Table 5.18 Conditional distribution $f(y_j|x_i)$, rounded (1991)

Economic situation X	Residence Y West	Residence Y East	
Very good	0.563	0.437	1.000
Good	0.558	0.442	1.000
Reasonable	0.399	0.601	1.000
Bad	0.462	0.538	1.000
Very bad	0.286	0.714	1.000

Table 5.19 Current economic situation (X) and the place of residence of the respondent (Y) in 1996

Economic situation X		Residence Y West	Residence Y East	MD X
Very good	Observed	0.006	0.002	0.008
	Expected	0.05	0.003	
Good	Observed	0.082	0.036	0.118
	expected	0.078	0.040	
Reasonable	Observed	0.314	0.175	0.489
	Expected	0.323	0.166	
Bad	Observed	0.215	0.104	0.319
	Expected	0.211	0.108	
Very bad	Observed	0.044	0.022	0.066
	Expected	0.044	0.022	
MD Y		0.661	0.339	1.000

In order to determine whether these random variables were independent we also computed the conditional distribution. The conditional distribution, $f(y_j|x_i)$, is provided in Table 5.18.

Table 5.17 implies: A person from the West considers the economic situation "good" with probability 0.257. If the place of residence and the assessment of economic conditions were independent this probability would have to equal 0.228.

Table 5.18 implies: A person who considers the current economic situation as good is from the West with probability 0.558 and the East with probability 0.442. These probabilities also differ from the marginal distribution of Y in the last row of Table 5.17.

This indicates that the random variables X and Y are not independent, i.e., the assessment of the economic situation depends on the respondent's place of residence.

This survey was also conducted with another 3000 people in year 1996. The joint distribution from study is presented in Table 5.19, along with the values that would be obtained under independence (denoted as "expected"). Table 5.20 contains the conditional distributions. Figure 5.19 presents the joint distribution of these two random variables.

Using the data from 1996, there are also differences between the observed probabilities and the probabilities expected if the random variables were independent. In addition, the conditional distribution $f(y_j|x_i)$ differs from the marginal distribution

Table 5.20 Conditional distribution $f(y_j|x_i)$, rounded (1996)

	Residence Y		
Economic situation X	West	East	
Very good	0.750	0.250	1.000
Good	0.558	0.305	1.000
Reasonable	0.358	0.601	1.000
Bad	0.462	0.326	1.000
Very bad	0.667	0.333	1.000

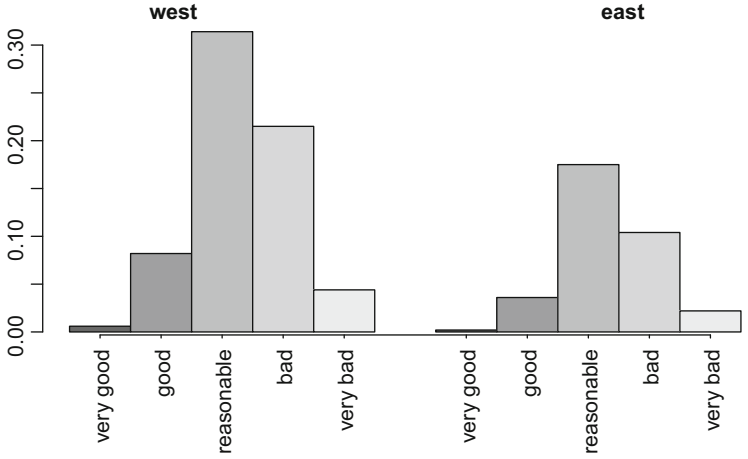


Fig. 5.19 The current economic situation (X) and the residence of the respondent (Y) in 1996

of Y. This suggests that we cannot conclude that the assessment of the economic condition is independent of the place of residence in 1996.

The conclusions concerning the independence of the random variables "economic situation" and "place residence" are valid only for the 3000 people included in the experiment! This example will be further examined in the chapter discussing the "χ^2-test of independence."

5.7 Parameters of Two-Dimensional Distributions

We can easily compute the expected values and variances of the marginal and conditional distributions—we simply use the formulas for the expected value and variance of one-dimensional random variable.

There are some other parameters that contain important information about the joint distribution of a pair of random variables. The most important of these are the covariance and the correlation coefficient.

Covariance

The covariance is based on the product of the differences of random variables X and Y from their expected values: $(X - E(X))(Y - E(Y))$. The covariance $Cov(X, Y)$ is defined as the expected value of this product:

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$$

The covariance measures the dependence between two random variables. Note: the covariance can be either positive or negative! In general, the covariance is not bounded. The following theorem is very important:

The covariance of two (stochastically) independent random variables X and Y is equal to zero. In general, the converse does not hold, i.e., zero covariance does not imply independence.

Correlation Coefficient

The correlation coefficient is used to evaluate the magnitude of the dependence between two random variables. We standardize the random variables X and Y in order to obtain a measure that is unit free:

$$\frac{[X - E(X)]}{\sigma_x}; \frac{[Y - E(Y)]}{\sigma_y}$$

The expected value of this product is called the correlation coefficient:

$$\rho(X, Y) = E\left[\frac{[X - E(X)]}{\sigma_x} \cdot \frac{[Y - E(Y)]}{\sigma_y}\right] = \frac{Cov(X, Y)}{\sigma_x \cdot \sigma_y}$$

for $\sigma_x > 0$, $\sigma_y > 0$. It can be shown that:

$$-1 \leq \rho(X, Y) \leq +1$$

Properties of the Correlation Coefficient

- The correlation coefficient and the covariance will have the same sign, because the standard deviation cannot be negative (square root of the variance)
- The correlation coefficient is always in the interval $[-1; +1]$
- The correlation coefficient measures degree of linear dependence between two random variables

- $|\rho(X, Y)| = 1$ if and only if X and Y fulfill

$$Y = a + bX,\ b \neq 0,\ \text{resp.}\ X = c + dY,\ d \neq 0$$

  for some a, b, c, d
- If X and Y are independent, then $\rho(X, Y) = 0$. Zero correlation does not imply independence. If $\rho(X, Y) = 0$, then we say that X and Y are uncorrelated. Two uncorrelated random variables can still be dependent, but the dependence is not linear.

More Information

Expected Values and Variances of Marginal Distributions

a) For two discrete random variables

$$\begin{aligned}
E(X) &= \sum_i \sum_j x_i \cdot f(x_i, y_j) = \sum_i x_i \sum_j f(x_i, y_j) = \sum_i x_i f(x_i), \\
E(Y) &= \sum_j \sum_i y_j \cdot f(x_i, y_j) = \sum_j y_j \sum_i f(x_i, y_j) = \sum_j y_j \cdot f(y_j), \\
Var(X) &= E[(X - E(X))]^2 = \sum_i \sum_j [x_i - E(X)]^2 f(x_i, y_j) \\
&= \sum_i [x_i - E(X)]^2 \sum_j f(x_i, y_j) \\
&= \sum_i [x_i - E(X)]^2 f(x_i) = \sum_i x_i^2 f(x_i) - [E(X)]^2, \\
Var(Y) &= E[(Y - E(Y))]^2 = \sum_j \sum_i [y_j - E(Y)]^2 f(x_i, y_j) \\
&= \sum_j [y_j - E(Y)]^2 \sum_i f(x_i, y_j) \\
&= \sum_j [y_j - E(Y)]^2 f(y_j) = \sum_j y_j^2 f(y_j) - [E(Y)]^2.
\end{aligned}$$

b) For two continuous random variables

$$\begin{aligned}
E(X) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xf(x,y)\,dx\,dy \\
&= \int_{-\infty}^{+\infty} x\left[\int_{-\infty}^{+\infty} f(x,y\,dy\right]dx = \int_{-\infty}^{+\infty} xf(x)\,dx, \\
E(Y) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} yf(x,y)\,dx\,dy \\
&= \int_{-\infty}^{+\infty} y\left[\int_{-\infty}^{+\infty} f(x,y)\,dx\right]dy = \int_{-\infty}^{+\infty} yf(y)\,dy, \\
Var(X) &= \int_{-\infty}^{+\infty} [x-E(X)]^2\cdot f(x)\,dx = \int_{-\infty}^{+\infty} x^2 f(x)\,dx - [E(X)]^2, \\
Var(Y) &= \int_{-\infty}^{+\infty} [y-E(Y)]^2\cdot f(y)\,dy = \int_{-\infty}^{+\infty} y^2 f(y)\,dy - [E(Y)]^2
\end{aligned}$$

Expected Values and Variances of Conditional Distribution

a) For two discrete random variables

$$\begin{aligned}
E(X|y_j) &= \sum_i x_i f(x_i|y_j), \\
E(Y|x_i) &= \sum_j y_j f(y_j|x_i), \\
Var(X|y_j) &= \sum_i [x_i - E(X|y_j)]^2 f(x_i|y_j) = \sum_i x_i^2 f(x_i|y_j) - [E(X|y_j)]^2, \\
Var(Y|x_i) &= \sum_j [y_j - E(y|x_i)]^2 f(y_j|x_i) = \sum_j y_j^2 f(y_j|x_i) - [E(Y|x_i)]^2.
\end{aligned}$$

b) For two continuous random variables

$$\begin{aligned}
E(X|y) &= \int_{-\infty}^{+\infty} xf(x|y)\,dx, \\
E(Y|x) &= \int_{-\infty}^{+\infty} yf(y|x)\,dx, \\
Var(X|y) &= \int_{-\infty}^{+\infty} [x-E(X|y)]^2\cdot f(x|y)\,dx, \\
Var(Y|x) &= \int_{-\infty}^{+\infty} [y-E(Y|x)]^2\cdot f(y|x)\,dy,
\end{aligned}$$

Covariance

Calculation of the covariance of X and Y:

a) X and Y discrete:

$$\begin{aligned} Cov(X,Y) &= \sum_i \sum_j [x_i - E(X)][y_j - E(Y)] f(x_i, y_j) \\ &= \sum_i \sum_j x_i y_j f(x_i, y_j) - E(X)E(Y) \end{aligned}$$

b) X and Y continuous:

$$\begin{aligned} Cov(X,Y) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} [x - E(X)][y - E(Y)] f(x,y)\, dx\, dy \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy f(x,y)\, dx\, dy - E(X)E(Y) \end{aligned}$$

The definition of the covariance immediately implies that the covariance of a random variable with itself is equal to the variance:

$$Cov(X,X) = E[(X - E(X))(X - E(X))] = E[(X - E(X))^2]\,.$$

From the definition of the covariance:

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$$

we obtain:

$$E(XY) = E(X)E(Y) + Cov(X,Y) \text{ if } X \text{ and } Y \text{ are dependent and}$$
$$E(XY) = E(X)E(Y) \text{ if } X \text{ and } Y \text{ are independent.}$$

Furthermore:

$$\begin{aligned} Var(XY) &= E\{[XY - E(XY)]^2\} \\ &= E\{(XY)^2 - 2XYE(XY) + E(XY)E(XY)\} \\ &= E[(XY)^2] - 2E(XY)E(XY) + E(XY)E(XY) \end{aligned}$$

$$\Longrightarrow$$

$$Var(XY) = E[(XY)^2] - [E(XY)]^2 \text{ for } X \text{ and } Y \text{ dependent and}$$
$$Var(XY) = E(X^2)E(Y^2) - [E(X)E(Y)]^2 \text{ for } X \text{ and } Y \text{ independent.}$$

Explained: Parameters of Two-Dimensional Random Variables

Example for Two Discrete Random Variables

The police department collected data on the number of mechanical problems (denoted by the random variable X) and the age of the cars, which is measured in years (denoted by the random variable Y). Only cars that are between 1 and 3 years were selected for further investigation. The joint and marginal density functions for these variables are given in Table 5.21.

The expected values and the variances of the marginal distributions are:

$$\begin{aligned} E(X) &= 0 \cdot 0.46 + 1 \cdot 0.3 + 2 \cdot 0.24 = 0.78, \\ Var(X) &= 0 \cdot 0.46 + 1 \cdot 0.3 + 4 \cdot 0.24 - 0.78^2 = 0.6516, \\ E(Y) &= 1 \cdot 0.6 + 2 \cdot 0.3 + 3 \cdot 0.1 = 1.5, \\ Var(Y) &= 1 \cdot 0.6 + 4 \cdot 0.3 + 9 \cdot 0.1 - 1.5^2 = 0.45. \end{aligned}$$

On average, a car has 0.78 mechanical problem(s), with a variance of deviance 0.65. The average age of a car is 1.5 years, with a variance of 0.45.

The covariance and the correlation coefficient are calculated as:

$$\begin{aligned} E(XY) &= 0 \cdot 1 \cdot 0.3 + 0 \cdot 2 \cdot 0.14 + 0 \cdot 3 \cdot 0.02 + 1 \cdot 1 \cdot 0.18 + 1 \cdot 2 \cdot 0.1 \\ &\quad + 1 \cdot 3 \cdot 0.02 + 2 \cdot 1 \cdot 0.12 + 2 \cdot 2 \cdot 0.06 + 2 \cdot 3 \cdot 0.06 \\ &= 1.28, \\ Cov(X, Y) &= 1.28 - 0.78 \cdot 1.5 = 0.11, \\ \rho(X, Y) &= 0.11/(0.6516 \cdot 0.45)^{0.5} = 0.2031. \end{aligned}$$

This means that the number of mechanical problems and the age of the car are positively correlated.

Table 5.21 Joint marginal distribution of X="Number of mechanical problems" and Y="Age of the car"

Number of (X)	Age (Y)			
Mechanical problems	1	2	3	MD X
0	0.30	0.14	0.02	0.46
1	0.18	0.10	0.02	0.30
2	0.12	0.06	0.06	0.24
MD Y	0.60	0.30	0.10	1.00

Example for Two Continuous Random Variables

Let us consider two continuous random variables X and Y with the joint density

$$f(x,y) = \begin{cases} \frac{x+3y}{2} & \text{for } 0 < x < 1, \text{ and } 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

and the marginal distributions

$$f(x) = \begin{cases} \frac{x}{2} + \frac{3}{4} & \text{for } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$f(y) = \begin{cases} \frac{3y}{2} + \frac{1}{4} & \text{for } 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

The expected values and variances are:

$$E(X) = \int_0^1 x\left(\frac{x}{2} + \frac{3}{4}\right) dx = \left[\frac{x^3}{6} + \frac{3x^2}{8}\right]_0^1 = \frac{1}{6} + \frac{3}{8} = \frac{13}{24}$$

$$E(Y) = \int_0^1 y\left(\frac{3y}{2} + \frac{1}{4}\right) dy = \left[\frac{y^3}{2} + \frac{y^2}{8}\right]_0^1 = \frac{1}{2} + \frac{1}{8} = \frac{5}{8}$$

$$Var(X) = \int_0^1 x^2\left(\frac{x}{2} + \frac{3}{4}\right) dx + \left(\frac{13}{24}\right)^2$$

$$= \left[\frac{x^4}{8} + \frac{x^3}{4}\right]_0^1 + \left(\frac{13}{24}\right)^2 = \frac{3}{8} + \frac{169}{576} = 0,\,6684$$

$$Var(Y) = \int_0^1 y^2\left(\frac{3y}{2} + \frac{1}{4}\right) dy + \left(\frac{5}{8}\right)^2$$

$$= \left[\frac{3y^4}{8} + \frac{y^3}{12}\right]_0^1 + \left(\frac{5}{8}\right)^2 = \frac{11}{24} + \frac{25}{64} = 0,\,849$$

The covariance:

$$Cov(X,Y) = \int_0^1 \int_0^1 xy\left(\frac{x+3y}{2}\right) dx\,dy - \left(\frac{13}{24}\right)\left(\frac{5}{8}\right)$$

$$= \frac{1}{2}\int_0^1 \int_0^1 (x^2y + 3xy^2)\,dx\,dy - \left(\frac{13}{24}\right)\left(\frac{5}{8}\right)$$

$$
\begin{aligned}
&= \frac{1}{2}\int_0^1 \left[\frac{x^2y^2}{2} + xy^3\right]_0^1 dx - \left(\frac{13}{24}\right)\left(\frac{5}{8}\right) \\
&= \frac{1}{2}\int_0^1 \left(\frac{x^2}{2} + x\right) dx - \left(\frac{13}{24}\right)\left(\frac{5}{8}\right) \\
&= \left[\frac{x^3}{6} + \frac{x^2}{2}\right]_0^1 dx - \left(\frac{13}{24}\right)\left(\frac{5}{8}\right) \\
&= \frac{1}{3} - \frac{65}{192} = -\frac{1}{192}
\end{aligned}
$$

And the correlation coefficient:

$$\rho(X, Y) = \frac{-\frac{1}{192}}{\sqrt{0,6684 \cdot 0,849}} = -0,007$$

Enhanced: Investment Funds

An investment advisor offers a client two investment funds: Securia (S) and Technoinvest (T). The expected return is usually taken as a measure of profitability and the variance (or, equivalently, the standard deviation) is measure of the risk. The expected return may be related with future economic conditions. To examine the portfolio weightings on these investments, which contain different levels of risk, we need to consider the correlation between the expected profits. The investment advisor offers the following probabilities for three possible scenarios for the path of the economy (1—no change, 2—recession, 3—growth) and, depending on the state of the economy, an estimate of the expected return on the investment funds S(ecuria) and T(echnoinvest) (Table 5.22).

Expected return for the two investment funds over possible states of the world:

$$E(S) = 3.5 \cdot 0.5 + 4 \cdot 0.3 + 2 \cdot 0.2 = 3.35\%$$

$$E(T) = 5 \cdot 0.5 - 1 \cdot 0.3 + 7 \cdot 0.2 = 3.6\%$$

Table 5.22 Expected returns of two investment projects

Scenario	Probability	Expected return S (%)	Expected return T (%)
1	0.5	3.5	5.0
2	0.3	4.0	−1.0
3	0.2	2.0	7.0

Variance for these investment funds:

$$\begin{aligned} Var(S) &= (3.5-3.35)^2 \cdot 0.5 + (4-3.35)^2 \cdot 0.3 + (2-3.35)^2 \cdot 0.2 \\ &= 0.5025, \\ \sigma(S) &= 0.7089\,\% \end{aligned}$$

$$\begin{aligned} Var(T) &= (5-3.6)^2 \cdot 0.5 + (-1-3.6)^2 \cdot 0.3 + (7-3.6)^2 \cdot 0.2 \\ &= 9.64, \\ \sigma(T) &= 3.1048\,\% \end{aligned}$$

The variability of the expected return, as well as the risk, is larger for the investment fund Technoinvest (T).

Now we calculate the covariance of the expected returns for these funds:

$$\begin{aligned} Cov(S,T) = &(3.5-3.35)(5-3.6) \cdot 0.5 + (4-3.35)(-1-3.6) \cdot 0.3 \\ &+(2-3.35)(7-3.6) \cdot 0.2 = -1.71. \end{aligned}$$

And we obtain the correlation coefficient:

$$\rho(S,T) = -1.71/(0.7089 \cdot 3.1048) = -0.7769.$$

The expected returns on the funds based on the scenarios provided by the investment advisor are negatively correlated.

The expected return of portfolio Z depends on the weights attached to each of the funds. Using the weights a and b ($a+b=1$), we obtain:

$$\begin{aligned} E(Z) &= aE(S) + bE(T), \\ Var(Z) &= a^2 Var(S) + b^2 Var(T) + 2abCov(S,T) \\ &= a^2 Var(S) + b^2 Var(T) + 2ab \cdot \sigma(S) \cdot \sigma(T) \cdot \rho(S,T). \end{aligned}$$

If we know the risk associated with both investment funds, the risk of the portfolio is decreasing if the two funds are negatively correlated. We can now calculate $E(Z)$ and $Var(Z)$ for given a and b, e.g., $a = b = 0.5$:

$$\begin{aligned} E(Z) &= 0.5 \cdot 3.35 + 0.5 \cdot 3.6 = 3.475, \\ Var(Z) &= 0.25 \cdot 0.5025 + 0.25 \cdot 9.64 - 2 \cdot 0.5 \cdot 0.5 \cdot 0.7089 \cdot 3.1048 \cdot 0.7769 \\ &= 1.6806, \\ \sigma(Z) &= 1.296. \end{aligned}$$

For different values, e.g., $a = 0.8$ $b = 0.2$, we obtain:

$$\begin{aligned} E(Z) &= 0.8 \cdot 3.35 + 0.2 \cdot 3.6 = 3.4, \\ Var(Z) &= 0.64 \cdot 0.5025 + 0.04 \cdot 9.64 - 2 \cdot 0.8 \cdot 0.2 \cdot 0.7089 \cdot 3.1048 \cdot 0.7769 \\ &= 0.16, \\ \sigma(Z) &= 0.4. \end{aligned}$$

The risk is much smaller if we invest 80 % in Securia (S) and 20 % in Technoinvest (T), instead of an equal weighting in both funds. In addition, the expected return of these two portfolios is equal. The risk of this portfolio is also smaller than the risk associated with the safer fund.

Chapter 6
Probability Distributions

6.1 Important Distribution Models

In the following section we present some important probability distributions, which are often used in statistics. These distributions can be described using at most three parameters. In general, the greater the number of parameters describing a distribution, the more flexible the distribution will be to model data.

6.2 Uniform Distribution

Discrete Uniform Distribution

A discrete random variable X with a finite number of outcomes x_i $(i = 1, 2, \ldots, n)$ follows a uniform distribution, if each value of X can occur with an **equal** probability, which depends on n.

The probability density function of a uniform random variable is:

$$f(x_i) = \begin{cases} \frac{1}{n} \text{ for } i = 1, \ldots, n \\ \\ 0 \text{ otherwise} \end{cases}$$

The distribution function for a uniform random variable is:

$$F(x) = \begin{cases} 0 \text{ for } x < x_1 \\ \\ \frac{i}{n} \text{ for } x_i \leq x \leq x_{i+1}; i = 1, \ldots n-1 \\ \\ 1 \text{ for } x_n \leq x \end{cases}$$

W.K. Härdle et al., *Introduction to Statistics*, DOI 10.1007/978-3-319-17704-5_6

The expected value and variance of discrete uniform random variable X are:

$$E(X) = \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$Var(X) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Continuous Uniform Distribution

A continuous random variable X on the interval $[a, b]$ is uniformly distributed if the density function assigns equal values to each point in that interval. Hence, the density function will have the following form:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

The **distribution function** for a continuous uniform random variable is:

$$F(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } a \le x \le b \\ 1 & \text{for } b \le x \end{cases}$$

The expected value and variance of continuous uniform random variables are:

$$E(X) = \frac{b+a}{2}$$

$$Var(X) = \frac{(b-a)^2}{12}$$

The parameters of a continuous uniform distribution are a and b.

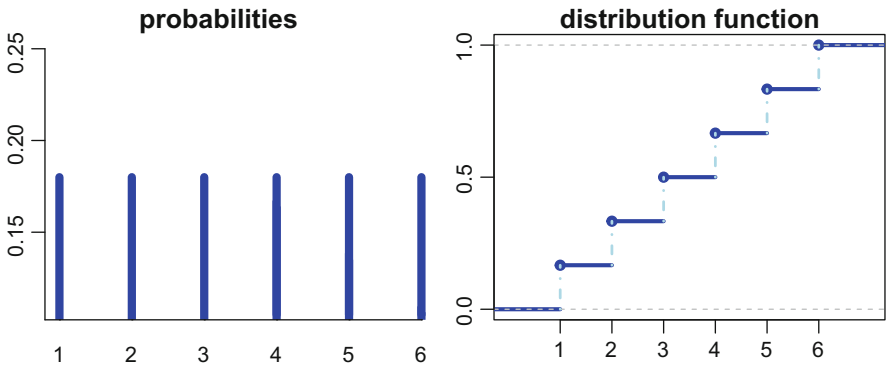


Fig. 6.1 Density and distribution function of a discrete uniform random variable on the integers between 1 and 6

More Information

Discrete Uniform Distribution

The probability density function of discrete Uniform random variable can be illustrated with a bar chart. The distribution function of this random variable, on the other hand, will be a step function (Fig. 6.1).

A common example of a discrete uniform random variable is the outcomes associated with the roll of a fair die. The discrete random variable X (= result of the throw) can take integer numbers between 1 and 6. If the dice are "fair," the probability of each outcome of X is $f(x_i) = 1/6, i = 1, \ldots, 6$.

Continuous Uniform Distribution

Let us verify whether

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

is a density function: First, $b > a$, so $f(x) \ge 0$ for all x, i.e., the function is nonnegative. Furthermore we have:

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{a}^{b} \frac{1}{b-a}\,dx = \left[\frac{x}{b-a}\right]_a^b = \frac{b-a}{b-a} = 1.$$

This indicates that $f(x)$ is a density. The distribution function $F(x)$ can be computed as:

$$F(x) = \int_a^x \frac{1}{b-a}\, dv = \left[\frac{v}{b-a}\right]_a^x = \frac{x-a}{b-a}$$

The expected value and the variance for this random variable are:

$$\begin{aligned} E(X) &= \int_a^b x \frac{1}{b-a}\, dx = \left[\frac{x^2}{2(b-a)}\right]_a^b = \frac{b^2-a^2}{2(b-a)} \\ &= \frac{(b-a)(b+a)}{2(b-a)} = \frac{(b+a)}{2} \\ Var(X) &= \int_a^b x^2 \frac{1}{b-a}\, dx - \left(\frac{(b+a)}{2}\right)^2 = \left[\frac{x^3}{3(b-a)}\right]_a^b - \left(\frac{(b+a)}{2}\right)^2 \\ &= \frac{b^3-a^3}{3(b-a)} - \frac{b+a}{4} = \frac{(b-a)^2}{12} \end{aligned}$$

Figure 6.2 illustrates the density and distribution function of a continuous uniform random variable.

Explained: Uniform Distribution

A man arrives at a tram stop, but does not know the schedule of the tram. The tram arrives at that stop every 20 min. Define the random variable X: "waiting time for

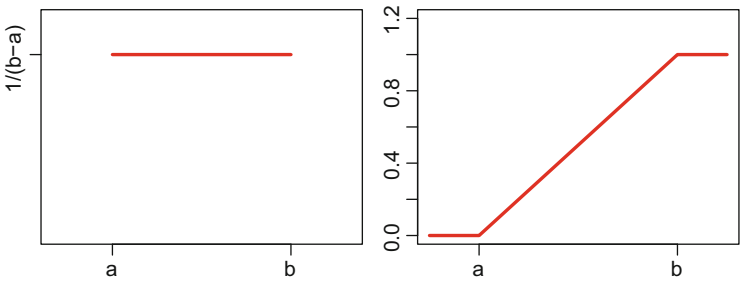


Fig. 6.2 Density (*left*) and distribution function (*right*) of a continuous uniform random variable between a and b

a tram in minutes." This random variable can take any value in the interval $[0, 20]$. This implies: $P(0 \leq X \leq 20) = 1, a = 0, b = 20$.

The random variable X="waiting time" will have a uniform distribution.
Density of X:

$$f(x) = \begin{cases} \frac{1}{20} \text{ for } 0 < x \leq b \\ \\ 0 \quad \text{otherwise} \end{cases}$$

Distribution function:

$$F(x) = \begin{cases} 0 \qquad \text{for } x < 0 \\ \\ \frac{1}{20} \cdot x \text{ for } 0 \leq x \leq 20 \\ \\ 1 \qquad \text{otherwise} \end{cases}$$

The expected value of X is:

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx = \int_{0}^{20} x\frac{1}{20}\,dx$$
$$= \frac{1}{20}\left[\frac{1}{20}x^2\right]_0^2 0 = \frac{1}{20}\left[\frac{1}{2}20^2 - \frac{1}{2}0^2\right] = 10$$

On average a person will have to wait 10 min for a tram.
The variance is:

$$Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx = \int_{0}^{20} (x-10)^2 \cdot \frac{1}{20}\,dx$$
$$= \frac{1}{20}\int_{0}^{20} (x^2 - 20x + 100)\,dx$$
$$= \frac{1}{20}\left[\frac{1}{3}x^3 - \frac{1}{2}20x^2 + 100x\right]_0^{20}$$
$$= \frac{1}{20}\left[\frac{1}{3}20^3 - \frac{1}{2}20^3 + 100 \cdot 20\right] = 33.33.$$

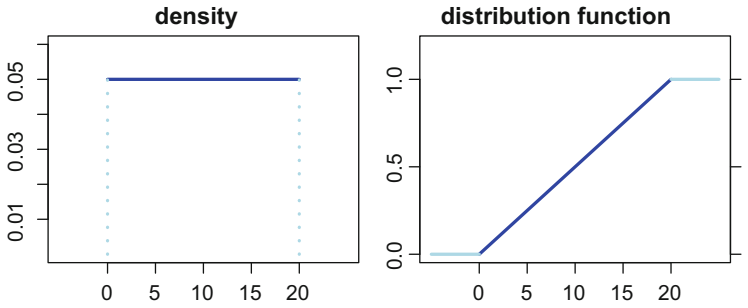


Fig. 6.3 Density and distribution function of a continuous uniform random variable between 0 and 20

Since $Var(X) = 33.33$, the standard deviation is given by $\sigma = \sqrt{Var(X)} = 5.77$ (Fig. 6.3).

6.3 Binomial Distribution

A binomial distribution is derived from a random experiment in which we either obtain event A with constant probability p, or the complementary event $\bar{A}$ with probability $1 - p$.

Suppose this experiment is repeated n times.

A discrete random variable that contains the number of successes A after n repetitions of this experiment has a binomial distribution with parameters n and p. Its probability density function is:

$$f(x; n, p) = \begin{cases} \binom{n}{x} \cdot p_x \cdot (1-p)^{n-x} & \text{for } x = 0, 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

We write $X \sim B(n; p)$. The distribution function is given as:

$$F(x; n, p) = \begin{cases} 1 & \text{for } x > n \\ \sum_{k=0}^{x} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} & \text{for } n \geq x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

The expected value and the variance of a binomial distribution $B(n;p)$ are:

$$E(X) = n \cdot p$$
$$Var(X) = n \cdot p \cdot (1-p)$$

The **properties** of the binomial distribution include:

- *Reproduction property:*
  If $X \sim B(n;p)$ and $Y \sim B(m;p)$ are independent random variables, then the random variable $Z = X + Y$ has binomial distribution with parameters $n+m$ and p, i.e., $Z \sim B(n+m;p)$.
- *Symmetry:*
  If $X \sim B(n;p)$ and $Y = n - X$ then $Y \sim B(n;1-p)$.
  The binomial distribution has been tabulated for selected values of the parameters n and p ($p \leq 0.5$).

More Information

Derivation of the Binomial Distribution

The random experiment can be described by the following properties:

- Only two events, A and $\bar{A}$, are possible.
- The probabilities of these events are $P(A) = p$ and $P(\bar{A}) = 1-p$.
- The experiment is repeated n times, the repetitions are mutually independent, and the probabilities are constant.

Each component of this experiment is called Bernoulli experiment. For each Bernoulli experiment, we define the random variable, $X_i (i = 1, \ldots, n)$, which takes the values 0 (if we obtain event $\bar{A}$) and 1 (if we obtain the event A). The probabilities for the events in this experiment will be $P(A) = p$ and $P(\bar{A}) = 1-p$ and the random variable X_i has the following probability function (i.e., Bernoulli distribution):

$$f(x;p) = \begin{cases} p^x(1-p)^{1-x} & \text{for} x = 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

$$E(X_i) = p, \quad Var(X_i) = p(1-p)$$

After repeating the Bernoulli experiment n times, we obtain the number of occurrences of the event A, i.e., we observe random variable X={number of occurrence of event A in n trials}:

$$X = \sum_{i=1}^{n} X_i$$

X is a function (linear combination) of n random variables. The event $X = x$ occurs if and only if the event A is observed exactly x times and event $\bar{A}$ is observed $(n - x)$ times in the n trials, e.g.,

$$A_1 \cap A_2 \cap \cdots \cap A_x \cap \bar{A}_{x+1} \cap \bar{A}_{x+2} \cap \cdots \cap \bar{A}_n$$
$$| \quad x - timesA \quad | \quad (n-x) - times\bar{A} \quad |$$

The index of the event shows the number of trials. The independence of the Bernoulli experiments means that the probability that $X = x$ is

$$\begin{aligned} f(x) = P(X = x) &= P(A_1 \cap A_2 \cap \cdots \cap A_x \cap \bar{A}_{x+1} \cap \bar{A}_{x+2} \cap \cdots \cap \bar{A}_n) \\ &\cdot P(A_1) \cdot P(A_2) \cdot \cdots \cdot P(A_x) \cdot P(\bar{A}_x + 1) \cdot P(\bar{A}_x + 2) \cdot \cdots \cdot P(\bar{A}_n) \\ &= p \cdot p \cdot \cdots \cdot p \cdot (1-p) \cdot (1-p) \cdot \cdots \cdot (1-p) \\ &= p^x \cdot (1-p)^{n-x} \end{aligned}$$

This probability is computed only for the specified ordering of the event A. The probability of this specific ordering is $f(x) = p^x \cdot (1-p)^{n-x}$. The number of different orderings of these events is denoted as binomial coefficient and it is computed as:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Notice that the different orderings are disjoint events. Hence, we obtain the following probability function:

$$P(X = x) = f(x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$$

The binomial distribution is discrete, the probability function can be displayed as a histogram, and the distribution function as a step function. The following diagrams illustrate the density function for various values of p, holding n constant.

For $p < 0.5$ the distribution is skewed to the left. The skew is greater for smaller values of p. The distribution is symmetric for $p = 0.5$, with np being the center of the distribution. For $p > 0.5$ the diagrams are skewed to the right. For large values of n, we can approximate this density function using a normal distribution with parameters $\mu = np$ and $\sigma^2 = np(1-p)$. The quality of approximation improves the closer p is to 0.5. The approximation follows from the Central Limit Theorem, which will be explained later.

Explained: Drawing Balls from an Urn

There are 10 balls in a box, 3 are white and 7 are red;

$$A = \{\text{white ball}\} \rightarrow P(A) = 0.3;$$
$$\bar{A} = \{\text{red ball}\} \rightarrow P(\bar{A}) = 0.7.$$

After each draw, we return the ball to the box. We draw five balls ($n = 5$) in total. The assumptions of a Bernoulli experiment are obviously fulfilled:

- There are only 2 possible outcomes for each draw
- The probabilities associated with each outcome are constant because we return the balls into the box
- The draws are mutually independent

We want to compute the probability of drawing two white balls, i.e., $P(X = 2)$.

$$X_i = \{\text{number of white balls in draw}i\}$$

Then, $P(X_i = 1) = 0.3$ and $P(X_i = 0) = 0.7$ for all $i = 1, \ldots, 5$. Using five repetitions, we obtain the following random variables: X_1, X_2, X_3, X_4, X_5. Consider:

$$X = \{\text{number of white balls from } n = 5 \text{ draws}\}$$
$$X = \sum_i X_i$$
$$X \sim B(n; p) = B(5; 0.3)$$

The number of all possible permutations of the draws when we select 2 white and 3 red balls is:

$$\binom{5}{2} = \frac{5!}{2! \cdot 3!} = 10$$

The probability is:

$$P(X = 2) = f_B(2; 5; 0.3) = \binom{5}{2} \cdot 0.3^2 \cdot 0.7^3 = 0.3087$$

Table 6.1 contains the density and the distribution function of the binomial distribution for this experiment. Figure 6.4 shows the probability distribution function $B(5; 0, 3)$. The probability of a certain event can be calculated using the

Table 6.1 Binomial distribution with $n = 8$ and $p = 0.3$

x	$f_B(x; 5; 0.3)$	$F_B(x; 5; 0.3)$
0	0.1681	0.1681
1	0.3601	0.5282
2	0.3087	0.8369
3	0.1323	0.9692
4	0.0284	0.9976
5	0.0024	1.0000

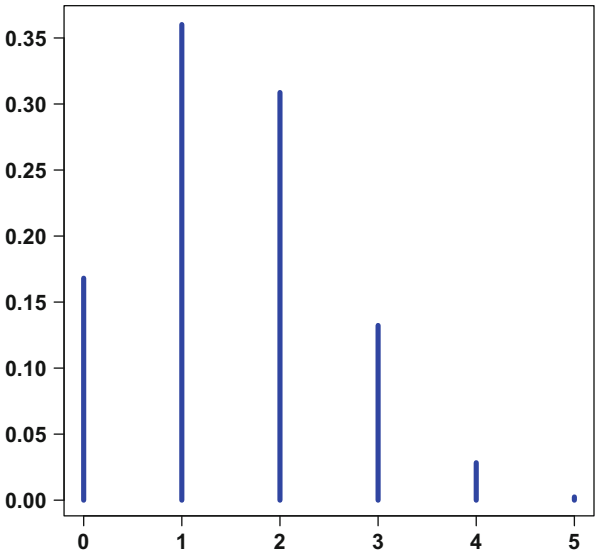


Fig. 6.4 Probability distribution function $B(5; 0, 3)$

distribution function:

$$\begin{aligned} f_B(2; 5; 0.3) &= F_B(2; 5; 0.3) - F_B(1; 5; 0.3) \\ &= 0.8369 - 0.5282 = 0.3087 \end{aligned}$$

The probability that we draw 2 white balls in 5 trials is equal to 0.3087.

Enhanced: Better Chances for Fried Hamburgers

A TV commercial for Hamburger-Land contained following sentence: "Our research showed that 75 % of people prefer fried hamburgers." In the same commercial, the announcer also said: "If you ask four Hamburger-Land customers,

Table 6.2 Binomial distribution with $n = 4$ and $p = 0.25$

x	$f_B(x; 4; 0.25)$	$F_B(x; 4; 0.25)$
0	0.3164	0.3164
1	0.4219	0.7383
2	0.2109	0.9492
3	0.0469	0.9961
4	0.0039	1.0000

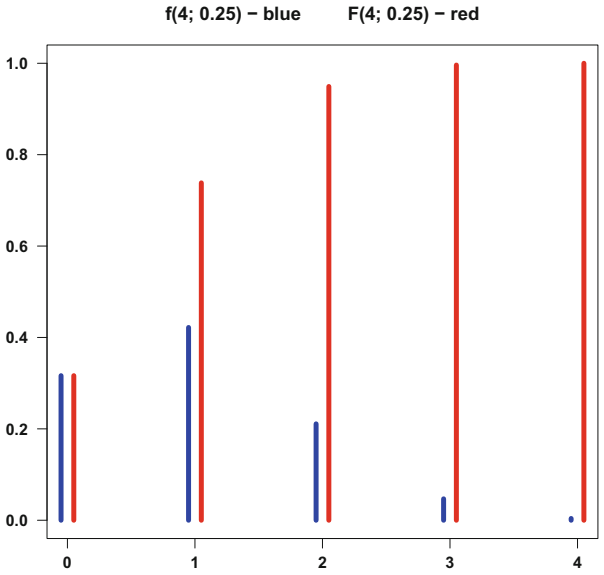


Fig. 6.5 Density (*blue*) and distribution function (*red*) of the binomial distribution with $n = 4$ and $p = 0.25$

at most one of them would choose nonfried hamburger." Are these sentences saying exactly the same?

The assumptions of a Bernoulli experiment are satisfied. The outcome of each experiment can take one of only two values: $A = \{\text{nonfried hamburger}\}$ and $\bar{A} = \{\text{fried hamburger}\}$ with probabilities $P(A) = 0.25$ and $P(\bar{A}) = 0.75$.

A sample of customer can be very large. Therefore, it is not important whether the sampling is done with or without "replacement." The probabilities associated with each outcome can be considered to be constant and the experiments independent.

Define the random variable $X = \{\text{number of nonfried hamburgers in 4 decisions}\}$, which has a binomial distribution with parameters $n = 4$, $p = 0.25$; i.e., $X \sim B(4; 0.25)$. The probability $P(X \leq 1)$ can be computed as

$$P(X \leq 1) = P(X = 0) + P(X = 1) = F_B(1; 4; 0.25).$$

The probability that the event "nonfried hamburger" occurs at most once is the sum of the probabilities that the "nonfried hamburger" will be chosen by none or by

only one customer out of four randomly chosen customers of Hamburger-Land. In other words, it is the value of the distribution function of the binomial distribution at $x = 1$.

The binomial distribution with $n = 4$ and $p = 0.25$ is summarized in Table 6.2.

The last column of the table implies that $F_B(1; 4; 0.25) = 0.7383$. Assuming that the probabilities for fried (P(fried hamburger) $= 0.75$) and non-fried hamburgers (P(nonfried hamburger) $= 0.25$) are accurate, the statement from the commercial is correct with probability 0.7383 (Fig. 6.5).

Enhanced: Student Jobs

Students from a university (HU Berlin) completed a questionnaire. 65 % of the students responded that they have a part time job. What is the probability that at most 4 out of 8 randomly chosen students from this university have a part time job?

The assumptions of a Bernoulli experiment are satisfied. Each "experiment" can produce only two outcomes: $A = \{$ student has a part time job $\}$; $\bar{A} = \{$ student does not have a part time job $\}$, $P(A) = 0.65$; $P(\bar{A}) = 0.35$.

We assume that the sample number of students is large compared with the number of all students, which makes it possible to use a binomial distribution. The probabilities associated with the events can be considered to be constant and the responses of the students are independent (the probability of choosing one student two times is very close to zero).

The outcome of this experiment is the random variable $X = \{$ number of students with a part time job$\}$. This random variable has a binomial distribution: $X \sim B(n; p) = B(8; 0.65)$. We need to compute the probability $P(X \leq 4)$, i.e., the distribution function $F(4)$.

The value of the distribution function $B(8; 0.65)$ is not tabulated. The calculation of the distribution function by hand would be very difficult, since we would have to calculate and then sum up five probabilities $f(x)$, $x = 0, 1, \ldots, 4$. Therefore, we evaluated the distribution function numerically (see the second column of Table 6.3):

Table 6.3 Binomial distributions with $n = 8$, $p = 0.65$ and $n = 8$, $p = 0.35$

x	$B(8; 0.65)$	$B(8; 0.35)$
0	0.0002	0.0319
1	0.0036	0.1691
2	0.0253	0.4278
3	0.1061	0.7064
4	0.2936	0.8939
5	0.5722	0.9747
6	0.8309	0.9964
7	0.9681	0.9998
8	1.0000	1.0000

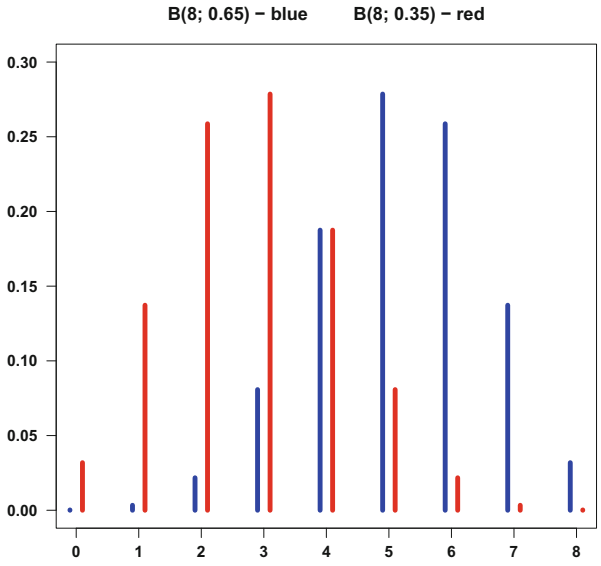


Fig. 6.6 Density functions for $B(8; 0.35)$ and $B(8; 0.65)$

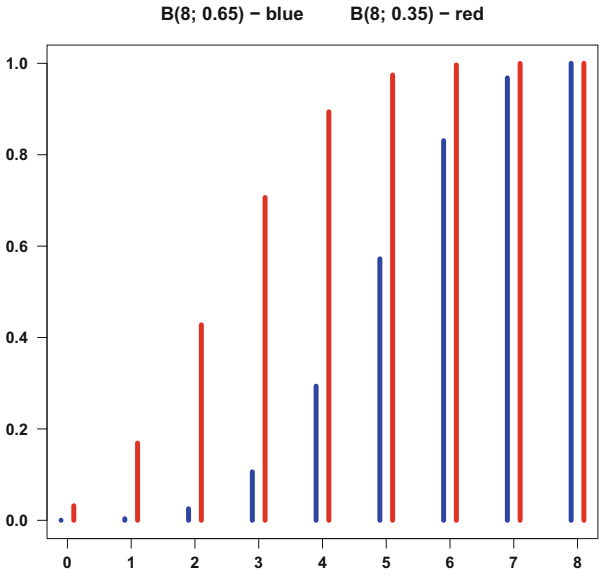


Fig. 6.7 Distribution functions for $B(8; 0.35)$ and $B(8; 0.65)$

The probability that at most 4 students from $n = 8$ randomly chosen students will have a part time job is equal to 0.2936 (Fig. 6.6).

If you are unable to evaluate the distribution numerically, it is possible to use the tabulated values of the binomial distribution and the symmetry of the binomial distribution to obtain the probabilities we require (Fig. 6.7). Consider:

$$X = \{\text{number of students with a part time job}\} \sim B(8; 0, 65),$$

$$Y = \{\text{number of students without a part time job}\} \sim B(8; 0, 35).$$

Then, $X \leq 4$, i.e., $x \in \{0, 1, 2, 3, 4\}$ corresponds to $Y \geq 4$, i.e., $y \in \{4, 5, 6, 7, 8\}$. Instead of computing the probability $P(X \leq 4)$, we can compute $P(Y \geq 4) = 1 - P(Y \leq 3)$. Using the table for a binomial distribution, we find in the third column $P(Y \leq 3) = 0.7064$ and this implies that

$$P(Y \leq 4) = 1 - 0.7064 = 0.2936.$$

Interactive: Binomial Distribution

The binomial distribution depends on the parameters n and p that determine

- its shape
- its location, i.e., expected value $E(X) = np$, and
- its variance, i.e., $\sigma = \sqrt{np(1-p)}$

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please select

- the number of draws n
- the probability of success per draw p

Moreover, choose one of the following functions:

- Probability mass function
- Cumulative distribution function

Output

This interactive example allows you to change either one or both parameters of the distribution. The plot in Fig. 6.8 displays the probability distribution function (or cumulative distribution function) of the binomial distribution $B(n; p)$.

We recommend to only change value of one parameter at a time, to explore the effect of this change on the probability plot.

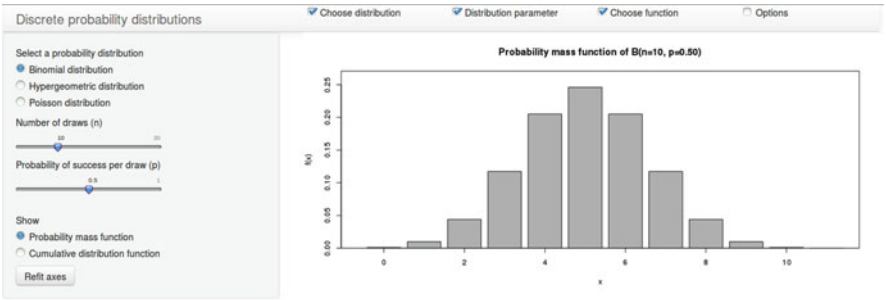


Fig. 6.8 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_bin

6.4 Hypergeometric Distribution

The Hypergeometric distribution is based on a random event with the following characteristics:

- Total number of elements is N
- From these N elements, M elements have a certain property of interest, and $N-M$ elements do not have this property, i.e., only two events, A and $\bar{A}$ are possible
- We randomly choose n elements out of N without replacement

This means that the probability $P(A)$ is not constant and the draws (events) are not independent in this sort of experiment. The random variable X, which contains the number of successes A after n repetitions of the experiment, follows a hypergeometric distribution with parameters N, M, and n, with probability density function:

$$f_H(x; N, M, n) = \begin{cases} \dfrac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}} & \text{for } x = a, \ldots, b \\ 0 & \text{otherwise} \end{cases}$$

$$\text{where } a = \max[0, n - (N - M)], \text{and } b = \min[n, M]$$

Shorthand notation is: $X \sim H(N, M, n)$. The expected value and the variance of the hypergeometric distribution $H(N, M, n)$ are:

$$E(X) = n \cdot \frac{M}{N}$$

$$Var(X) = n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right) \cdot \frac{N-n}{N-1}$$

More Information

Like the binomial distribution, the hypergeometric distribution is based on an experiment with only two possible outcomes.

The hypergeometric distribution differs from the binomial distribution in that we draw without replacement, which means the draws from the hypergeometric distribution are not independent. This implies that the number of objects in the pool is decreasing with each draw and furthermore that $n \leq N$.

In addition, the number of outcomes with property A also changes and this, in turn, changes the probability of drawing an object with property A.

Explanation of the Probability Function

- Assuming n draws, we are interested in the total number of objects with the property A, i.e., the random variable $X =$ "number of outcomes with the property A among n draws."

  The order of the drawn objects is not important. Using combinatorics, we can calculate the number of possible combinations in which we can draw n out of N objects without replacements:

$$\binom{N}{n}$$

- How many different ways are there to obtain $X = x$ objects with property A? We have $x \leq M$, i.e., we cannot draw more objects with property A than there are in total (no repetition). Again, keep in mind that the order in which these objects are drawn is not of interest. The total number of combinations of drawing x objects with property A out of M is:

$$\binom{M}{x}$$

  Analogously, $n - x \leq N - M$, and the number of possible combinations to draw $n - x$ objects without property A out of $N - M$ objects is:

$$\binom{N-M}{n-x}$$

  The number of possible combinations to draw x objects with property A and $n-x$ objects without property A is given by the product of the two previous terms:

$$\binom{M}{x} \cdot \binom{N-M}{n-x}$$

The desired probability can be obtained using the classical (Laplace) definition of the probability as the ratio:

$$P(X = x) = f(x) = \frac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}}.$$

Determining the Range of Values of X

The largest possible value of X is n for $n \leq M$, and M for $M < n$. This implies that:

$$x_{\max} = \min(n; M).$$

The smallest possible value of X is: $x \geq 0$ (it can never be smaller than that. No surprise!). If n is greater than the number of elements without the property A, then we have that $x \geq n - (N - M)$. This implies that:

$$x_{\min} = \max[0; n - (N - M)].$$

The Expected Value and the Variance

Let $M/N = p$, we have the following:

$$E(X) = n \cdot \frac{M}{N} = n \cdot p$$

$$Var(X) = n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right) \cdot \frac{N-n}{N-1} = n \cdot p \cdot (p-1) \cdot \frac{N-n}{N-1}$$

The distribution $H(M, N, n)$ will have the same expected value as the corresponding binomial distribution $B(n, M/N)$. However, its variance will be smaller because it is multiplied by the ratio $(N - n)/(N - 1)$ because drawing without replacement implies that we cannot use anymore the information we start with initially. The constant $(N - n)/(N - 1)$ is called a continuity correction.

The probability function of the hypergeometric distribution is illustrated in Fig. 6.9. We choose the following parameters for this example: $N = 100$, $M = 20$, $n = 10$ and $N = 16$, $M = 8$, $n = 8$.

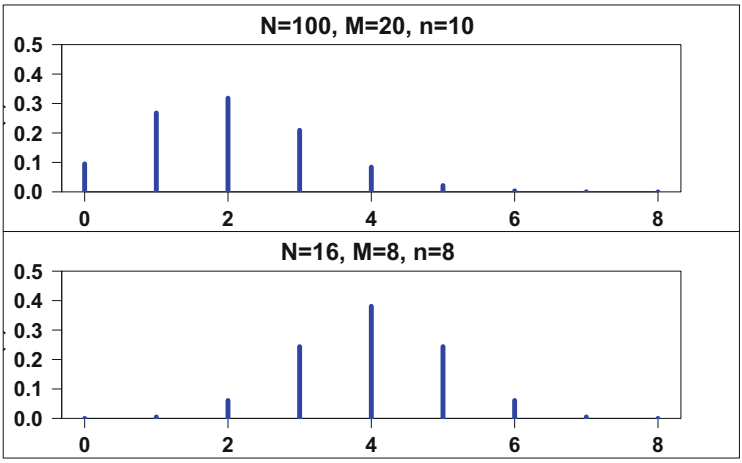


Fig. 6.9 Two probability functions for the hypergeometric distribution

Explained: Choosing Test Questions

A student has to complete a test with ten questions. The student must answer three randomly chosen questions. The student knows that six of the ten questions are so difficult that no one has a chance to answer them:

- $N = 10$ questions
- $M = 4$ questions have property A, they can be answered
- $n = 3$ randomly chosen questions the student must answer
- X = "number of questions with property A between n randomly chosen questions"

Possible values of X are: $\max[0, n-(N-M)] \leq x \leq \min(n, M)$, i.e., $0 \leq X \leq 3$. Motivation of the use of hypergeometric distribution:

- finite number of questions,
- returning (repeating) of the questions does not make any sense in this situation,
- hence, the draws are not independent,
- this implies that $P(A)$ depends on the previously drawn questions.

What is the probability that the student draws three "good" questions?

$$f_H(3; 10, 4, 3) = \frac{\binom{4}{3} \cdot \binom{10-4}{3-3}}{\binom{10}{3}} = \frac{4 \cdot 1}{120} = \frac{1}{30}$$

What is the probability that the student chooses at least one question that he can answer? $P(X \geq 1) = 1 - P(X = 0)$:

$$P(X = 0) = f_H(0; 10, 4, 3) = \frac{\binom{4}{0} \cdot \binom{10-4}{3-0}}{\binom{10}{3}} = \frac{1 \cdot 20}{120} = \frac{1}{6}$$

It follows that:

$$P(X \geq 1) = 1 - \frac{1}{6} = \frac{5}{6}.$$

Enhanced: Selling Life Insurances

An insurance agent arrives in a town and sells 100 life insurances: 40 are term life policies and the remaining 60 are permanent life policies. He chooses (randomly and without returning) five life insurance policies. What is the probability that he chooses exactly two term life policies.

There are $N = 100$ policies. The outcomes of this experiment (type of the insurance policy) can take one of two values: the term life type (property A) with $M = 40$ and the permanent life type (complementary event) with $N - M = 60$.

The random variable X is defined as "number of the term life policies in five randomly chosen insurance policies." The random variable X is based on random sampling without replacement and so follows a hypergeometric distribution $H(N; M; n) = H(100; 40; 5)$.

The smallest value of X is $0 = (\max[0, n - (N - M)])$, i.e., none of the five randomly chosen contracts is a term life policy. The largest possible value of X is $n < M$, i.e., 5. The set of possible values of X is such that $0 \leq \mathrm{x} \leq 5$.

We need to compute the value of the probability function for $x = 2$, i.e., $P(X = 2) = f_H(2; 100; 40; 5)$:

$$f_H(2; 100, 40, 5) = \frac{\binom{40}{2} \cdot \binom{100-40}{5-2}}{\binom{100}{5}} = \frac{\frac{40!}{2!\cdot 38!} \cdot \frac{60!}{3!\cdot 57!}}{\frac{100!}{5!\cdot 95!}} = 0.3545$$

Suppose we increase the number of draws (randomly chosen contracts) to $n = 10$. The only thing that would change in the example is the range of the random variable X, which becomes $0 \leq x \leq 10$. The random variable X now has the following hypergeometric distribution $H(100; 40; 10)$.

If we compute the probability that there are exactly 4 term life policies in the 10 randomly chosen policies, i.e., $P(X = 4)$, we obtain the following result:

$$f_H(4; 100, 40, 10) = \frac{\binom{40}{4} \cdot \binom{100-40}{10-4}}{\binom{100}{10}} = 0.2643$$

Enhanced: Insurance Contract Renewal

An insurance agent knows from experience that 70 % of his clients renew their contracts. Suppose this agent has 20 clients. What is the probability that at least one half of four randomly chosen clients will renew their contract?

We have total of $N = 20$ clients. Of these clients, $M = 14$ clients renew their policies (property A) and $N - M = 6$ clients do not. The experiment has only two possible outcomes.

We choose $n = 4$ clients randomly. Clearly, it does not make sense to model this random variable with replacement.

The random variable X is defined as "number of clients who renew their contract." X has hypergeometric distribution, $H(N; M; n) = H(20; 14; 4)$. The smallest possible value of X is $0 = (\max[0, n - (N - M)])$, i.e., none of the 4 clients renew their contracts. Since $n < M$, the largest possible value of X in this example is 4. Hence, X can take the following values: $0 \leq \mathrm{x} \leq 4$.

We need to find the probability $P(X \geq 2)$, which can be computed as: $P(X = 2) + P(X = 3) + P(X = 4)$.

$$f_H(2; 20, 14, 4) = \frac{\binom{14}{2} \cdot \binom{20-14}{4-2}}{\binom{20}{4}} = \frac{91 \cdot 15}{4845} = 0.2817$$

$$f_H(3; 20, 14, 4) = \frac{\binom{14}{3} \cdot \binom{20-14}{4-3}}{\binom{20}{4}} = \frac{364 \cdot 6}{4845} = 0.4508$$

$$f_H(4; 20, 14, 4) = \frac{\binom{14}{4} \cdot \binom{20-14}{4-4}}{\binom{20}{4}} = \frac{1001 \cdot 1}{4845} = 0.2066$$

This implies that: $P(X \geq 2) = 0.2817 + 0.4508 + 0.2066 = 0.9391$. The probability that at least two of the four chosen clients (out of the 20 clients) decides to renew their policy is 0.9391.

Interactive: Hypergeometric Distribution

A hypergeometric distribution depends on parameters N, M, and n. These parameters influence its shape, location, and variance.

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please select the

- population size N
- number of success states in the population M
- number of draws n

Moreover, choose one of the following functions:

- Probability mass function
- Cumulative distribution function

Output

This interactive example allows you to change either one or more parameters of the distribution. The plot in Fig. 6.10 displays the probability distribution function (or cumulative distribution function) of the hypergeometric distribution function $H(N; M; n)$.

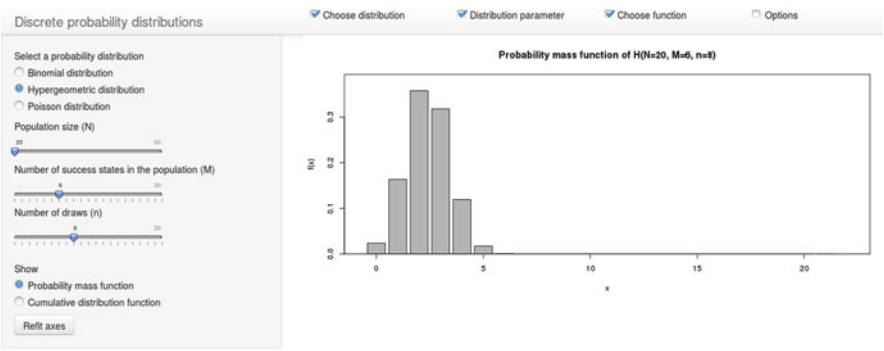


Fig. 6.10 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_hyp

We suggest that you only change the value of one parameter, holding the others constant, which will better illustrate the effects of the parameters on the shape of the hypergeometric distribution.

6.5 Poisson Distribution

The Poisson distribution can describe an experiment in which an event can be observed a number of times (e.g., accidental deaths).

The random variable X denotes the number of occurrences and is discrete in nature. This random variable will be described by a probability density function referred to as a Poisson distribution with parameter λ:

$$f_{PO}(x;\lambda) = \begin{cases} \frac{\lambda^x}{x!}e^{-\lambda} & \text{for } x = 0, 1, 2, \ldots; \lambda > 0 \\ 0 & \text{otherwise} \end{cases}$$

The distribution function is:

$$F_{PO}(x;\lambda) = \begin{cases} \sum_{k=0}^{x} \frac{\lambda^x}{x!}e^{-\lambda} & \text{for } k \geq 0; \lambda > 0 \\ 0 & \text{for } k \leq 0 \end{cases}$$

The expected value and variance of the Poisson distribution are:

$$E(X) = \lambda \quad Var(X) = \lambda\,.$$

Properties of the Poisson Distribution

- Reproductivity: Consider two independent variables $X \sim PO(\lambda_1)$ a $Y \sim PO(\lambda_2)$, then the random variable Z = X+Y is Poisson distributed with parameter $\lambda_1 + \lambda_2$: $Z \sim PO(\lambda_1 + \lambda_2)$
- Poisson distribution for an arbitrary interval length: If the number of occurrences in a unit interval is Poisson distributed, then the number of occurrences in an interval of length t units will also be Poisson distributed with parameter λt:

$$f_{PO}(x;\lambda \cdot t) = \frac{(\lambda t)^x}{x!}e^{-\lambda t}$$

More Information

The following are some examples for the application of the Poisson distribution:

- The number of printing defaults per page in books.
- The number of twining breaks of a weaving machine in a given interval of time.
- The number of received calls at a telephone center.
- The number of vehicles that drive past an intersection per minute.
- The number of patients arriving at an emergency department per hour.
- The number of alpha-particles emitted by a radioactive substance in a specific time interval.
- The number of fish caught during a day.
- The number of reported accidents to an insurance firm per year.
- The number of bank customers applying for credit in a month.

The following assumptions are needed.

- The possibility of occurrence is always based on an interval. The use of an appropriate scale will ascertain that the given size is made up of continuous interval units.
- The occurrence of an outcome is purely random in the sense that it cannot be predetermined.
- The independence of the outcomes means that an occurrence (or nonoccurrence) of an outcome cannot influence the occurrence of the same outcome in another trial. Subsequently the number of outcomes in 2 disjoint intervals are independent.
- Two outcomes cannot occur at the same time, i.e., in any arbitrary interval, the possibility of obtaining more than one outcome should be 0.
- The "intensity" of occurrence of an outcome must be constant with a parameter $\lambda > 0$, i.e., the average number of outcomes in an interval must be independent of the interval chosen. Consequently, the probability of occurrence in a specific interval will only be dependent on the size of the interval.

If these assumptions are true, then the variable is described by a Poisson process. The Poisson distribution can also be derived using a binomial distribution using the following assumptions:

- The number of trials; n, is large.
- The probability of occurrence of an outcome A, $P(A) = p$, in a single trial is very small.
- $E(X) = np = \lambda$, then with increasing number of trials n; $((n \to \infty))$, p will approach zero ($p \to 0$).

Consequently, the Poisson distribution PO(λ = np) can be used to approximate a binomial distribution. With large n and small p the Poisson distribution is often referred to as the distribution of rare occurrences.

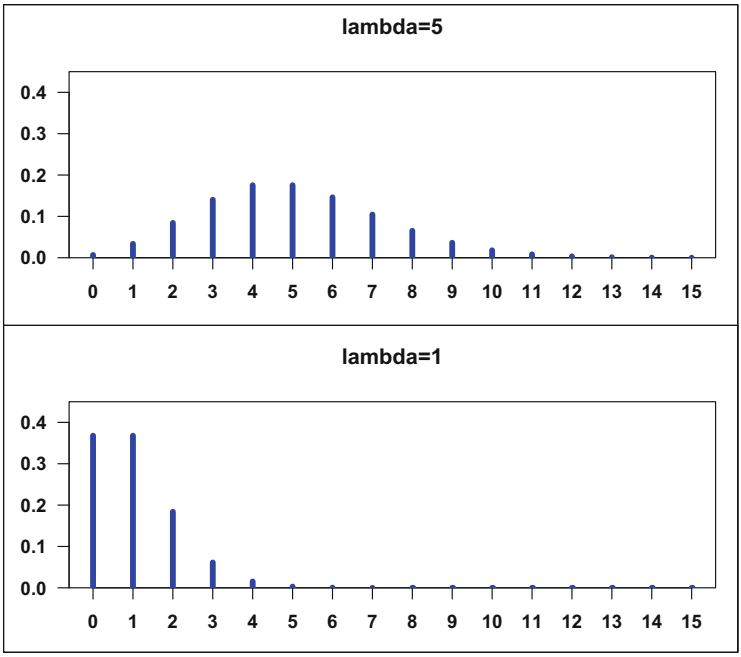


Fig. 6.11 Poisson probability density functions for $\lambda = 5$ and $\lambda = 1$

As a rule of thumb, the approximation of a binomial distribution by a Poisson distribution requires $n > 30$ and $p \leq 0.05$.

Figure 6.11 presents a plots of Poisson probability density functions for $\lambda = 5$ and $\lambda = 1$. The smaller the value of λ, the more the Poisson distribution is skewed to the left. However, as λ increases density function becomes more symmetric.

Explained: Risk of Vaccination Damage

A town has 20,000 inhabitants who need to be vaccinated. The probability that the vaccine provokes an adverse reaction in an inoculated person is 0.0001.

In fact, this is a Bernoulli experiment, where:

1. A = "Occurrence of adverse effect" and $\bar{A}$ = "No adverse effects from vaccine",
2. $P(A) = 0.0001$ is constant, and
3. Independence of trials, i.e., of vaccinations.

To obtain the probabilities for the number of adverse reactions, the binomial distribution could be used. However, the small probability associated with an outcome and the large number of trials suggest that the Poisson distribution could be used as an approximation, since $n > 30$ and $p \leq 0..05$. This approximation rule

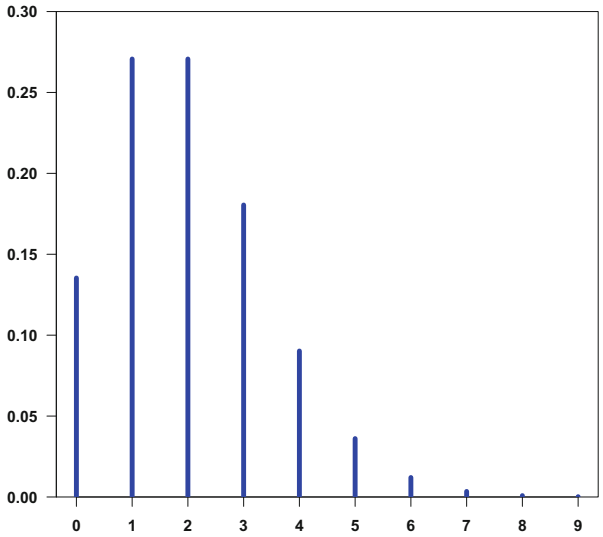


Fig. 6.12 Probability density function $PO(2)$

of thumb will be explained later in the text. We use the following parameter:

$$\lambda = np = 20000 \cdot 0.0001 = 2$$

This is the expected number of cases with adverse reactions. The probability density function $f_{PO(2)}$ is plotted in Fig. 6.12.

- The probability that no one suffers adverse effects is $P(X = 0) = P(X \leq 0) = F(0) = 0.1353$.
- The probability that one person has a bad reaction to the vaccination is: $P(X = 1) = P(X \leq 1) - P(X \leq 0) = F(1) - F(0) = 0.2707$.
- The probability that more than 4 persons have adverse effects is: $P(X > 4) = 1 - F(4)$. The value of $F(4)$ can be found in the tables for a Poisson distribution for $\lambda = 2$ and $X = 4$: $F(4) = 0.9473 \Rightarrow P(X > 4) = 1 - 0.9473 = 0.0527$.

Enhanced: Number of Customers in Service Department

Through experience, the customer service department of a major supermarket knows that it receives on average 1 customer per hour between 9 a.m. and 2 p.m., and 2 customers per hour between 2 p.m. and 7 p.m. Since a request for service from any customer can be considered to be random, as well as independent of other customer requests, the random variable X_1 = "number of customers per hour between 9 a.m. and 2 p.m." will follow a Poisson distribution with parameter $\lambda_1 = 1$.

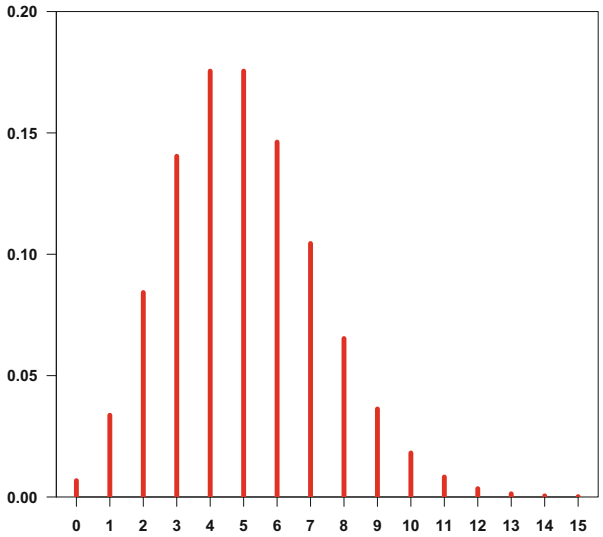


Fig. 6.13 Probability density function $PO(5)$

The random variable X_2 = "number of customers per between 2 p.m. and 7 p.m." will also follow a Poisson distribution, but with parameter $\lambda_2 = 2$. Notice that for both time intervals, t = 5.

Using this information, we can compute the probability of having a certain number of customers between 9 a.m. and 2 p.m. Let's denote this variable by Y_1. For example, if $Y_1 = 6$, we get:

$$P(Y_1 = 6) = f_{PO}(6; 1 \cdot 5) = \frac{(\lambda_1 t)^y}{y!} e^{-\lambda_1 t} = \frac{(1 \cdot 5)^6}{6!} e^{-1 \cdot 5} = 0.1462$$

The probability of having more than 4 customers at the customer department will be (Fig. 6.13):

$$\begin{aligned} P(Y_1 > 4) &= 1 - P(Y_1 \leq 4) = 1 - e^{-5} \left(\frac{5^0}{0!} + \frac{5^1}{1!} + \frac{5^2}{2!} + \frac{5^3}{3!} + \frac{5^4}{4!} \right) \\ &= 1 - 0.4405 = 0.5595. \end{aligned}$$

We can also obtain the probabilities for the number of customers between 2 p.m. and 7 p.m., denoted Y_2 (Fig. 6.14). For $Y_2 = 6$ or $Y_2 > 4$:

$$P(Y_2 = 6) = f_{PO}(6; 2 \cdot 5) = \frac{(\lambda_2 t)^x}{x!} e^{-\lambda_2 t} = \frac{(2 \cdot 5)^6}{6!} e^{-2 \cdot 5} = 0.063$$

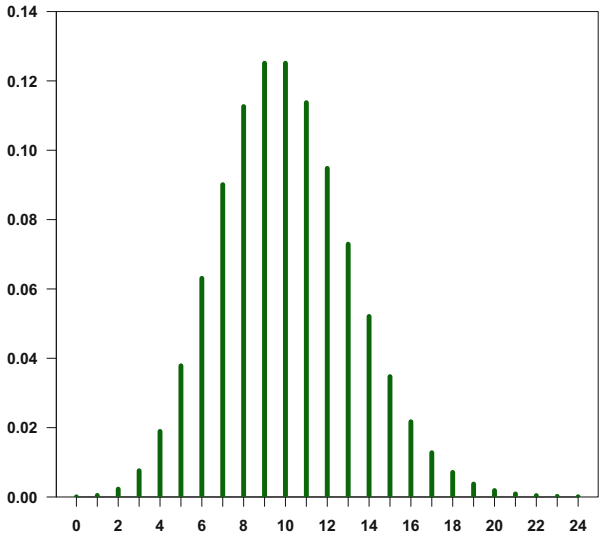


Fig. 6.14 Probability density function $PO(10)$

$$P(Y_2 > 4) = 1 - P(Y_2 \leq 4) = 1 - e^{-10}\left(\frac{10^0}{0!} + \frac{10^1}{1!} + \frac{10^2}{2!} + \frac{10^3}{3!} + \frac{10^4}{4!}\right)$$
$$= 1 - 0.0293 = 0.9707$$

Y_1 and Y_2, are independent. Using the above results, we can obtain the probability of receiving more than 4 customers between 9 a.m. and 2 p.m. and 2 p.m. and 7 p.m. as follows:

$$P(Y_1 > 4, Y_2 > 4) = P(Y_1 > 4) \cdot P(Y_2 > 4) = 0.5595 \cdot 0.9707 = 0.5431.$$

To obtain the total number of customers between 9 a.m. and 7 p.m., we create the random variable $Y = Y_1 + Y_2$. Since Y_1 and Y_2 are independent, Y will also have a Poisson distribution with parameter $\lambda_1 t + \lambda_2 t = 5 + 10 = 15$.

Interactive: Poisson Distribution

The Poisson distribution is completely described by the parameter λ, which influences its shape, position, and variance.

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please select a value for the parameter λ.

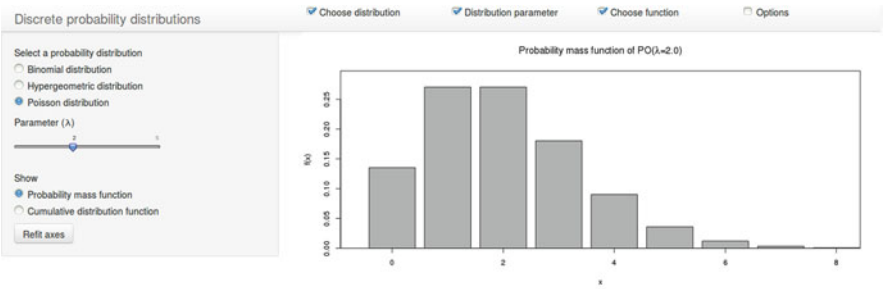


Fig. 6.15 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_poi

Moreover, choose one of the following functions:

- Probability mass function
- Cumulative distribution function

Output

This interactive example allows you to explore the effect of the parameter λ on shape of this distribution. The plot in Fig. 6.15 displays the probability distribution function (or cumulative distribution function) of the Poisson distribution function $PO(\lambda)$.

6.6 Exponential Distribution

A continuous random variable X follows an exponential distribution with parameter $\lambda > 0$ if its probability density function can be defined as:

$$f_{EX}(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x =\geq 0; \lambda > 0 \\ 0 & \text{for } x < 0 \end{cases}$$

This is denoted as $X \sim EX(\lambda)$. The distribution function is given as:

$$F_{EX}(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & \text{for } x \geq 0; \lambda > 0 \\ 0 & \text{for } x < 0 \end{cases}$$

The expected value and variance of an exponentially distributed random variable are:

$$E(X) = \frac{1}{\lambda} \quad Var(X) = \frac{1}{\lambda^2}$$

As $\lambda \to \infty$, the faster the density function approaches 0 and the distribution function approaches 1.

More Information

The Poisson distribution is used to compute the probability associated with a random variable Y that is defined as the number of occurrences of a certain event within a specified continuous time interval with an intensity λ.

But if we are interested in the time between these occurrences, an exponential distribution can be used to make probability statements. The exponential distribution provides the probability of the "distance" between two subsequent Poisson random events. We denote this new continuous random variable by X="the time interval between 2 subsequent events."

The probability that X takes on a maximum value of x is $P(X \leq x) = 1 - P$ (no outcome within the interval of length x). But P (no outcome within the interval of length x) simply represents the probability that a Poisson distributed random variable Y with the interval of length x takes on a value of 0, $P(Y = 0)$ so that:

$$f_{PO}(y; \lambda x) = \frac{(\lambda x)^y}{y!} e^{-\lambda x}$$

$$P(Y = 0) = f_{PO}(0; \lambda x) = \frac{(\lambda x)^0}{0!} e^{-\lambda x} = e^{-\lambda x}.$$

We obtain the distribution function of the exponential distribution, i.e., X is exponentially distributed:

$$P(X \leq x) = 1 - e^{-\lambda x}.$$

Therefore, there exists a relationship between the exponential and Poisson distributions. The exponential distribution is often used to model the length of time for continuous processes as well as waiting times.
For example:

- The waiting time before service in a restaurant, bank, or filling station.
- The time taken before a component within a technical system fails.
- Service time (time to load a truck, time to carry out a repair).
- Half life (life span) of a component (person).

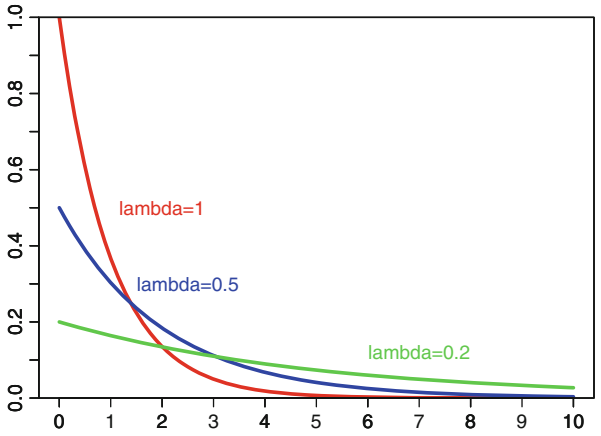


Fig. 6.16 Density functions of the exponential distribution for different parameter values

- Time taken for a telephone conversation.
- Time taken before the next report on damages at property insurance firm.

The following condition is often associated with an exponential distribution.

$$P(X \leq t + s | X \geq t) = P(X \leq s).$$

This condition means that the time associated with an outcome does not depend on previous times. We say that the exponential distribution is memoryless.

The graphical presentation of an exponentially distributed random variable will be given in the form of a density function, since it refers to the case of a continuous random variable (Fig. 6.16).

Explained: Number of Defects

On the basis of the relationship between the exponential and the Poisson distribution, the Poisson distribution defines the probability of the number of outcomes Y of a specific phenomenon, in a fixed and continuous length or interval with the intensity λ.

The following example illustrates the relationship between Poisson and exponential distribution. Suppose there is a machine for which 2 defects, on average, are recorded per week. Let t denote the number of weeks.

- The probability that no defects are recorded in a week is:
  Y_1= "number of defects in one week" ($t = 1$)

$$E(Y_1) = \lambda = 2\,, \qquad Y_1 \sim PO(2)$$

$$f_{PO}(y_1;\lambda) = \frac{(\lambda t)^{y_1}}{y_1!}e^{-\lambda x} = \frac{(2\cdot 1)^0}{0!}e^{-2\cdot 1} = e^{-2} = 0.1353$$

- The probability of recording no defects in two weeks: Y_2= "number of defects in two weeks" ($t = 2$)

$$E(Y_2) = \lambda t = 2\cdot 2\,, \qquad Y_2 \sim PO(4)$$

$$P(Y_2 = 0) = \frac{(2\cdot 2)^0}{0!}e^{-2\cdot 2} = \frac{4^0}{0!}e^{-4} = e^{-4} = 0.0183$$

- In general, the probability that no defect is recorded in t weeks is:
  Y= "number of defects in t weeks."

$$E(Y) = \lambda t \qquad Y \sim PO(\lambda t)$$

$$P(Y = 0) = \frac{(\lambda t)^0}{0!}e^{-\lambda t} = e^{-\lambda t}$$

- If we are interested in finding the probability associated with the time until the next defect occurs, for example, the probability that the next defect occurs in more than two weeks: X= "Waiting time till next defect."
  To calculate $P(X > 2)$ we use the exponential distribution:

$$P(X > 2) = 1 - P(X \le 2) = 1 - F_{EX}(x;\lambda) = 1 - (1 - e^{-\lambda x})$$
$$= e^{-\lambda x} = e^{-2\cdot 2} = 0.0183$$

  This value is the same as the probability $P(Y_2 = 0)$ from the Poisson distribution, for the random variable Y="in 2 weeks no defects is recorded (Fig. 6.17)."

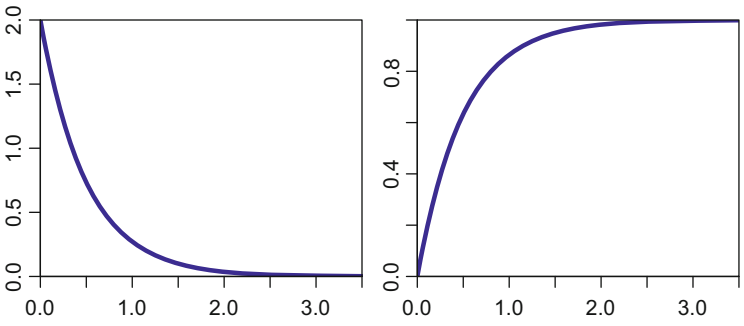


Fig. 6.17 The probability density function (*left*) and distribution function (*right*) of $EX(2)$

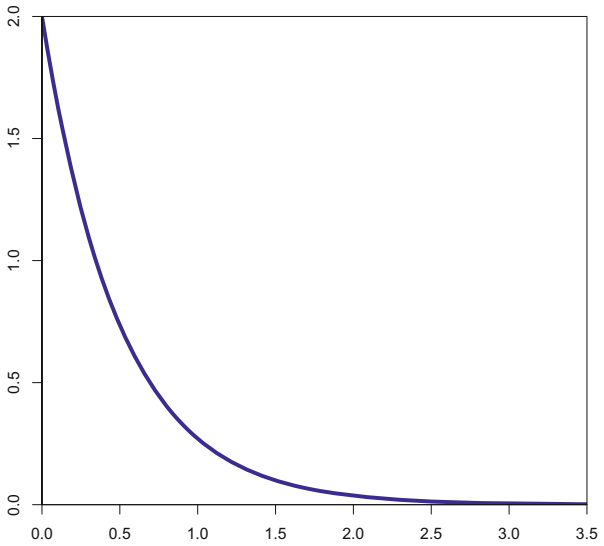


Fig. 6.18 Probability density function for $EX(2)$

Enhanced: Equipment Failures

In a manufacturing plant, 48 equipment failures are expected per day (=24 h). These failures are purely randomly and independent. On average, λ=48/24 =2 failures are expected per hour. Define the random variable T for the time between 2 failures, which will have an exponential distribution: $T \sim EX(2)$. The probability density function for $EX(2)$ is displayed in Fig. 6.18.

The probability the next equipment failure will occur in two 2 h is:

$$P(t > 2) = 1 - F_{EX}(2) = 1 - (1 - e^{-2 \cdot 2}) = e^{-4} = 0.01832$$

Suppose that a plant uses two of these systems. The plant comes to a halt as soon as one of the systems stops to function. Let T_1 = "Time between 2 failures for the first component." T_2 = "Time between 2 failures for the second component."

$T_1 \sim EX(2)$ and $T_2 \sim EX(2)$

Since the plant can only function while both components are operating, both will need more than 2 h to operate.

$$\begin{aligned} & P(\text{The system operates for more than 2 h}) \\ = \; & P[(\text{first component operates for more than 2 h}) \\ & \cap P(\text{second component operates for more than 2 h})] \\ = \; & P(\text{first component operates for more than 2 h}) \end{aligned}$$

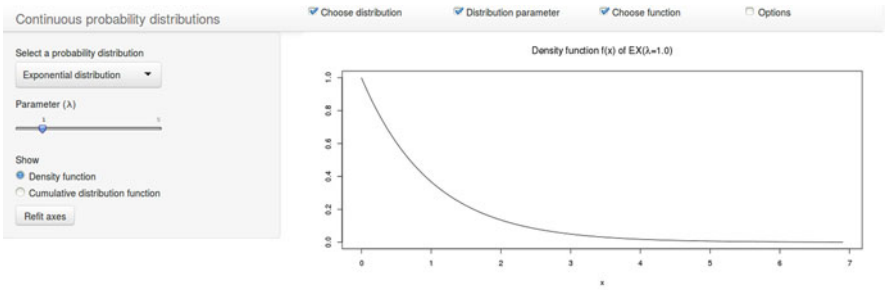


Fig. 6.19 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_exp

$$\cdot P(\text{second component operates for more than 2 h})$$
$$= P(T_1 \geq 2) \cdot P(T_2 \geq 2) = (0.01832)^2 = 0.000336$$

We use the multiplicative property for independent outcomes here since both components operate independently.

Interactive: Exponential Distribution

The exponential distribution is completely described by the parameter λ.

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please select a value for the parameter λ.

Moreover, choose one of the following functions:

- Probability mass function
- Cumulative distribution function

Output

This interactive example allows you to explore the effect of the parameter λ on shape of the distribution. The plot in Fig. 6.19 displays the probability distribution function (or cumulative distribution function) of the exponential distribution.

6.7 Normal Distribution

A continuous random variable X is normally distributed with parameters μ and σ, denoted $X \sim N(\mu, \sigma)$, if and only if its **density** function is:

$$f_N V(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad -\infty < x < +\infty, \sigma > 0$$

the distribution function is :

$$F_N V(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(t-\mu)^2/2\sigma^2}\, dt$$

The normal distribution depends on two parameters μ and σ, which are the expected value and the standard deviation of the random variable X.

The **expected value**, **variance**, and **standard deviation** are given by:

$$E(X) = \mu = \int_{-\infty}^{+\infty} x f(x)\, dx, \quad Var(X) = \sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)\, dx, \quad \sigma = \sqrt{\sigma^2}$$

Two important **properties** of normally distributed random variables are:

- Linear transformation

  Let X be normally distributed, $X \sim N(\mu,\sigma)$ and Y be a linear combination of X: $Y = a + bX\,, b \neq 0$. Then, the random variable Y also follows a normal distribution:

$$Y \sim N(a + b\mu, |b| \cdot \sigma)$$

  The values of the parameters of the transformed random variable follow from the rules for calculating with expected values and variances:

$$E(a + bX) = a + b \cdot E(X),$$
$$Var(a + bX) = b^2 Var(X) = b^2\sigma^2.$$

- Reproduction property

  Let us consider n random variables $X_1, X_2 \ldots, X_n$ with normal distributions: $X_i \sim N(\mu_i, \sigma_i), E(X_i) = \mu_i, Var(X_i) = \sigma_i^2$.

  The sum of independent, normally distributed random variables $X_1, \ldots, X_n$, i.e.,

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n, \quad a_i \neq 0,$$

  for at least one i, is again normally distributed.

$$Y = \sum_{i=1}^{n} a_iX_i \sim N\left(\sum_{i=1}^{n} a_i\mu_i, \sum_{i=1}^{n} a_i^2\sigma_i^2\right)$$

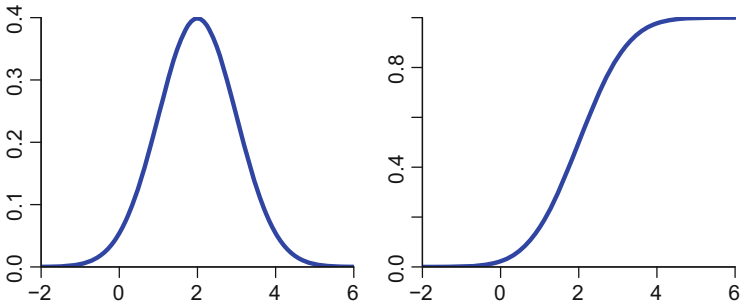


Fig. 6.20 Density (*left*) and distribution function (*right*) of $N(2; 1)$

Figure 6.20 displays a density and distribution function for a $N(2; 1)$ random variable.

Standardized Random Variable

Imagine a random variable X with $E(X) = \mu$ and $Var(X) = \sigma^2$. If we transform this variable in the following way

$$Z = \frac{X - \mu}{\sigma},$$

then random variable Z denotes a standardized random variable, which has been shifted by its mean and scaled by its standard deviation. We have $E(Z) = 0$ and $Var(Z) = 1$. If X is normally distributed, then Z also follows a normal distribution.

Standard Normal Distribution

A standardized normal variable is said to follow a standard normal distribution. The density function of a standard normal distribution $N(0; 1)$ is given by:

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

The distribution function of a standard normal distribution is:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-v^2/2}\, dv$$

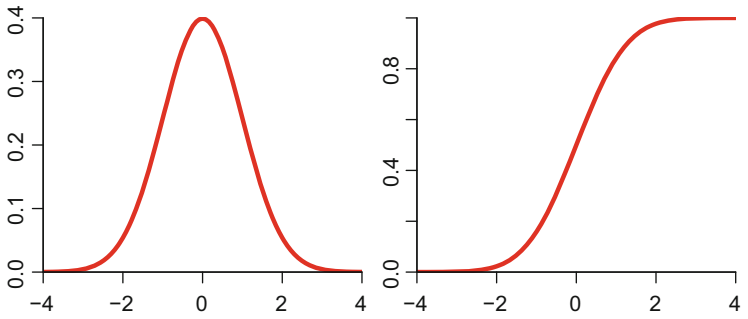


Fig. 6.21 Density (*left*) and distribution function (*right*) of $N(0; 1)$

As mentioned above, the expected value and variance of the standard normal distribution are given by:

$$E(Z) = 0, \quad Var(Z) = 1$$

The density and distribution function for a standard normal random variable are plotted in Fig. 6.21.

The relation between the distribution $N(\mu, \sigma)$ and the standard normal distribution:

$$x = \mu + z \cdot \sigma, \quad z = \frac{x - \mu}{\sigma}$$

which implies:

$$F_{NV}(x; \mu, \sigma) = P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = P(Z \leq z) = \Phi(z)$$

Confidence Interval

A confidence interval for the random variable X is the interval with boundaries x_l and $x_u (x_l \leq x_u)$, which will contain the value of the random variable X with probability 1 - α, i.e., $(1 - \alpha) \cdot 100\,\%$ of all values of X will fall in this interval and $\alpha \cdot 100\,\%$ will fall outside this interval. 1-α is usually referred to as the confidence level.

For known values of μ, the expected value of X, the interval is constructed to make the probability that X falls outside this region (there are 2 such regions) with probability $\alpha/2$. We call the interval

$$[x_u \leq x_o] = [\mu - k \leq X \leq \mu + k]$$

the (symmetric) confidence interval with confidence level

$$P(x_u \leq X \leq x_o) = 1 - \alpha.$$

To stress the importance of the standard deviation, as the parameter of scale, the deviation of X from its expected value μ is often measured in multiples of σ. The confidence interval has then this form:

$$[\mu - c\sigma \leq X \leq \mu + c\sigma].$$

If the random variable X is $N(\mu, \sigma)$, then for $x = \mu + c\sigma$ the following holds:

$$\frac{x - \mu}{\sigma} = \frac{\mu + c\sigma - \mu}{\sigma} = c = z$$

and

$$P(Z \leq z) = \Phi(z) = 1 - \frac{\alpha}{2}.$$

The critical value $z_{1-\alpha/2}$ for the probability 1 - $\alpha/2$ can be obtained from the tabulated values of a standardized normal distribution. Using these values, we can obtain the confidence interval for a normally distributed random variable:

$$[\mu - z_{1-\alpha/2}\sigma \leq X \leq \mu + z_{1-\alpha/2}\sigma]$$

and the probability of this interval is:

$$P(\mu - z_{1-\alpha/2}\sigma \leq X \leq \mu + z_{1-\alpha/2}\sigma) = 1 - \alpha.$$

For the confidence level of a normally distributed random variable we have (Fig. 6.22):

$$\begin{aligned} P(-z \leq Z \leq z) &= P(Z \leq z) - P(Z \leq -z) \\ &= P(Z \leq z) - [1 - P(Z \leq z)] \\ &= 2P(Z \leq z) - 1, \end{aligned}$$

which implies that

$$P(\mu - z_{1-\alpha/2}\sigma \leq X \leq \mu + z_{1-\alpha/2}\sigma) = 2\Phi(z) - 1.$$

For given z we can calculate the confidence levels of the interval:

$$
\begin{aligned}
P(\mu - z\sigma \leq X \leq \mu + z\sigma) \quad &= 0.6827 \quad \text{for} \quad z = 1 \\
&= 0.9545 \quad \text{for} \quad z = 2 \\
&= 0.9973 \quad \text{for} \quad z = 3
\end{aligned}
$$

On the other hand, we could also find the value z that produces the desired confidence level 1-α, e.g., $P(\mu - z_{1-\alpha/2}\sigma \leq X\mu + z_{1-\alpha/2}\sigma) = 0.95$, $z = 1.96$.

More Information

The normal distribution is one of the most important continuous distributions because:

- *approximate normality can be assumed in many applications*
- *it can be used to approximate other distributions*
- *many variables have normal distributions if there is a large number of observations*

A random variable with a normal distribution can take all values between $-\infty$ and $+\infty$. The normal distribution is also sometimes referred to as a Gaussian distribution. The density of a normal distribution is sometimes called the Bell curve.

The formulas for the density (or the distribution function) imply that a normal distribution will depend on the parameters μ and σ. By varying these parameters we can obtain a range of distributions. Figure 6.23 shows 5 normal densities with various parameters μ and σ.

The parameter μ specifies the location of the distribution. If we change the parameter μ, the location of the distribution will shift but its shape remains the

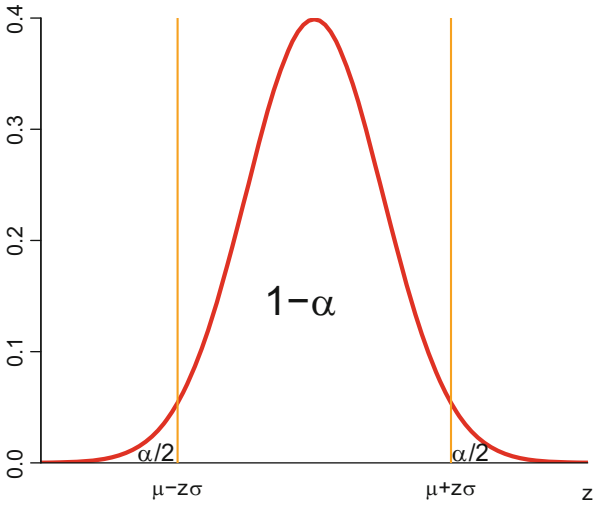


Fig. 6.22 The confidence interval for a normally distributed random variable

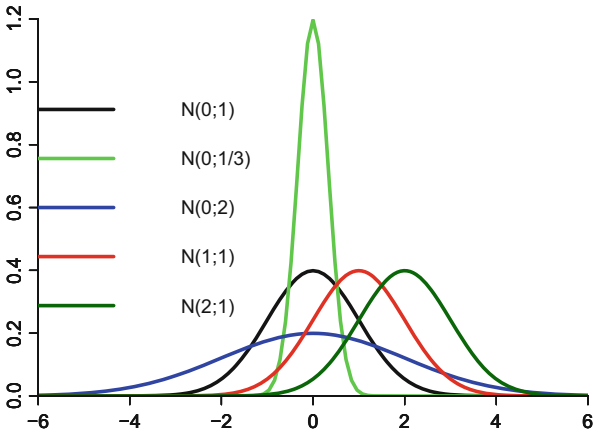


Fig. 6.23 Density functions of normal distribution with different parameter values

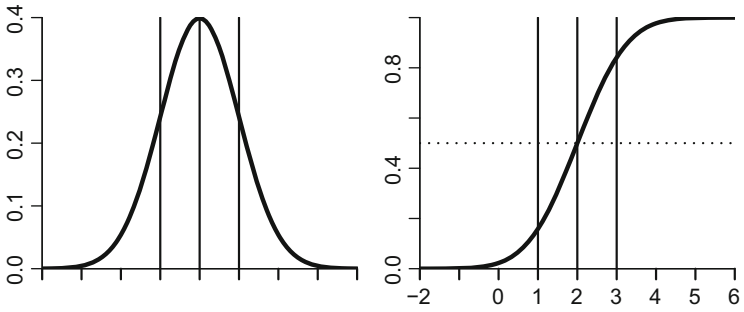


Fig. 6.24 Density (*left*) and distribution function (*right*) of $N(2; 1)$

same. By increasing or decreasing the parameter σ, the density "spreads" or "concentrates." Large values of σ produce flatter and wider densities. Small values of σ produce distributions that are narrow and tight.

Other Properties of the Normal Distribution

- The density has global maximum (the mode) at point $x = \mu$
- The density is symmetric around the point $x = \mu$. The symmetry implies that the median is $x_{0.5} = \mu$.
- The density has inflexion points at $x_1 = \mu - \sigma$ and $x_2 = \mu + \sigma$
- The density is asymptotically equal to 0 as $x \to -\infty$ or $x \to \infty$.

Figure 6.24 contains a plot of a $N(2; 1)$ distribution

Standard Normal Distribution

Tabulating the distribution function of the normal distribution for all values of μ and σ is not possible.

However, since we can transform a normal random variable to obtain another normal random variable, we need only to tabulate one distribution. The obvious choice is the normal distribution with expected value 0, $E(X) = \mu = 0$ and standard deviation 1, $\sigma = 1$.

This distribution is called a standard normal distribution, denoted $N(0, 1)$-distribution. The corresponding random variables are usually denoted by the letter Z.

The random variable Z is the random variable X centered at its mean and divided by its standard deviation. Hence $E(Z) = 0$ and $Var(Z) = 1$. If X is normally distributed, then Z also has a (standard) normal distribution.

The standard normal distribution is important because each random variable X with arbitrary normal distribution can be linearly transformed to a random variable Z with standard normal distribution.

In most tables for the density and distribution function of the standard normal distribution, you can find only positive values of Z. The tables of standard normal distribution for negative Z is unnecessary since the normal distribution is symmetric.

$$\Phi(-z) = P(Z \leq -z) = 1 - P(Z \leq z) = 1 - \Phi(z)$$

Explained: Normal Distributed Random Variable

Let us consider random variable X with normal distribution $N(100;\ 10)$.

1. We want to compute $P(X \leq x)$ for $x = 125$:

$$z = (x - \mu)/\sigma = (125 - 100)/10 = 2,5$$

$$P(X \leq 125) = F(125) = \Phi\left(\frac{125 - 100}{10}\right) = \Phi(2.5) = 0.99379$$

   There is a 99.38 % probability that the random variable X is smaller than 125 (Fig. 6.25).
2. We want to calculate the probability $P(X \geq x)$ for $x = 115.6$:

$$z = (x - \mu)/\sigma = (115.6 - 100)/10 = 1.56$$

$$\begin{aligned} P(X \geq 115.6) &= 1 - P(X \leq 115.6) = 1 - F(115.6) \\ &= 1 - \Phi\left(\frac{115.6 - 100}{10}\right) = 1 - \Phi(1.56) \\ &= 1 - 0.94062 = 0.05938 \end{aligned}$$

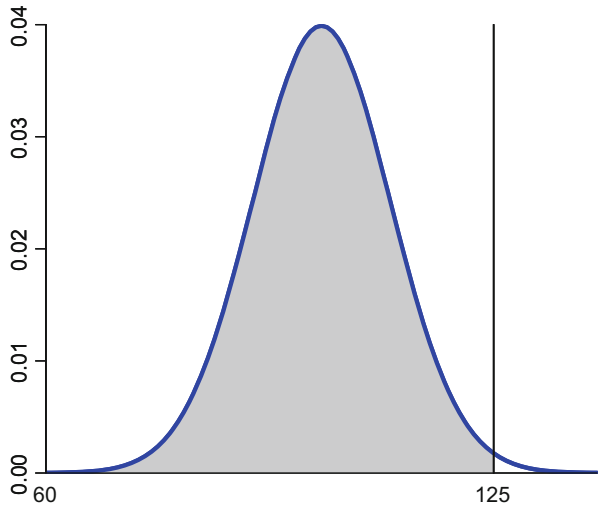


Fig. 6.25 Density of X following $N(100;\ 10)$; area $P(X \leq 125)$ in gray

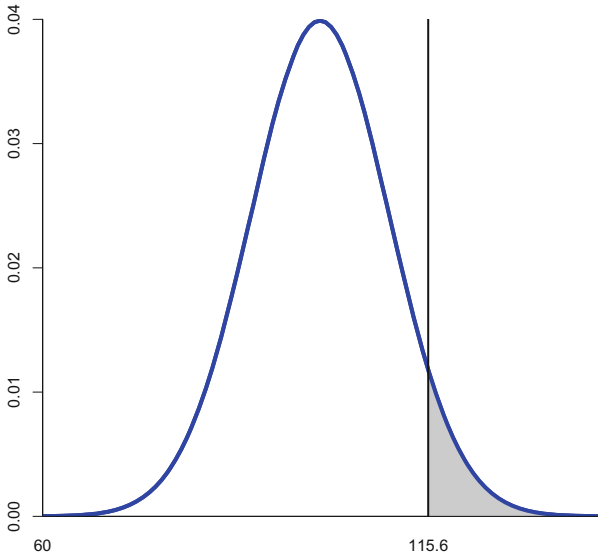


Fig. 6.26 Density of X following $N(100;\ 10)$; area $P(X \geq 115.6)$ in gray

There is a 5.94 % probability that the random variable X is greater than 115.6 (Fig. 6.26).

3. Let us calculate the probability $P(X \leq x)$ for $x = 80$:

$$z = (x - \mu)/\sigma = (80 - 100)/10 = -2$$

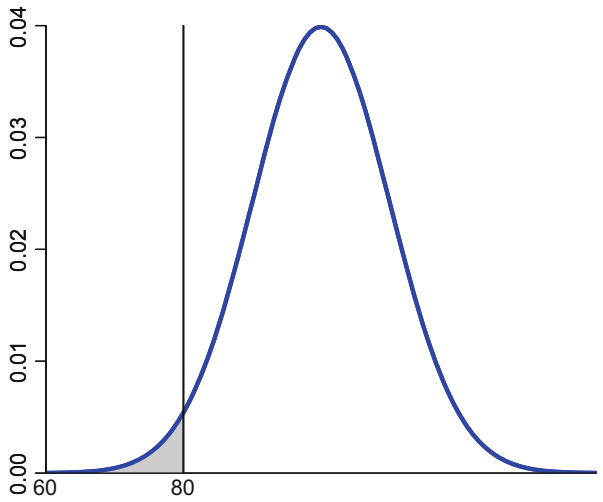


Fig. 6.27 Density of X following $N(100;\ 10)$; area $P(X \leq 80)$ in gray

$$P(X \leq 80) = F(80) = \Phi\left(\frac{80-100}{10}\right) = \Phi(-2) = 1 - \Phi(2)$$
$$= 1 - 0.97725 = 0.02275$$

The random variable X is smaller than 80 with probability of 2.275 % (Fig. 6.27).

4. Let us compute $P(X \geq x)$ for $x = 94.8$:

$$z = (x - \mu)/\sigma = (94.8 - 100)/10 = -0.52$$

$$P(X \geq 94.8) = 1 - P(X \leq 94.8) = 1 - F(94.8)$$
$$= 1 - \Phi\left(\frac{94.8-100}{10}\right) = 1 - \Phi(-0.52)$$
$$= 1 - (1 - \Phi(0.52)) = \Phi(0.52) = 0.698468$$

The probability that the random variable X is greater than 94.8 is 69.85 % (Fig. 6.28).

5. We compute the probability $P(x_u \leq X \leq x_o)$ for $x_u = 88.8$ and $x_o = 132$:

$$z_u = (x_u - \mu)/\sigma = (88.8 - 100)/10 = -1.12$$
$$z_o = (x_o - \mu)/\sigma = (132 - 100)/10 = 3.2$$

$$P(88.8 \leq X \leq 132) = P(X \leq 132) - P(X \leq 88,8)$$
$$= F(132) - F(88.8)$$

Fig. 6.28 Density of X following $N(100;\ 10)$; area $P(X \geq 94.8)$ in gray

$$
\begin{aligned}
&= \Phi(3.2) - \Phi(-1.12) \\
&= \Phi(3.2) - (1 - \Phi(1.12) \\
&= 0.999313 + 0.868643 - 1 \\
&= 0.867956
\end{aligned}
$$

The random variable X falls in the interval $[88.8\ ;\ 132]$ with probability 86.8 % (Fig. 6.29).

6. Let us calculate $P(x_u \leq X \leq x_o)$ for $x_u = 80.4$ and $x_o = 119.6$ (centered probability interval):

$$
\begin{aligned}
z_u &= (x_u - \mu)/\sigma = (80.4 - 100)/10 = -1.96 \\
z_o &= (x_o - \mu)/\sigma = (119.6 - 100)/10 = 1.96
\end{aligned}
$$

$$
\begin{aligned}
P(80.4 \leq X \leq 119.6) &= P(X \leq 119.6) - P(X \leq 80.4) \\
&= F(119.6) - F(80.4) \\
&= \Phi(1.96) - \Phi(-1.96) \\
&= \Phi(1.96) - (1 - \Phi(1.96) \\
&= 2\Phi(1.96) - 1 \\
&= 2 \cdot 0.975 - 1 = 0.95
\end{aligned}
$$

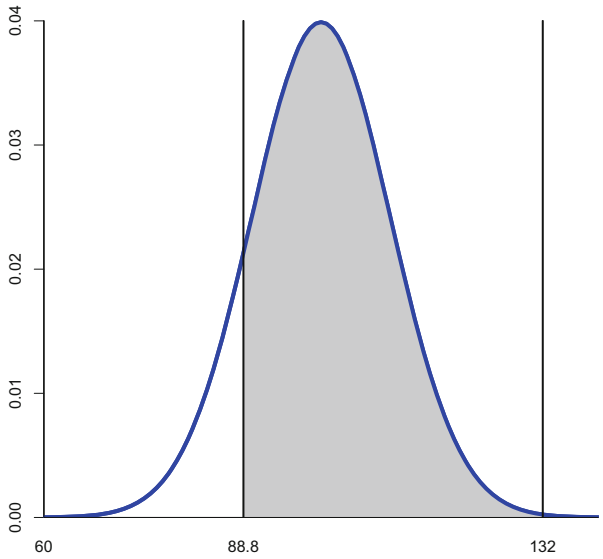


Fig. 6.29 Density of X following $N(100;\ 10)$; area $P(88.8 \leq X \leq 132)$ in gray

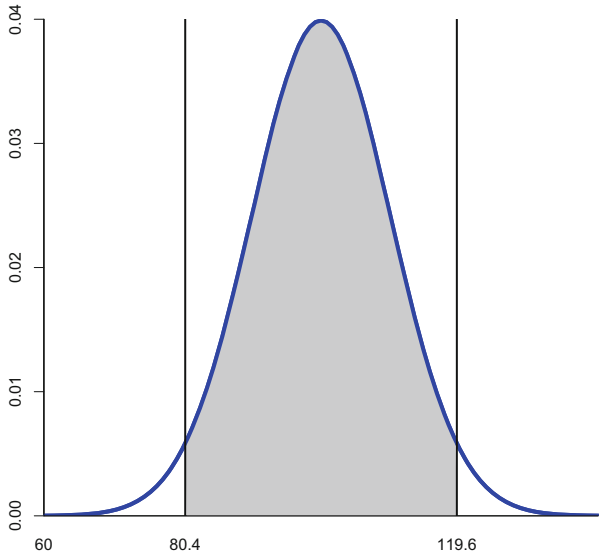


Fig. 6.30 Density of X following $N(100;\ 10)$; area $P(80.4 \leq X \leq 119.6)$ in gray

The random variable X falls into the interval $[88.4\,;\,119.6]$ with probability 95 % (Fig. 6.30).

7. We want to calculate an interval, which is symmetric around the expected value, such that it will contain 99 % of the realizations of X:

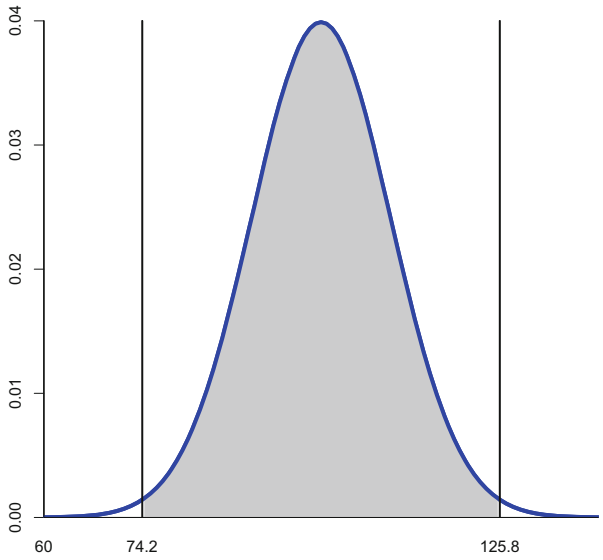


Fig. 6.31 Density of X following $N(100;\ 10)$; area $P(74.2 \leq X \leq 125.8)$ in gray

$$\begin{aligned} P(x_u \leq X \leq x_o) &= 0.99 \\ &= P\left(\frac{x_u - 100}{10} \leq Z \leq \frac{x_o - 100}{10}\right) \\ &= P(-z \leq Z \leq z) = 2\Phi(z) - 1, \\ \text{with } \Phi(z) &= \frac{1.99}{2} = 0.995 \end{aligned}$$

For the value (the probability) 0.995 we find in the tables of the distribution function of standard normal distribution function that $z = 2.58$. This implies:

$$x_o = \mu + z\sigma = 100 + 2.58 \cdot 10 = 125.8$$
$$x_u = \mu - z\sigma = 100 - 2.58 \cdot 10 = 74.2$$

take $P(74.2 \leq X \leq 125.8) = 0.99$.

The random variable X falls into the interval $[74.2\,;\,125.8]$ with a 99 % probability (Fig. 6.31).

8. Let us find an x such that 76.11 % of the realizations of X are smaller than x:

$$\begin{aligned} P(X \leq x) &= 0.7611 \\ &= P\left(Z \leq \frac{x - 100}{10}\right) = P(Z \leq z) \end{aligned}$$

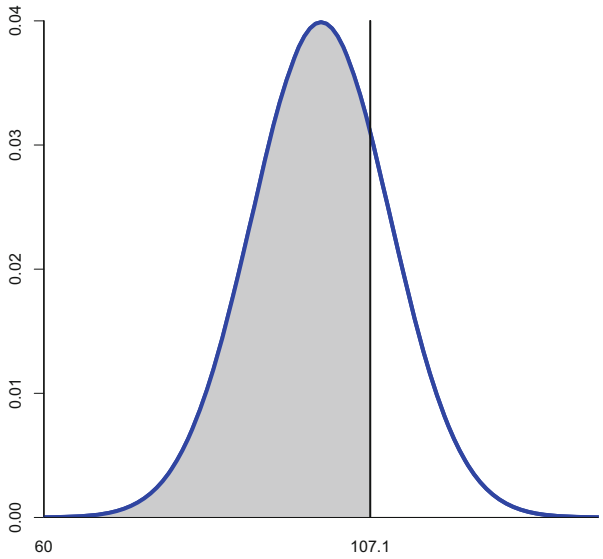


Fig. 6.32 Density of X following $N(100;\ 10)$; area $P(X \leq 107.1)$ in gray

For the value 0.7611 we obtain from the standard normal distribution tables that $z = 0.71$. Hence:

$$x = \mu + z\sigma = 100 + 0.71 \cdot 10 = 107.1$$

so that $P(X \leq 107.1) = 0.7611$.

There is a 76.11 % probability that the random variable X will be smaller than 107.1 (Fig. 6.32).

9. We calculate x such that 3.6 % of realizations of X are greater than x:

$$\begin{aligned} P(X \geq x) &= 0.036 \\ &= P\left(Z \geq \frac{x - 100}{10}\right) = P(Z \geq z) \end{aligned}$$

Since $P(Z \geq z) = 1 - P(Z \leq z) = 0.964$, using the standard normal distribution tables the value $z = 1.8$ for the probability 0.964. Hence,

$$x = \mu - z\sigma = 100 - 1.8 \cdot 10 = 118$$

so that $P(X \geq 118) = 0.036$.

There is a 3.6 % probability that the random variable X is greater than 118 (Fig. 6.33).

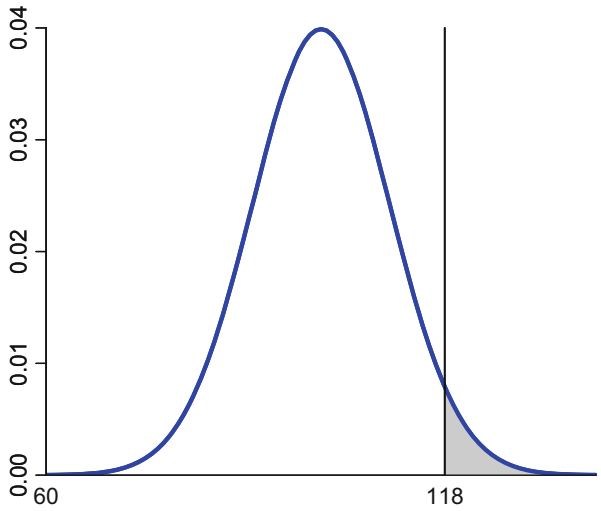


Fig. 6.33 Density of X following $N(100;\ 10)$; area $P(X \geq 118)$ in gray

Interactive: Normal Distribution

The normal distribution is described by two parameters which define its shape, location, and scale (variance).

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please select a value for

- the mean μ
- the variance σ^2

Moreover, choose one of the following functions:

- Probability mass function
- Cumulative distribution function

Output

This interactive example allows you to explore the effect of the parameters μ and σ^2 on shape of the distribution. The plot in Fig. 6.34 displays the probability distribution function (or cumulative distribution function) of the normal distribution function $N(\mu; \sigma^2)$.

We recommend that you only change one parameter at a time to better observe their effects on the distribution function.

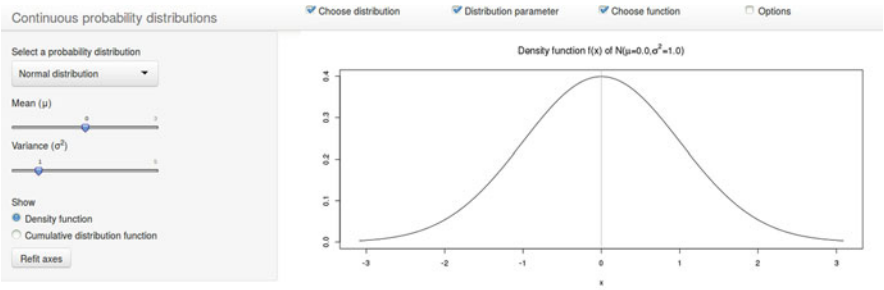


Fig. 6.34 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_norm

6.8 Central Limit Theorem

One property of the normal distribution is that the sum of n independent random variables $X_1, X_2, \ldots, X_n$ with normal distribution is also normally distributed. This property remains true for any value of n. If the random variables $X_1, X_2, \ldots, X_n$ are not normally distributed, then this property is not exactly true, but it remains approximately correct for large n.

Let $X_1, X_2, \ldots, X_n$ be independently and identically distributed random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 > 0$ for i = 1,... ,n. Then the sum of these random variables is for large n approximately normally distributed:

$$E(X_1 + X_2 + \cdots + X_n = n\mu \text{ and } Var(X_1 + X_2 + \cdots + X_n = n\sigma^2,$$

$$X_1 + X_2 + \cdots + X_n \approx N(n\mu, n\sigma^2),$$

where $\approx$ means approximately for large n.

Let $X_1, X_2, \ldots, X_n$ be independently and identically distributed random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 > 0$ for i= 1,... , n. Then the mean of these random variables is for large n approximately normally distributed:

$$E\left(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right) = E(\bar{x}) = \mu \text{ and } Var(\bar{x}) = \frac{\sigma^2}{n}$$

$$\bar{x} \approx N(\mu, \frac{\sigma^2}{n}).$$

This result requires that none of the random variables are responsible for most of the variance. The distribution $N(\mu, \frac{\sigma^2}{n})$ depends on the number of the summands n and for infinite n it would have infinite expected value and infinite variance. The meaning of this theorem can be described more clearly if we use standardized sums of random variables.

Central Limit Theorem

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables: $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 > 0$. Then, the distribution function $F_n(z) = P(Z_n \leq z)$ of

$$Z_n = \frac{\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$$

converges as n$\rightarrow \infty$ to a standard normal distribution:

$$\lim_{n\to\infty} F_n(z) = \Phi(z)$$

The "standardized" random variable Z_n is approximately standard normal distributed:

$$Z_n \approx N(0; 1).$$

More Information

The (Lindeberg and Lévy) Central Limit Theorem is the main reason that the normal distribution is so commonly used. The practical usefulness of this theorem derives from the fact that a sample of identically distributed independent random variables follows approximately a normal distribution as the sample increases. Usually $n \geq 30$ is deemed to be sufficiently large for a reasonably good approximation. This theorem becomes particularly important when deriving the sampling distribution of test statistics.

The convergence towards the normal distribution will be very quick if the distribution of the random variables is symmetric. If the distribution is not symmetric, then the convergence will be much slower.

The Central Limit Theorem has various generalizations (e.g., Lyapunov CLT for independent, but not identically distributed random variables). Furthermore, there are also limit theorems that describe convergence towards other sorts of distributions.

Explained: Application to a Uniform Random Variable

In this example, we will try to illustrate the principle of the Central Limit Theorem. Let us consider continuous random variables $X_1, X_2, \ldots$ random variables which are independently and identically uniformly distributed on the interval $[-0, 5; 0, 5]$:

$$f(x) = \begin{cases} 1 \text{ for } -0.5 \le x \le 0.5 \\ 0 \text{ otherwise.} \end{cases}$$

The expected value and the variance are:

$$E(X) = \frac{b+a}{2} = \frac{0.5-0.5}{2} = 0$$
$$Var(X) = \frac{(b-a)^2}{12} = \frac{[0.5-(-0.5)]^2}{12} = \frac{1}{12}.$$

Let us consider a sequence of the sum of these variables; the index of the variable Y denotes the number of observations in the sample:

$$Y_n = \sum_{i=1}^{n} X_i \qquad n = 1, 2, 3, \ldots .$$

For example, for $n = 1$, $n = 2$, and $n = 3$ we get:

$$Y_1 = X_1$$
$$Y_2 = X_1 + X_2$$
$$Y_3 = X_1 + X_2 + X_3.$$

The densities are:

$$f(y_1) = \begin{cases} 1 \text{ for } -0.5 \le y_1 \le 0.5 \\ 0 \text{ otherwise} \end{cases}$$

$$f(y_2) = \begin{cases} 1 + y_2 \text{ for } -1 \le y_2 \le 0 \\ 1 - y_2 \text{ for } 0 \le y_2 \le 1 \\ 0 \qquad \text{otherwise} \end{cases}$$

$$f(y_3) = \begin{cases} 0.5(1.5 + y_3)^2 & \text{for } -1.5 \le y_3 \le -0.5 \\ 0.5 + (0.5 + y_3)(0.5 - y_3) & \text{for } -0.5 < y_3 \le 1.5 \\ 0.5(1.5 - y_3)^3 & \text{for } 0.5 < y_3 \le 1.5 \\ 0 & \text{otherwise} \end{cases}$$

All these densities are plotted in Fig. 6.35, which also contains a plot of a $N(0, 1)$ density for comparison.

The convergence towards of these distributions to a normal density can be clearly seen. As the number of observations increases the distribution becomes more similar to a normal distribution. In fact, for $n \ge 30$ we can hardly see any differences.

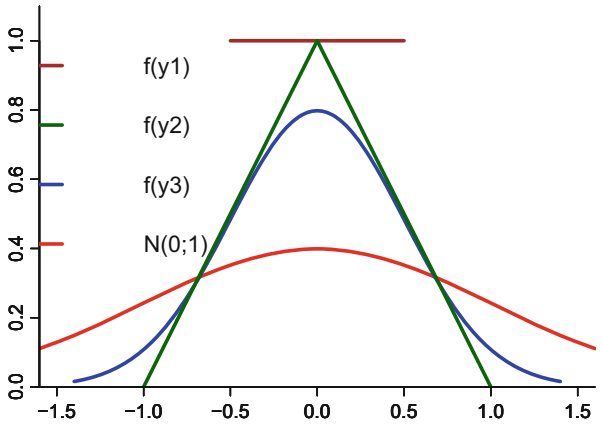


Fig. 6.35 Illustration of the Central Limit Theorem

6.9 Approximation of Distributions

Approximation means that, under certain conditions, another distribution provides a description of the data which is similar to the distribution the data were sampled from. The limit theorems (e.g., Central Limit Theorem) provide a theoretical tool for deriving such approximations. These limit theorems can be used to approximate a number of common distributions. Since we are dealing with approximations of the true distribution, there are some errors. However, there are methods for evaluating the quality of the approximation. In the following we present approximations for a number of distributions as well as some of the criteria that can be used to evaluate the quality of these approximations.

Normal Distribution as Limit of Other Distributions

- **Approximation of Binomial distribution by normal distribution:**
  This approximation is based on Laplace and DeMoivre's limit theorem.
  Let $X_1, \ldots, X_n$ be independent, Bernoulli distributed random variables with $E(X_i) = p$ and $Var(X_i) = p(1-p)$ for all i. Then $X = X_1 + \cdots + X_n$ is random variable with binomial distribution $B(n, p)$, expected value $E(X) = np$ and variance $Var(X) = np(1-p)$.
  For $n \to \infty$, the distribution of the standardized random variable

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

converges to a standard normal distribution $N(0;1)$. For large n we have:

$$X_n \approx N(np; \sqrt{np(1-p)})$$

with the expected value $\mu = np$ and variance $\sigma^2 = np(1-p)$.

Since the binomial distribution is discrete and the normal distribution is continuous we improve the quality of approximation by using a continuity adjustment:

$$P(X \leq x) = F_B(x;n,p) \approx \Phi\left(\frac{x+0.5-np}{\sqrt{np(1-p)}}\right)$$

$$P(X = x) = f_B(x;n,p) \approx \Phi\left(\frac{x+0.5-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{x-0.5-np}{\sqrt{np(1-p)}}\right)$$

A rough rule of thumb for a good approximation for the binomial distribution requires: $np \geq 5$ or $n(1-p) \geq 5$.

- **Approximation of the Poisson distribution by the normal distribution**

  The Poisson distribution with $\lambda = np$ can be derived from a binomial distribution. Since the binomial distribution can be approximated by the normal distribution this suggests that the normal distribution can also approximate the Poisson distribution.

  Let X be a random variable with the distribution $PO(\lambda)$. Then for large λ, we approximate the Poisson distribution using a normal distribution with expected value $\mu = \lambda$ and variance $\sigma^2 = \lambda$ (with the continuity correction):

$$P(X \leq x) = F_{PO}(x;\lambda) \approx \Phi\left(\frac{x+0.5-\lambda}{\sqrt{\lambda}}\right)$$

  The rule of thumb for a "reasonable" approximation requires: $\lambda \geq 10$

- **Approximation of hypergeometric distribution by normal distribution**

  Let $nM/N \geq 5$, $n(1-M/N) \geq 5$, and $n/M \leq 0.05$. Then a random variable with Hypergeometric distribution can be approximated using a normal distribution with the parameters:

$$E(X) = \mu = n \cdot \frac{M}{N} \quad Var(X) = \sigma^2 = n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)$$

  We can also use the continuity correction to improve the approximation.

- **Approximation of hypergeometric distribution by binomial distribution**

  The binomial and hypergeometric distributions use different sampling methods: the binomial distribution uses draws with replacement and the hypergeometric distribution uses draws without replacement. As M and N increase, M/N converges to a constant p, the difference between these two distributions becomes

much smaller. As $N \to \infty$ (and $M \to \infty$) the hypergeometric distribution converges to a binomial distribution. This implies: for large N and M as well as small n/N, the hypergeometric distribution can be approximated by the binomial distribution with parameters $p = M/N$. The rule of thumb requires: $n/M \leq 0.05$.

- **Approximation of binomial distribution by Poisson distribution**
  The Poisson distribution can also be derived from a binomial distribution. Consequently, the binomial distribution can also be approximated by a Poisson distribution $PO(\lambda = np)$, if n is large and the probability p is small. We are using the following rule of thumb: $n > 30$ and $p \leq 0.05$.

Explained: Wrong Tax Returns

Based on experience, we know that 10 % of tax returns from a certain town have errors. Using a sample of 100 tax returns from this town—what is the probability that 12 of them contain errors?

There are only two possible outcomes for the experiment—"wrong" or "correct," with corresponding probabilities $p = 0.1$ and $1 - p = 0.9$. The random variables X—"number of wrong tax returns from 100 randomly chosen ones" has the binomial distribution $B(n, p) = B(100; 0.1)$. We need to compute the probability $P(X = 12) = f_B(12)$:

$$f_B(12;\, 100;\, 0,1) = \binom{2000}{10} \cdot 0.1^{12} \cdot 0.9^{88} = 0.0988\,.$$

If the value $f_B(12;\, 100;\, 0,1)$ is not contained in the tables, we would have to compute it, which might be fairly cumbersome. However, since the conditions for the validity of an approximation using the normal distribution are satisfied ($np = 10 \geq 5$ and $n(1-p) = 90 \geq 5$), we could approximate the probability with a normal distribution $N(\mu;\, \sigma)$. The expected value and the variance of the binomial distribution are:

$$\mu = np = 100 \cdot 0.1 = 10\,, \quad \sigma^2 = np(1-p) = 100 \cdot 0.1 \cdot 0.9 = 9\,.$$

so we could use a $N(10;\, 3)$ distribution (see the diagram).

Recall: for a continuous random variable, the probabilities are given by the area under the density and thus the probability of one specific value is always equal to zero, e.g., $P(X = 12) = 0$.

Therefore, we subtract and add 0.5 to 12; this is a sort of continuity correction. Instead of $x = 12$ (for the discrete variable) we use an interval for the continuous $11.5 \leq x \leq 12.5$ and $f_B(12;\, 100;\, 0,1)$ is then approximated by $P(11.5 \leq x \leq 12.5)$, i.e., the area under the density of a $N(10;\, 3)$ between the points 11.5 and 12.5 shown in Fig. 6.36.

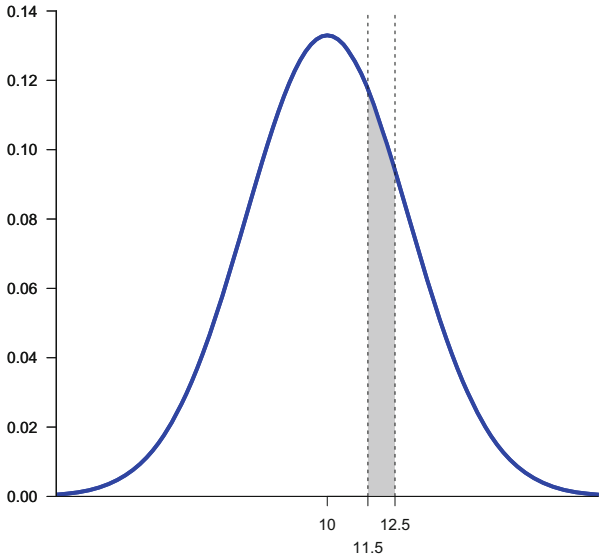


Fig. 6.36 Approximation of the binomial distribution by the normal distribution

The tables only contain the distribution function of a $N(0, 1)$ random variable, so we have to standardize the random variable X:

$$z_1 = (12.5 - 10)/3 = 0.83 \text{ and } z_2 = (11.2 - 10)/3 = 0.5\,.$$

Using the normal tables, we obtain $\Phi(0.83) = 0.7967$ and $\Phi(0.5) = 0.6915$. Hence,

$$P(11.5 \leq x \leq 12.5) = \Phi(0.83) - \Phi(0.5) = 0.7967 - 0.6915 = 0.1052\,.$$

The approximation works reasonably well, the error of the approximation is only $0.1052 - 0.0988 = 0.0064$. We can also see that:

- the approximate probability of having at most 12 wrong tax is

$$P(X \leq 12) = \Phi[12 + 0.5 - 10]/3] = \Phi(0.83) = 0.7967$$

- the approximate probability of obtaining more than 12 wrong tax forms is

$$P(X > 12) = 1 - \Phi[12 + 0.5 - 10]/3] = 1 - \Phi(0.83) = 1 - 0.7967 = 0.2033$$

- the approximate probability of obtaining at least 12 wrong tax forms is

$$P(X \geq 12) = 1 - \Phi[12 - 0.5 - 10]/3] = 1 - \Phi(0.5) = 1 - 0.6915 = 0.3085$$

Enhanced: Storm Damage

In a certain town, one house in each 100 is damaged every year because of storms. What is the probability that storms damage four houses in a year if the town contains 100 houses?

For each house, there are only two possible outcomes—"damage" and "no damage." The probabilities of these outcomes are constant: $p = 0.01$ and $1 - p = 0.99$. The random variable $X = \{\text{number of damaged houses}\}$ has the binomial distribution $B(n, p) = B(100;\, 0,01)$. We compute the probability $P(X = 4)$:

$$P(X = 4) = f_B(4;\, 100;\, 0,01) = \binom{100}{4} \cdot 0.01^4 \cdot 0.99^{96} = 0.01494\,.$$

We could also use the Poisson distribution (with parameter $\lambda = np = 1$) to approximate this probability since the conditions for a good approximation are satisfied:

$$F_{PO}(4;\, 1) = \frac{1^4}{4\,!} e^{-1} = 0.01533\,.$$

We see that the probabilities $f_B(4)$ and $F_{PO}(4)$ are fairly close. More generally, the approximation is also good at other points in the distribution (Table 6.4).

After a storm, there are 300 damaged houses out of a total of 2000 in a given region (Fig. 6.37). What is the probability that there are exactly 2 damaged houses among 10 randomly chosen houses?

Again, there are only two possible outcomes for each house—"damage" and "no damage." Furthermore, $N = 2000$, $M = 300$, and $N - M = 1700$. The probability $P(X = 2)$ is equal to

$$P(X = 2) = f_H(2) = \frac{\binom{300}{2} \cdot \binom{1700}{8}}{\binom{2000}{10}} = 0.2766\,.$$

Table 6.4 Approximation of the binomial distribution $B(100;\, 0.1)$ by the Poisson distribution $PO(1)$

x	$B(100;\, 0.1)$	$PO(1)$
0	0.36603	0.36788
1	0.36973	0.36788
2	0.18486	0.18394
3	0.06100	0.06131
4	0.01494	0.01533
5	0.00290	0.00307
6	0.00046	0.00051
7	0.00006	0.00007
8	0.00000	0.00000

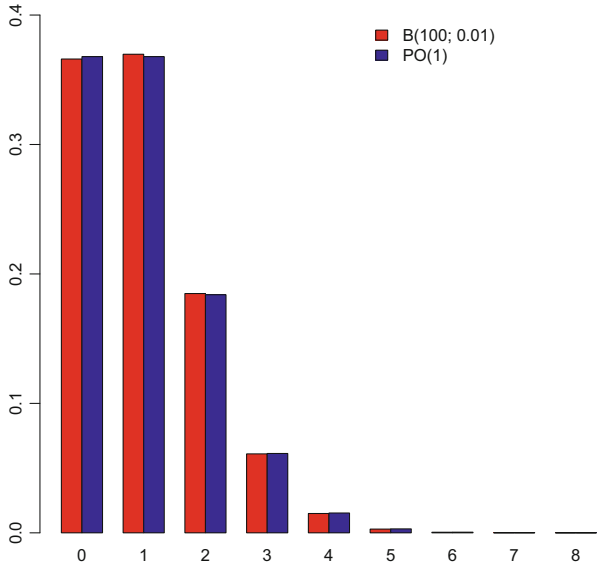


Fig. 6.37 Approximation of the binomial distribution $B(100; 0.1)$ by the Poisson distribution $PO(1)$

This calculation is fairly demanding. Fortunately, we can use binomial distribution (with parameter $p = M/N = 0.15$) to approximate this probability:

$$P(X = 2) - f_B(2) = \binom{10}{2} \cdot 0.15^2 \cdot 0.85^8 = 0.2759\,.$$

6.10 Chi-Square Distribution

Suppose we have n independently and identically distributed standard normal random variables $X_1, \ldots, X_n : X_i \sim N(0; 1)$ for $i = 1, \ldots, n$, where n is a positive integer.

The distribution of the sum of the squared $X_i s$

$$Y = X_1^2 + X_2^2 + \ldots X_n^2$$

is referred to as the Chi-square distribution with parameter df, or written shortly as $\chi^2(df)$.

The parameter df represents the degrees of freedom of the distribution, with $df > 0$. The expected value and variance of Chi-square distribution are given as

$$E(Y) = df \text{ a } Var(Y) = 2df.$$

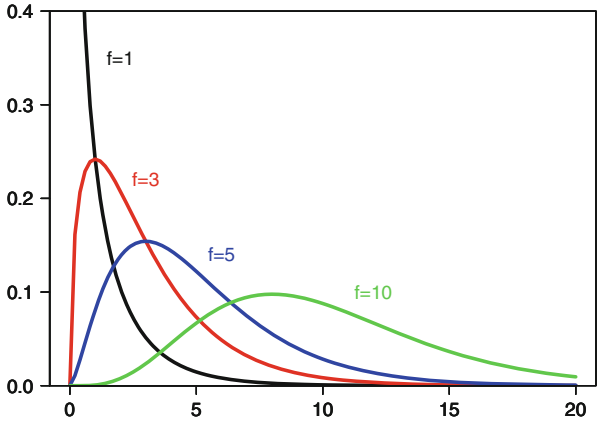


Fig. 6.38 Density functions of the chi-square distribution with different degrees of freedom

Figure 6.38 shows the density functions for some chi-square distributions, with different values for the degrees of freedom *df*.

More Information

The chi-square, t-, and F- distributions are distributions that are functions of normal random variables that are particularly useful in statistics.

On the Chi-Square Distribution

The parameter *df* denotes the degrees of freedom. The degrees of freedom reflects the number of independent random variables included in the sum Y. If the random variables $X_i, i = 1, \ldots, n$ are independent from each other, then squaring and summing them does not change their properties. In this example, the random variable Y:

$$Y = X_1^2 + X_2^2 + \cdots + X_n^2$$

will have the chi-square distribution with $df = n$ degrees of freedom.

The shape of the density function will depend on the parameter *df*. *For* $df = 1$ *and* $df = 2$, *the* χ^2 *distribution follows a monotone structures.* For small values of *df*, the χ^2-distribution will be skewed to the right. However, as *df* increases the

χ^2-distribution will tend towards the normal density function. The χ^2-distribution is tabulated for a number of values of df.

6.11 t-Distribution (Student t-Distribution)

The t-Distribution is also known as the Student t-Distribution. If Z has a standard normal distribution $N(0; 1)$ and Y, the sum of df squared standard normal random variables, has a χ^2-distribution with df degrees of freedom, then we define

$$T = \frac{Z}{\sqrt{\frac{Y}{df}}}$$

as the t-distribution with parameter df (shortly written as t(df)), if Z and Y are independent. The parameter df represents the degrees of freedom for the χ^2 random variable Y. The random variable T has range $-\infty \leq T \leq +\infty$ and expected value and variance:

$$E(T) = 0, \text{ for } df > 1$$
$$Var(T) = f/(f-2), \text{ for } df > 2$$

Figure 6.39 shows the density functions of the t-distribution for different numbers of degrees of freedom df.

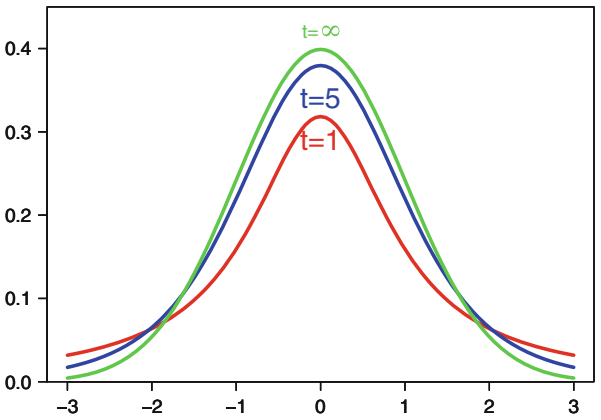


Fig. 6.39 Density functions of the t-distribution with different degrees of freedom

More Information

The Chi-square, t-, and F- distributions are distributions that are functions of normal random variables that are particularly useful in statistics.

On the t-Distribution

The density function of a t-distribution is a bell-shaped symmetric distribution with expected value $E(T) = 0$ (as a standard normal distribution). However, a t-distribution has heavier tails than a standard normal distribution. In other words, the t-distribution will be more dispersed than a standard normal distribution. The variance of the standard normal distribution is 1, but the variance of a t-distribution equals $Var(T) = df/(df - 2)$ (for $df > 2$).

As $df \to \infty$, the density function of the t-distribution converges to the standard normal distribution. For $df \geq 30$, a normal distribution can produce a good approximation to a t-distribution. The t-distribution is tabulated different values of df.

6.12 F-Distribution

Consider two independent χ^2 random variables Y_1 and Y_2, with df_1 and df_2 degrees of freedom respectively, then the random variable:

$$X = \frac{\frac{Y_1}{df_1}}{\frac{Y_2}{df_2}}$$

will have a F-distribution (denoted as $F(df_1, df_2)$) with parameters df_1 and df_2. The df_1 and df_2 parameters represent the degrees of freedom for the χ^2 distributed random variables in the numerator and the denominator.

An F-distribution with parameters df_1 and df_2 has expected value and variance

$$E(X) = \frac{df_2}{df_2 - 2}, \quad \text{for } df_2 > 2$$

$$Var(X) = \frac{2df_2^2(df_1 + df_2 - 2)}{df_1(df_2 - 2)^2(df_2 - 4)}, \quad \text{for } df_2 > 4$$

Figure 6.40 shows densities of the F-distribution for different values of df_1 and df_2.

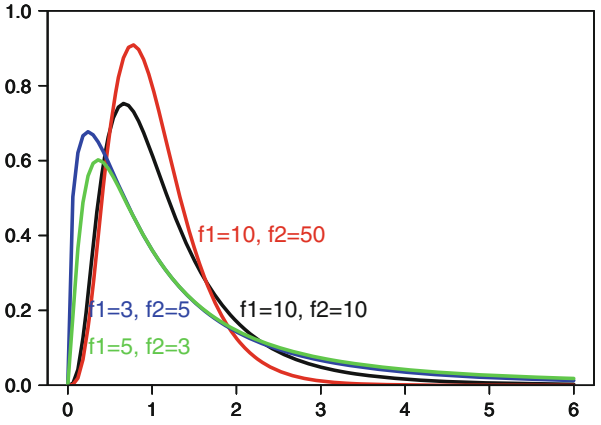


Fig. 6.40 Density functions of the F-distribution with different degrees of freedom

More Information

The Chi-square, t-, and F-distributions are distributions that are functions of normal random variables that are particularly useful in statistics.

On the F-Distribution

The density function of an F-distribution is right-skewed. Increasing the values of f_1 and f_2 reduces this skewness. As $f_1 \to \infty$ and $f_2 \to \infty$, the density of the F-distribution will tend to a standard normal distribution. The F-distribution is plotted for different values of df_1 and df_2.

Chapter 7
Sampling Theory

7.1 Basic Ideas

Population

One of the major tasks of statistics is to obtain information about populations. The set of all elements that are of interest for a statistical analysis is called a population. The population must be defined precisely and comprehensively so that one can immediately determine whether an element belongs to it or not.

Size of the Population The size of the population, N, is simply the number of elements in the population. Populations can be of finite or infinite in size and may even be hypothetical.

Suppose that a random variable X takes on J distinct values $x_j (j = 1, \ldots, J)$ in a finite population with certain absolute and relative frequencies $h(x_j)$ and $f(x_j)$ respectively. The absolute frequency $h(x_j)$ is the total number of elements in the population for which $X = x_j$. The relative frequency is related to the absolute frequency as follows: $f(x_j) = h(x_j)/N$.

To easily describe the population or distribution, certain characteristics can be computed. They are often denoted with Greek letters:

- The mean

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}\sum_{j=1}^{J} x_j h(x_j) = \sum_{j=1}^{J} x_j f(x_j)$$

- The variance

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{N}\sum_{j=1}^{J}(x_j - \mu)^2 h(x_j) = \sum_{j=1}^{J}(x_j - \mu)^2 f(x_j)$$

W.K. Härdle et al., *Introduction to Statistics*, DOI 10.1007/978-3-319-17704-5_7

- The standard deviation

$$\sigma = \sqrt{\sigma^2}$$

- Suppose that a random variable X is binary, that is, it takes on the two distinct values $x_1 = 0$ and $x_2 = 1$. Then, the proportion is defined as

$$\pi = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}\sum_{j=1}^{2} x_j h(x_j) = \sum_{j=1}^{2} x_j f(x_j).$$

Each characteristic takes a fixed value for the population. (As we shall see below, their sample counterparts which we will call statistics, such as the sample mean, sample variance, and sample proportion, will vary from sample to sample.)

The distribution of the variable X and its characteristics are typically unknown. To learn about them one could try to look at all elements of a population, i.e., conduct a census.

Census: In a census, data are collected on all elements of a population. Only in this case can the distribution and characteristics of X be determined exactly.

Sample

Any finite subset of observations drawn from the population is called a sample. The number of elements of a sample is called the sample size and denoted with n.

Inductive Inference Since a sample only contains a subset of the elements of the population it can merely provide incomplete information about the distribution of the variable X in the population. Yet, results obtained from analyzing the sample can be used to draw inferences about the population. This type of inference (from the sample to the population) is called inductive inference. Inductive inferences cannot be made with certainty and may be wrong. Often, the laws of probability can be used to calculate the degree of uncertainty of these conclusions. That is, inductive inference provides a set of tools for drawing probabilistic conclusions about a population from a sample. Using these tools requires that the sample is drawn in a way that can be formalized by a probability model. This is assured if the selection of elements into the sample is done randomly.

Random Sampling There are two basic approaches to random sampling from a finite population:

– without replacement
– with replacement

In sampling without replacement, each element of the population has the same probability of being selected as each observation is drawn. However, the draws are not independent because the population distribution of X changes as observations are removed.

In sampling with replacement, each observation has the same probability of being selected as each observation is drawn. In this case, the draws are independent of each other. However, because observations are being replaced, (and therefore the population and distribution of X does not change), the same element may occur more than once in the sample.

Drawing a random sample of size n can be viewed as a sequence of n random experiments. Each draw thus corresponds to a random variable and the entire sample is a collection of n random variables $X_1, \ldots, X_n$.

The simplest sampling scheme involves sampling with replacement. In this case,

- The random variables $X_1, \ldots, X_n$ are identically distributed and all have the same distribution function $F(x)$ as the variable X in the population;
- the random variables $X_1, \ldots, X_n$ are independent random variables.

The n actual realizations of $X_1, \ldots, X_n$ are denoted as $x_1, \ldots, x_n$.

Statistic

A function $U = U(X_1, \ldots, X_n)$ of the random variables $X_1, \ldots, X_n$ is called a statistic. A statistic, being a function of random variables, is a random variable itself, with its own distribution, called the sampling distribution.

The expected value, variance, and standard deviation of the sampling distribution are denoted as follows:

– expected value $E(U) = \mu_U$
– variance $Var(U) = \sigma_U^2$
– standard deviation $\sigma_U = \sqrt{Var(U)}$

After the sample is actually drawn the n realizations $x_1, \ldots, x_n$ of the random variables $X_1, \ldots, X_n$ are observed. Calculating $u = U(x_1, \ldots, x_n)$ as a function of the n actual realizations $x_1, \ldots, x_n$ yields a realization of the statistic $U = U(X_1, \ldots, X_n)$.

If one repeatedly draws samples of a given size n from the same population, then the corresponding realizations of X and U will vary from sample to sample.

When discussing statistics, it is common to use lowercase for both the random variable and its realization. The context determines which object is being described.

The purpose of calculating statistics is to use them for drawing inferences about unknown population characteristics. The specific rule for calculating a statistic is usually obtained by analogy to its population counterpart. Important statistics are:

- sample average (or sample mean) by analogy to population mean μ, is calculated using

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{j=1}^{J} x_j \hat{h}(x_j) = \sum_{j=1}^{J} x_j \hat{f}(x_j),$$

  where $\hat{h}(x_j)$ and $\hat{f}(x_j)$ are the absolute and relative frequencies in the sample.

  In Chap. 2, we outlined descriptive statistics which are applied to a given body of data. In that case we had not yet distinguished between populations and samples. Here, we denote population absolute and relative frequencies using $h(x_j)$ and $f(x_j)$ respectively, while their sample counterparts are distinguished with the "hat" symbol above. The hat notation is commonly used in statistics to denote estimators (see Chap. 8), and indeed, we can think of the sample relative frequencies $\hat{f}(x_j)$ as estimates or approximations of their population counterparts the $f(x_j)$.
- mean squared deviation by analogy to population variance σ^2

$$MSD = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{j=1}^{J}(x_j - \bar{x})^2 \hat{h}(x_j) = \sum_{j=1}^{J}(x_j - \bar{x})^2 \hat{f}(x_j)$$

  the closely related sample variance divides by $n - 1$ instead of n (for further explanation of this subtle difference see Chap. 8):

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{j=1}^{J}(x_j - \bar{x})^2 \hat{h}(x_j)$$

$$= \frac{n}{n-1}\sum_{j=1}^{J}(x_j - \bar{x})^2 \hat{f}(x_j)$$

- sample proportion by analogy to population proportion π

$$\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{j=1}^{2} x_j \hat{h}(x_j) = \sum_{j=1}^{2} x_j \hat{f}(x_j)$$

We re-emphasize (see Sect. 5.1) that uppercase letters are used to denote random variables and lowercase letters are used to denote their realizations.

More Information

Reasons for Drawing Samples

Although complete information about the distribution of some variable X in a population can only be obtained by means of a census, there are important reasons to still draw samples. A census is often not feasible, too expensive or too time consuming:

- Conducting the census means destroying the elements of the population.
  Example: X is the life times of batteries or light bulbs. In this case, each element of the population would be used until it becomes unusable.
- The population is very large.
  Example: to write a report about the state of North American forests, it is impossible to inspect every tree.
- The population is hypothetical or of infinite size.
- The population has elements that do not yet exist.
  Example: the population of all items that have or will be produced by a certain machine.

Regarding Random Sampling

We have discussed two types of random sampling: sampling with replacement and sampling without replacement. The distinction between these becomes irrelevant if the size of the sample n is small relative to the size of the population N, In this case, removal of observations results in small changes to the remaining population.

There are many other types of sampling schemes, such as stratified sampling and cluster sampling.

Explained: Illustrating the Basic Principles of Sampling Theory

There are $N = 7$ participants in an examination for a course at the graduate level. Table 7.1 gives the results.

The variable X="Number of points in the exam" has the frequency distribution in the population shown in Table 7.2.

Table 7.1 Examination for a course at the graduate level

Student	A	B	C	D	E	F	G
Points	10	11	11	12	12	12	16

Table 7.2 Frequency distribution of X="number of points in the exam"

x	$h(x)$	$f(x) = h(x)/N$	$F(x)$
10	1	1/7	1/7
11	2	2/7	3/7
12	3	3/7	6/7
16	1	1/7	7/7

Table 7.3 Random sampling with replacement of size $n = 2$

	2. Exam						
1. Exam	10	11	11	12	12	12	16
10	10;10	10;11	10;11	10;12	10;12	10;12	10;16
11	11;10	11;11	11;11	11;12	11;12	11;12	11;16
11	11;10	11;11	11;11	11;12	11;12	11;12	11;16
12	12;10	12;11	12;11	12;12	12;12	12;12	12;16
12	12;10	12;11	12;11	12;12	12;12	12;12	12;16
12	12;10	12;11	12;11	12;12	12;12	12;12	12;16
16	16;10	16;11	16;11	16;12	16;12	16;12	16;16

From this distribution the mean, variance, and standard deviation of the variable X in the population can be calculated:

$$\mu = 12\,, \quad \sigma^2 = 22/7 = 3.143\,, \quad \sigma = 1.773$$

Randomly selecting an exam from this population and recording its points give rise to a random variable, which is also labeled X. The relative frequencies in the population correspond to the probabilities that an exam with a given score will be selected. The random variable X therefore has probability function $f(x)$ and cumulative distribution function $F(X)$ as laid out in Table 7.2, as well as expected value $\mu = 12$ and variance $\sigma^2 = 3.143$.

Random Sampling with Replacement

Suppose that two exams are randomly selected from the population and their scores recorded, but that after each draw, the selected exam is returned to the population before the next exam is selected. The random variables X_1 = "first exam score" and X_2 = "second exam score" can be defined accordingly. Table 7.3 shows all possible samples of size $n = 2$.

The probability of obtaining any one of these samples is 1/49. It is straightforward to infer the probability functions of X_1 and X_2 from Table 7.3.

The probability functions of X_1 and X_2 are identical to each other and also to the distribution of the variable X in the population. The two-dimensional probability distribution of $f(x1, x2)$ can also be deduced from Table 7.3.

Table 7.4 Probability functions for first and second draw with replacement, X_1 and X_2

x_1	$h(x_1)$	$f(x_1)$	x_2	$h(x_2)$	$f(x_2)$
10	7	7/49 = 1/7	10	7	7/49 = 1/7
11	14	14/49 = 2/7	11	14	14/49 = 2/7
12	21	21/49 = 3/7	12	21	21/49 = 3/7
16	7	7/49 = 1/7	16	7	7/49 = 1/7

Table 7.5 The two-dimensional probability distribution of $f(x_1, x_2)$

	X_2				
X_1	10	11	12	16	$f(x_1)$
10	1 / 49	2 / 49	3 / 49	1 / 49	1/7
11	2 / 49	4 / 49	6 / 49	2 / 49	2/7
12	3 / 49	6 / 49	9 / 49	3 / 49	3/7
16	1 / 49	2 / 49	3 / 49	1 / 49	1/7
$f(x_2)$	1/7	2/7	3/7	1/7	1

Table 7.6 Random sampling without replacement of size $n = 2$

	2. Exam						
1. Exam	10	11	11	12	12	12	16
10		10;11	10;11	10;12	10;12	10;12	10;16
11	11;10		11;11	11;12	11;12	11;12	11;16
11	11;10	11;11		11;12	11;12	11;12	11;16
12	12;10	12;11	12;11		12;12	12;12	12;16
12	12;10	12;11	12;11	12;12		12;12	12;16
12	12;10	12;11	12;11	12;12	12;12		12;16
16	16;10	16;11	16;11	16;12	16;12	16;12	

The last column of Table 7.5 contains the marginal distribution of X_1 and the last row contains the marginal distribution of X_2, which have already been given in Table 7.4.

For each cell of Table 7.5, i.e., for each pair $(x1, x2)$, we have:

$$f(x_1, x_2) = f(x_1) \cdot f(x_2)$$

The random variables X_1 and X_2 are therefore independent. Independently identically distributed (i.i.d.) sampling schemes are the simplest data generating mechanisms.

Random Sampling Without Replacement

Two exams are randomly drawn without replacement and the random variables X_1 and X_2 are defined as before. Table 7.6 shows all possible samples of size $n = 2$.

The probability of obtaining any one of these samples is 1/42. It is straightforward to infer the probability functions of X_1 and X_2 from Table 7.6.

It is not surprising that $f(x_1)$, the probability function of X_1, is identical to the distribution of X in the population. However, in random sampling without replacement, the population distribution changes after the first draw because the sampled item is not returned. The distribution of the second draw depends on the particular value of the first draw. If the first draw produced an exam with a score of 10 points ($X_1 = 10$), then—conditional on this turnout of the first draw—the probability of drawing an exam with a score of 10 in the second draw is also zero ($P(X_2 = 10|X_1 = 10) = 0$), because there is no exam left with a score of 10 points. Each column of Table 7.8 contains the conditional probability distribution for the second draw given a particular value for the first draw.

The unconditional probability (or equivalently the marginal probability) that X_2 takes on a specific value x_2 (i.e., $P(X_2 = x_2) = f(x_2)$) can be calculated from the law of total probability:

$$\begin{aligned} P(X_2 = 10) = {} & P(X_2 = 10|X_1 = 10) \cdot P(X_1 = 10) + \\ & P(X_2 = 10|X_1 = 11) \cdot P(X_1 = 11) + \\ & P(X_2 = 10|X_1 = 12) \cdot P(X_1 = 12) + \\ & P(X_2 = 10|X_1 = 16) \cdot P(X_1 = 16) \\ = {} & 0 \cdot 1/7 + 1/6 \cdot 2/7 + 1/6 \cdot 3/7 + 1/6 \cdot 1/7 = 6/42 = 1/7 \end{aligned}$$

$$\begin{aligned} P(X_2 = 11) = {} & P(X_2 = 11|X_1 = 10) \cdot P(X_1 = 10) + \\ & P(X_2 = 11|X_1 = 11) \cdot P(X_1 = 11) + \\ & P(X_2 = 11|X_1 = 12) \cdot P(X_1 = 12) + \\ & P(X_2 = 11|X_1 = 16) \cdot P(X_1 = 16) \\ = {} & 2/6 \cdot 1/7 + 1/6 \cdot 2/7 + 2/6 \cdot 3/7 + 2/6 \cdot 1/7 = 12/42 = 2/7 \end{aligned}$$

$$\begin{aligned} P(X_2 = 12) = {} & P(X_2 = 12|X_1 = 10) \cdot P(X_1 = 10) + \\ & P(X_2 = 12|X_1 = 11) \cdot P(X_1 = 11) + \\ & P(X_2 = 12|X_1 = 12) \cdot P(X_1 = 12) + \\ & P(X_2 = 12|X_1 = 16) \cdot P(X_1 = 16) \\ = {} & 3/6 \cdot 1/7 + 3/6 \cdot 2/7 + 2/6 \cdot 3/7 + 3/6 \cdot 1/7 = 18/42 = 3/7 \end{aligned}$$

Table 7.7 Probability functions for first and second draw without replacement, X_1 and X_2

x_1	$h(x_1)$	$f(x_1)$	x_2	$h(x_2)$	$f(x_2)$
10	6	$6/42 = 1/7$	10	6	$6/42 = 1/7$
11	12	$12/42 = 2/7$	11	12	$12/42 = 2/7$
12	18	$18/42 = 3/7$	12	18	$18/42 = 3/7$
16	6	$6/42 = 1/7$	16	6	$6/42 = 1/7$

Table 7.8 Conditional probability distribution for the second draw given a particular value for the first draw without replacement

x_2	$P(X_2 = x_2 \mid X_1 = 10)$	$P(X_2 = x_2 \mid X_1 = 11)$
10	0	3/6
11	2/6	1/6
12	3/6	3/6
16	1/6	1/6
$\sum$	1	1
x_2	$P(X_2 = x_2 \mid X_1 = 12)$	$P(X_2 = x_2 \mid X_1 = 16)$
10	1/6	1/6
11	2/6	2/6
12	2/6	3/6
16	1/6	0
$\sum$	1	1

$$
\begin{aligned}
P(X_2 = 16) = {} & P(X_2 = 16 \mid X_1 = 10) \cdot P(X_1 = 10) + \\
& P(X_2 = 16 \mid X_1 = 11) \cdot P(X_1 = 11) + \\
& P(X_2 = 16 \mid X_1 = 12) \cdot P(X_1 = 12) + \\
& P(X_2 = 16 \mid X_1 = 16) \cdot P(X_1 = 16) \\
= {} & 1/6 \cdot 1/7 + 1/6 \cdot 2/7 + 1/6 \cdot 3/7 + 0 \cdot 1/7 = 6/42 = 1/7
\end{aligned}
$$

These are the probabilities reported in Table 7.7. Hence, $f(x_2)$ is identical to $f(x_1)$ and both are identical to the population distribution. However, X_1 and X_2 are not independent. This can be seen from the conditional distributions of Table 7.8 (which are not identical) as well as from the two-dimensional joint distribution $f(x1, x2)$ calculated from Table 7.6.

Obviously, $f(x_1, x_2) \neq f(x_1) \cdot f(x_2)$, and hence X_1 and X_2 are not independent (Table 7.9).

Conclusion X_1 and X_2 are identically distributed and have the same distribution as the variable X in the population but they are not independent.

Table 7.9 Two-dimensional joint distribution $f(x_1, x_2)$ for two draws without replacement

	X_2				
X_1	10	11	12	16	$f(x_1)$
10	0	2 / 42	3 / 42	1 / 42	1/7
11	2 / 42	4 / 42	6 / 42	2 / 42	2/7
12	3 / 42	6 / 42	9 / 42	3 / 42	3/7
16	1 / 42	2 / 42	3 / 42	1 / 42	1/7
$f(x_2)$	1/7	2/7	3/7	1/7	1

7.2 Sampling Distribution of the Mean

The distribution of a statistic (which is itself a function of the sample) is called a sampling distribution. Statistics are used for estimating unknown population characteristics or parameters and for testing hypotheses. These tasks involve probability statements which can only be made if the sampling distributions of the statistics are known (or can be approximated). For the most important statistics, we now present in each case the sampling distribution its expected value and variance.

Distribution of the Sample Mean

Consider sampling from a population with distribution function $F(x)$, expected value $E(X) = \mu$, and variance $Var(X) = \sigma^2$. One of the most important statistics is the sample mean.

The sample mean (or sample average) is given by:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

The expected value, variance, and standard deviation of the sample mean are given by:

1. for a random sample with replacement

$$E(\bar{X}) = \mu$$

$$Var(\bar{X}) = \sigma^2(\bar{X}) = \frac{\sigma^2}{n}$$

$$\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

2. for a random sample without replacement

$$E(\bar{X}) = \mu$$
$$Var(\bar{X}) = \sigma^2(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$
$$\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

The factor $\frac{N-n}{N-1}$ is called the finite population correction.

If the population variance $Var(X) = \sigma^2$ is unknown it has to be estimated by the statistic s^2. In the above formulas σ^2 is replaced by s^2 which leads to an estimator of the variance of the sample mean given by:

- For a simple random sample:

$$\widehat{\sigma^2}(\bar{X}) = \frac{s^2}{n}$$

- For a random sample without replacement

$$\widehat{\sigma^2}(\bar{X}) = \frac{s^2}{n} \cdot \frac{N-n}{N-1}$$

These results for the expectation and variance of the sample mean hold regardless of the specific form of its sampling distribution.

Distribution of the Sample Mean

The sampling distribution $F(\bar{x})$ of the sample mean is determined by the distribution of the variable X in the population. In each case below we assume a random sample with replacement.

1. X has a normal distribution

   It is assumed that X is normally distributed with expected value μ and variance σ^2, that is:

$$X \sim N(\mu,\, \sigma^2)$$

   (a) The population variance σ^2 is known; in this case $\bar{x}$ has the following normal distribution:

$$\bar{X} \sim N(\mu,\, \sigma^2(\bar{X})) = \; N\left(\mu,\, \frac{\sigma^2}{n}\right)$$

and the standardized random variable

$$Z = \frac{\bar{X} - \mu}{\sigma(\bar{X})} = \sqrt{n}\frac{\bar{X} - \mu}{\sigma}$$

follows the standard normal distribution $Z \sim N(0; 1)$.

(b) The population variance σ^2 is unknown. In this case, it may be estimated by s^2. The transformed random variable:

$$T = \sqrt{n}\frac{\bar{X} - \mu}{S}$$

has a tabulated distribution with a parameter (the "degrees of freedom") which equals $n - 1$. This distribution is called the t-distribution and it is usually denoted by t_{n-1}.

As n increases, the t-distribution converges to a standard normal. Indeed the latter provides a good approximation when $n > 30$, as explained in Chap. 6.

2. The variable X has an arbitrary distribution. This is the most relevant case for applications in business and economics since the distribution of many interesting variables may not be well approximated by the normal or its specific form is simply unknown.

Consider n i.i.d. random variables $X_1, \ldots, X_n$ with unknown distribution. The random variables have expectation $E(X_i) = \mu$ and variance $Var(X_i) = \sigma^2$. According to the central limit theorem the following propositions hold:

- If σ^2 is known, then the random variable

$$Z = \sqrt{n}\frac{\bar{X} - \mu}{\sigma}$$

is approximately standard normal for sufficiently large n.

- If σ^2 is unknown, then the random variable

$$T = \sqrt{n}\frac{\bar{X} - \mu}{S}$$

is also approximately standard normal for sufficiently large n.

As rule of thumb, the normal distribution can be used for $n > 30$.

Calculating probabilities: If X is normally distributed with known μ and σ^2 so that $\bar{x}$ also follows the normal distribution, then the calculation of probabilities may be done as in Chap. 6. Calculations hold approximately if X is arbitrarily distributed and n is sufficiently large. More generally, if the distribution of X is not normal, but is known, then it is in principle possible to calculate the sampling distribution of $\bar{x}$ and the probabilities that falls in a given interval (though the results may be quite complicated).

Weak Law of Large Numbers

Suppose $X_1, \ldots, X_n$ are n independent and identically distributed random variables with expectation $E(X_i) = \mu$ and variance $Var(X_i) = \sigma^2$. Then, for each $\epsilon > 0$ it holds that:

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| < \epsilon) = 1 \ .$$

This can be proven by means of Chebyshev's inequality. It holds that

$$P(|\bar{X}_n - \mu| < \epsilon) \geq 1 - \frac{\sigma^2(\bar{X})}{\epsilon^2}.$$

After inserting $\sigma^2(\bar{X}) = \sigma^2/n$:

$$P(|\bar{X}_n - \mu| < \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2}$$

If n approaches infinity the second term on the right-hand side goes to zero.

Implication of This Law With increasing n, the probability that the sample mean $\bar{X}$ will deviate from its expectation μ by less than $\epsilon > 0$ converges to one. If the sample size is large enough the sample mean will take on values within a prespecified interval $[\mu - \epsilon; \mu + \epsilon]$ with high probability, regardless of the distribution of X.

More Information

Consider a population with distribution function $F(x)$, expected value $E(X) = \mu$, and variance $Var(X) = \sigma^2$. The random variables $X_i, i = 1, \ldots, n$ all have the same distribution function $F(x_i) = F(x)$, expectation $E(X_i) = \mu$ and variance $Var(X_i) = \sigma^2$.

Expectation of the Sample Mean $\bar{X}$

Using the rules for the expectation of a linear combination of random variables it is easy to calculate that

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} \cdot n \cdot \mu = \mu \ ,$$

with $E(X_i) = \mu$. This result holds under random sampling with or without replacement and is valid for any positive sample size n.

Variance of the Sample Mean $\bar{X}$

(1) Under random sampling with replacement

$$
\begin{aligned}
Var(\bar{X}) &= E[(\bar{X} - E(\bar{X}))^2] = E[(\bar{X} - \mu)^2] \\
&= E\left[\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\right)^2\right] \\
&= E\left[\left(\frac{1}{n}(X_1 - \mu) + \cdots + \frac{1}{n}(X_n - \mu)\right)^2\right] \\
&= \frac{1}{n^2}[E(X_1 - \mu)^2 + \cdots + E(X_n - \mu)^2 + \sum_i \sum_{j \neq i} E(X_i - \mu)(X_j - \mu)] \\
&= \frac{1}{n^2}[Var(X_1) + \cdots + Var(X_n) + \sum_i \sum_{j \neq i} Cov(X_i, X_j]
\end{aligned}
$$

For each $i = 1, \ldots, n$ holds $Var(X_i) = \sigma^2$. Furthermore, under random sampling with replacement the random variables are independent and therefore have $Cov(X_i, X_j) = 0$. The variance of the sample mean thus simplifies to

$$Var(\bar{X}) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}\,.$$

Note that the variance of $\bar{X}$ is equal to the variance of the population variable X divided by n. This implies that $Var(\bar{X})$ is smaller than $Var(X)$ and that $Var(\bar{X})$ is decreasing with increasing n. In other words, for large n the distribution of $\bar{X}$ is tightly concentrated around its expected value μ.

(2) Under random sampling without replacement

The derivation of $Var(\bar{X})$ in the case of random sampling without replacement is similar but more complicated because of the dependency of the random variables. Regarding the finite sample correction, for large populations the following approximation is quite accurate

$$\frac{N-n}{N-1} \approx \frac{N-n}{N}\,,$$

and the approximate correction $1 - n/N$ can be used. In sampling without replacement n cannot exceed N. For fixed n, the finite sample correction approaches 1 with increasing N :

$$\lim_{N \to \infty} \frac{N-n}{N-1} = 1\,.$$

In applications, the correction can be ignored if n is small relative to N. Rule of thumb: $n/N \leq 0.05$. However, this will only give an approximation to $Var(\bar{X})$.

On the Distribution of $\bar{X}$

Suppose that X follows a normal distribution in the population with expectation μ and variance σ^2: $X \sim N(\mu, \sigma^2)$.

In this case, the random variables $X_i, i = 1, \ldots, n$ are all normally distributed: $X_i \sim N(\mu, \sigma^2)$ for each $i = 1, \ldots, n$. The sum of n independent and identically normally distributed random variables also follows a normal distribution:

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)\,.$$

The statistic $\bar{X}$ differs from this sum only by the constant factor $1/n$ and, hence, is also normally distributed: $\bar{X} \sim N(\mu, \sigma^2(\bar{X}))$. Since only the standard normal distribution is tabulated the following standardized version of $\bar{X}$ is considered:

$$Z = \frac{\bar{X} - \mu}{\sigma(\bar{X})} = \sqrt{n}\frac{\bar{X} - \mu}{\sigma}\,,$$

which follows the standard normal distribution: $Z \sim N(0, 1)$. Evidently, using the standardized variable Z hinges on knowing the population variance σ^2. If the population variance σ^2 is unknown, the unknown variance σ^2 is estimated by

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Dividing both sides by σ^2 gives

$$\frac{S^2}{\sigma^2} = \frac{1}{\sigma^2}\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$
$$\frac{n-1}{\sigma^2}S^2 = \sum_{i=1}^{n}\frac{(X_i - \bar{X})^2}{\sigma}\,.$$

To simplify notation, set $Y = \frac{(n-1)S^2}{\sigma^2}$. In random sampling with replacement, the $X_i, i = 1, \ldots, n$ are independent and y is therefore the sum of squared independent standard normal random variables. It follows that Y is chi-square distributed with

degrees of freedom $n-1$. Using the standardized random variable Z to construct the ratio

$$T = \frac{Z}{\sqrt{\frac{Y}{n-1}}},$$

gives rise to the random variable T which follows the t-distribution with degrees of freedom $n-1$. (Recall from Chap. 6 that a t random variable is the ratio of a standard normal to the square root of an independent chi-square divided by its degrees of freedom.) Inserting the expressions for Z and Y and rearranging terms yield:

$$T = \frac{\sqrt{n}\frac{\bar{X}-\mu}{\sigma}}{\sqrt{\frac{1}{n-1}\left(\frac{n-1}{\sigma^2}S^2\right)}} = \sqrt{n}\frac{\bar{X}-\mu}{S}$$

Probability Statements About $\bar{X}$

If the sampling distribution of $\bar{X}$ including all its parameters are known, then probability statements about $\bar{X}$ can be made in the usual way. Suppose one wants to find a symmetric interval around the true mean which will contain $\bar{X}$ with probability $1-\alpha$. That is, we need to find c such that $P[\mu - c \leq \bar{x} \leq \mu + c] = 1-\alpha$.

It will be convenient to use the standardized random variable Z, the distribution of which we will assume to be symmetric.

$$\begin{aligned}
P(\mu - c \leq \bar{X} \leq \mu + c) &= 1-\alpha \\
P(-c \leq \bar{X} - \mu \leq c) &= 1-\alpha \\
P\left(\frac{-c}{\sigma(\bar{X})} \leq \frac{\bar{X}-\mu}{\sigma(\bar{X})} \leq \frac{c}{\sigma(\bar{X})}\right) &= 1-\alpha \\
P\left(\frac{-c}{\sigma(\bar{X})} \leq z \leq \frac{c}{\sigma(\bar{X})}\right) &= 1-\alpha \\
P\left(-z_{1-\frac{\alpha}{2}} \leq Z \leq z_{1-\frac{\alpha}{2}}\right) &= 1-\alpha \\
\frac{c}{\sigma(\bar{X})} &= z_{1-\frac{\alpha}{2}} \\
c &= z_{1-\frac{\alpha}{2}} \cdot \sigma(\bar{X})
\end{aligned}$$

Table 7.10 Examination for a course at the graduate level

Student	A	B	C	D	E	F	G
Score	10	11	11	12	12	12	16

Table 7.11 Frequency distribution of X="Number of points in the exam"

x	$h(x)$	$f(x) = h(x)/N$	$F(x)$
10	1	1/7	1/7
11	2	2/7	3/7
12	3	3/7	6/7
16	1	1/7	7/7

Thus, the deviation c from μ is a multiple of $\sigma(\bar{X})$. Inserting $\sigma(\bar{X})$ leads to the interval

$$\left[\mu - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

with probability

$$P\left(\mu - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

If X is normally distributed, then the central interval of variation with prespecified probability $1 - \alpha$ is determined by reading $z_{1-\alpha/2}$ from the standard normal table. The probability $1 - \alpha$ is approximately valid if X has an arbitrary distribution and the sample size n is sufficiently large.

Explained: Sampling Distribution

$N = 7$ students take part in an exam for a graduate course and obtain the scores given in Table 7.10.

The variable X = "score of an exam" has the population frequency distribution provided in Table 7.11.

The population parameters are $\mu = 12, \sigma^2 = 3.143$ and $\sigma = 1.773$.

Random Sampling with Replacement

$n = 2$ exams are sampled with replacement from the population. Table 7.12 contains all possible samples of size $n = 2$ with replacement and paying attention to the order of the draws.

For each possible sample, the sample mean can be calculated and is recorded in Table 7.13.

Table 7.12 Random sampling with replacement of size $n = 2$

	2. Exam						
1. Exam	10	11	11	12	12	12	16
10	10;10	10;11	10;11	10;12	10;12	10;12	10;16
11	11;10	11;11	11;11	11;12	11;12	11;12	11;16
11	11;10	11;11	11;11	11;12	11;12	11;12	11;16
12	12;10	12;11	12;11	12;12	12;12	12;12	12;16
12	12;10	12;11	12;11	12;12	12;12	12;12	12;16
12	12;10	12;11	12;11	12;12	12;12	12;12	12;16
16	16;10	16;11	16;11	16;12	16;12	16;12	16;16

Table 7.13 Sample means for all possible samples of size $n = 2$ with replacement

	2. Exam						
1. Exam	10	11	11	12	12	12	16
10	10.0	10.5	10.5	11.0	11.0	11.0	13.0
11	10.5	11.0	11.0	11.5	11.5	11.5	13.5
11	10.5	11.0	11.0	11.5	11.5	11.5	13.5
12	11.0	11.5	11.5	12.0	12.0	12.0	14.0
12	11.0	11.5	11.5	12.0	12.0	12.0	14.0
12	11.0	11.5	11.5	12.0	12.0	12.0	14.0
16	13.0	13.5	13.5	14.0	14.0	14.0	16.0

Table 7.14 Distribution of the sample mean for samples of size $n = 2$ with replacement

$\bar{x}$	$P(\bar{x})$	$\bar{x} - E(\bar{X})$	$[\bar{x} - E(\bar{X})]^2$	$[\bar{x} - E(\bar{X})]^2 \cdot P(\bar{x})$
10.0	1 / 49	−2.0	4.00	4 / 49
10.5	4 / 49	−1.5	2.25	9 / 49
11.0	10 / 49	−1.0	1.00	10 / 49
11.5	12 / 49	−0.5	0.25	3 / 49
12.0	9 / 49	0.0	0.00	0
13.0	2 / 49	1.0	1.00	2 / 49
13.5	4 / 49	1.5	2.25	9 / 49
14.0	6 / 49	2.0	4.00	24 / 49
16.0	1 / 49	4.0	16.00	16 / 49

$\bar{X}$ therefore can take on various values with certain probabilities. From Table 7.13 the distribution of $\bar{X}$ can be determined as given in the first two columns of Table 7.14.

The mean of this distribution, i.e., the expected value of $\bar{X}$, is given by

$$E(\bar{X}) = 588/49 = 12\,.$$

which is equal to the expected value of the variable X in the population: $E(X) = 12$. Using the intermediate results in columns three to five of Table 7.14 allows one to

Table 7.15 Random sampling without replacement of size $n = 2$

	2. Exam						
1. Exam	10	11	11	12	12	12	16
10		10;11	10;11	10;12	10;12	10;12	10;16
11	11;10		11;11	11;12	11;12	11;12	11;16
11	11;10	11;11		11;12	11;12	11;12	11;16
12	12;10	12;11	12;11		12;12	12;12	12;16
12	12;10	12;11	12;11	12;12		12;12	12;16
12	12;10	12;11	12;11	12;12	12;12		12;16
16	16;10	16;11	16;11	16;12	16;12	16;12	

Table 7.16 Sample means for all possible samples of size $n = 2$ without replacement

	2. Exam						
1. Exam	10	11	11	12	12	12	16
10		10.5	10.5	11.0	11.0	11.0	13.0
11	10.5		11.0	11.5	11.5	11.5	13.5
11	10.5	11.0		11.5	11.5	11.5	13.5
12	11.0	11.5	11.5		12.0	12.0	14.0
12	11.0	11.5	11.5	12.0		12.0	14.0
12	11.0	11.5	11.5	12.0	12.0		14.0
16	13.0	13.5	13.5	14.0	14.0	14.0	

calculate the variance of $\bar{X}$:

$$Var(\bar{X}) = \sigma^2(\bar{X}) = 77/49 = 11/7 = 1.5714$$

This result is in agreement with the formula for $\sigma^2(\bar{X})$ given above:

$$\sigma^2(\bar{X}) = \sigma^2/n = (22/7)/2 = 11/7\,.$$

It is easy to see that the variance of $\bar{X}$ is indeed smaller than the variance of X in the population.

Random Sampling Without Replacement

From the population, $n = 2$ exams are randomly drawn without replacement. Table 7.15 displays all possible samples of size $n = 2$ from sampling without replacement, paying attention to the order of the draws. For each possible sample, the sample mean is calculated and reported in Table 7.16.

Table 7.17 Distribution of the sample mean for samples of size $n = 2$ without replacement

$\bar{x}$	$P(\bar{x})$	$\bar{x} - E(\bar{X})$	$[\bar{x} - E(\bar{X})]^2$	$[\bar{x} - E(\bar{X})]^2 \cdot P(\bar{x})$
10.5	4 / 42	−1.5	2.25	9 / 42
11.0	8 / 42	−1.0	1.00	8 / 42
11.5	12 / 42	−0.5	0.25	3 / 42
12.0	6 / 42	0.0	0.00	0
13.0	2 / 42	1.0	1.00	2 / 42
13.5	4 / 42	1.5	2.25	9 / 42
14.0	6 / 42	2.0	4.00	24 / 42

The first two columns of Table 7.17 contain the probability distribution of the sample mean. The expected value $E(\bar{X})$ is

$$E(\bar{X}) = 504/42 = 12$$

and is equal to the expected value of X in the population.

The variance is equal to

$$Var(\bar{X}) = \sigma^2(\bar{X}) = 55/42 = 1.3095\,,$$

which is in agreement with the formula for calculating $\sigma^2(\bar{x})$ given earlier:

$$\begin{aligned} Var(\bar{X}) = \sigma^2(\bar{X}) &= \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} \\ &= \frac{22/7}{2} \cdot \frac{7-2}{7-1} = \frac{22 \cdot 5}{7 \cdot 2 \cdot 6} = \frac{55}{42}\,. \end{aligned}$$

Enhanced: Gross Hourly Earnings of a Worker

This example is devoted to formally explaining the sampling distribution of the sample mean, its expectation, and variance. To this end, certain assumptions must be made about the population. In particular, it is assumed that the mean hourly gross earnings of all 5000 workers of a company equals \$27.30 with a standard deviation of \$5.90 and variance of \$34.81.

Problem 1

Suppose that the variable X = "Gross hourly earnings of a (randomly selected) worker in this company" is normally distributed. That is, $X \sim N(27.3; 34.81)$.

From the population of all workers of this company, a random sample (with replacement) of n workers is selected. The sample mean gives the average gross hourly earnings of the n workers in the sample.

Calculate the expected value, variance, and standard deviation, and find the specific form of the distribution of $\bar{X}$ for the following sample sizes:

a) $n = 10$,
b) $n = 50$
c) $n = 200$.

Expected Value Regardless of n, the expected value of $\bar{X}$ is

$$E(\bar{X}) = \mu = \$27.30$$

Variance and Standard deviation The variance of the sample mean is equal to

$$Var(\bar{X}) = \sigma^2(\bar{X}) = \sigma^2/n\,.$$

Thus,

a) for a random sample of size $n = 10$
$Var(\bar{X}) = \sigma^2(\bar{X}) = 5.9^2/10 = 34.81/10 = 3.481$
$\sigma(\bar{X}) = \$1.8657$.
b) for a random sample of size $n = 50$
$Var(\bar{X}) = \sigma^2(\bar{X}) = 5.9^2/50 = 34.81/50 = 0.6962$
$\sigma(\bar{X}) = \$0.8344$
c) for a random sample of size $n = 200$
$Var(\bar{X}) = \sigma^2(\bar{X}) = 5.9^2/200 = 34.81/200 = 0.17405$
$\sigma(\bar{X}) = \$0.4172$.

Obviously, the standard deviation of $\bar{X}$ is smaller than the standard deviation of X in the population. Moreover, the standard deviation of $\bar{X}$ decreases from 1.8657 to 0.8344 and to 0.4172, as the sample size is increased from 10 to 50 and eventually to 200. Increasing the sample size by a factor of five cuts the standard deviation roughly by half. Increasing the sample size twentyfold reduces the standard deviation by more than 3/4.

Sampling Distribution of $\bar{X}$ Since X is assumed to be normally distributed it follows that the sample mean $\bar{X}$ is also normally distributed under random sampling with replacement, regardless of the sample size. Thus:

a) for random samples of size $n = 10$:

$$\bar{X} \sim N(27.3; 3.481)$$

The red curve in Fig. 7.1 corresponds to the distribution of X in the population while the blue curve depicts the distribution of the sample mean $\bar{X}$.

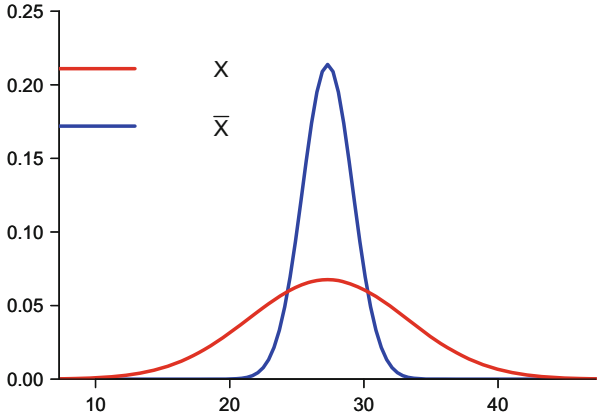


Fig. 7.1 The distribution of X in the population (red) and the distribution of the sample mean for $n = 10$

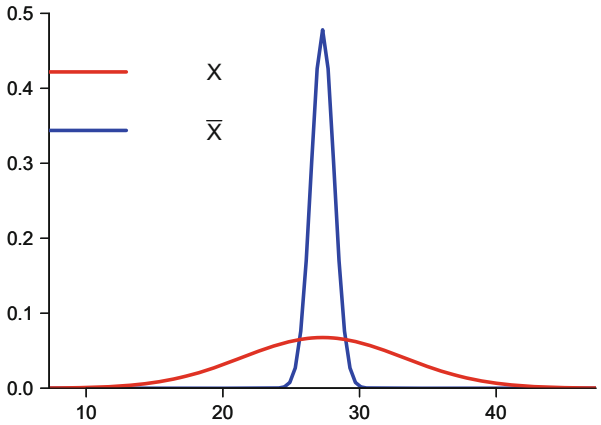


Fig. 7.2 The distribution of X in the population (red) and the distribution of the sample mean for $n = 50$

b) for random samples of size $n = 50$ (Fig. 7.2):

$$\bar{X} \sim N(27.3; 0.6962)$$

c) for random samples of size $n = 200$ (Fig. 7.3):

$$\bar{X} \sim N(27.3; 0.17405)$$

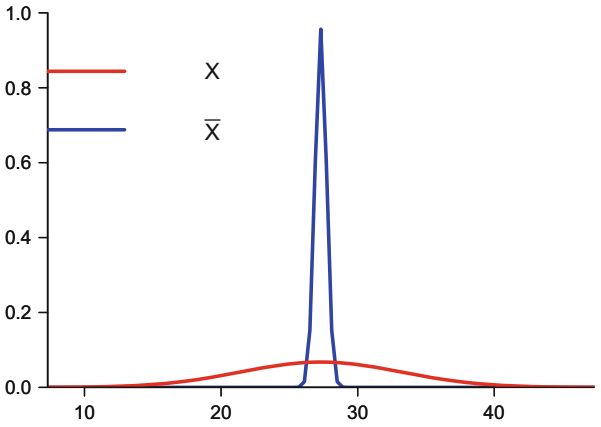


Fig. 7.3 The distribution of X in the population (red) and the distribution of the sample mean for $n = 200$

Problem 2

Suppose that the variable X = "gross hourly earnings of a (randomly selected) worker of this company" is normally distributed, say, $X \sim N(27.3;\ 34.81)$.

A sample of size n is randomly drawn without replacement. The sample mean gives the gross hourly earnings of the n workers in the sample. Calculate the expected value, variance, and standard deviation of $\bar{X}$ for the following sample sizes:

a) $n = 10$,
b) $n = 50$
c) $n = 1000$.

Expected Value All random samples without replacement, regardless of n, have the same expected value as in the first problem:

$$E(\bar{X}) = \mu = \$27.30$$

Variance and Standard deviation In the case of sampling without replacement, the variance of the sample mean is reduced by a "finite population correction factor." Specifically, the variance of the sample mean is given by

$$Var(\bar{X}) = \sigma^2(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}.$$

However, the finite population correction can be neglected if n is sufficiently small relative to N for example if $n/N \leq 0.05$. Thus,

a) for a random sample with replacement of size $n = 10$:
Since $n/N = 10/5000 = 0.002 < 0.05$ the variance can be calculated sufficiently accurately using $Var(\bar{X}) = \sigma^2(\bar{X}) = \sigma^2/n$. This leads to the same result as in problem 1:

$$Var(\bar{X}) = \sigma^2(\bar{x}) = 5.9^2/10 = 34.81/10 = 3.481,$$
$$\sigma(\bar{X}) = \$1.8657.$$

In comparison, the finite population correction yields $Var(\bar{X}) = \sigma^2(\bar{X}) = 3.4747$ and $\sigma(\bar{X}) = \$1.8641$, which demonstrates the negligibility of the correction.

b) for a random sample with replacement of size $n = 50$:
Since $n/N = 50/5000 = 0.01 < 0.05$ the variance can be calculated using $Var(\bar{X}) = \sigma^2(\bar{X}) = \sigma^2/n$. This leads to the same result as in problem 1:

$$Var(\bar{X}) = \sigma^2(\bar{X}) = 5.9^2/50 = 34.81/50 = 0.6962,$$
$$\sigma(\bar{X}) = \$0.8344,$$

which is very similar to the finite sample corrected result $\sigma(\bar{X}) = \$0.8303$.

c) for a random sample with replacement of size $n = 1000$:
Since $n/N = 1000/5000 = 0.2 > 0.05$ the variance and standard deviation should be calculated using the finite population correction:

$$Var(\bar{X}) = \sigma^2(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$
$$= \frac{5.9^2}{1000} \cdot \frac{5000-1000}{5000-1} = 0.0279$$
$$\sigma(\bar{X}) = \$0.1669.$$

Problem 3

Suppose that, more realistically, the distribution of X = "gross hourly earnings of a (randomly selected) worker from this company" is unknown. Hence, all that is known is $E(X) = \mu = \$27.0$ and $\sigma(X) = \$5.90$.

A sample of size n is randomly drawn. The sample mean gives the gross hourly earnings of the n workers in the sample. Calculate the expected value, variance, and standard deviation, and find the specific form of the distribution of $\bar{X}$ for the

following sample sizes:

a) $n = 10$,
b) $n = 50$
c) $n = 200$.

Expected Value How the expected value $E(\bar{X})$ is calculated does not depend on the distribution of X in the population. Hence, there are no new aspects in the present situation and the results are identical to the previous two problems:

$$E(\bar{X}) = \mu = \$27.30$$

Variance and Standard deviation How the variance of $\bar{X}$ is calculated does not depend on the distribution of X in the population but it does depend on the type and size of the random sample.

In the statement of problem 3 the sampling scheme has not been specified. However, for all three sample sizes $n/N < 0.05$ and, hence, if the sample is drawn without replacement the formula $Var(\bar{X}) = \sigma^2(\bar{X}) = \sigma^2/n$ can be used as an approximation.

for $n = 10$:	$Var(\bar{X}) = \sigma^2(\bar{X}) = 3.481$,	$\sigma(\bar{X}) = \$1.8657$
for $n = 50$:	$Var(\bar{X}) = \sigma^2(\bar{X}) = 0.6962$,	$\sigma(\bar{X}) = \$0.8344$
for $n = 200$:	$Var(\bar{X}) = \sigma^2(\bar{X}) = 0.17405$,	$\sigma(\bar{X}) = \$0.4172$

Sampling Distribution of $\bar{X}$ Since the distribution of X in the population is unknown no exact statement can be made about the distribution of $\bar{X}$. However, the central limit theorem implies that the standardized random variable Z

$$Z = \sqrt{n}\frac{\bar{X} - \mu}{\sigma}$$

is approximately standard normal if the sample size $n > 30$ and—in random sampling without replacement—the size of the population N is sufficiently large. This is satisfied for the cases b) $n = 50$ and c) $n = 200$.

7.3 Distribution of the Sample Proportion

Consider a dichotomous population with two types of elements and that the proportion of elements with property A is π while the proportion of elements that do not have property A is $1 - \pi$.

Randomly selecting an element for this population gives rise to a random variable that takes on the value 1 if the selected element possesses property A , and that takes

on the value 0 otherwise. n draws produce n random variables $X_1, \ldots, X_n$ all of which can only take on the values 1 or 0.

Let X denote the number of elements in the sample of size n with property A (i.e., X is equal to the absolute frequency of elements with property A in the sample). Then,

$$\hat{\pi} = \frac{X}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is the proportion (i.e., the relative frequency) of elements in the sample of size n with property A (sample proportion).

After actually drawing a sample a specific number x of sample elements with property A is observed and the sample proportion takes on the realization $\hat{\pi} = X/n$.

X and $\hat{\pi}$ vary from sample to sample (even if the sample size n is fixed). They are statistics (i.e., functions of the sample) and consequently random variables. Their sampling distributions, expected values, and variances will be determined below. The sampling distributions depend crucially on

- how the sample is drawn (with or without replacement) and
- the size of the population.

1. Random sampling with replacement: This corresponds to conducting n Bernoulli-experiments. All sample variables have the following distribution

$$f(x_i, \pi) = \begin{cases} \pi^{x_i}(1-\pi)^{1-x_i} & \text{if } \; x_i = 0,\ 1 \\ 0 & \text{otherwise} \end{cases}$$

with expectation $E(X_i) = \pi$ and variance $Var(X_i) = \pi \cdot (1-\pi)$.

In this case X follows a binomial distribution with parameters n and π: $X \sim B(n; \pi)$:

$$f_B(x|n;\pi) = \begin{cases} \binom{n}{x} \pi^x (1-\pi)^{n-x} & \text{if } \; x = 0, 1, \ldots, n \\ \\ 0 & \text{otherwise} \end{cases}$$

with

$$E(X) = n \cdot \pi, \; Var(X) = \sigma^2(X) = n \cdot \pi \cdot (1-\pi)$$

Since $\hat{\pi} = X/n$ and $1/n$ is just a constant factor it follows that the sample proportion $\hat{\pi}$ has a probability function closely related to that of X. Expected value and variance of $\hat{\pi}$ are equal to:

$$E(\hat{\pi}) = E\left(\frac{X}{n}\right) = \frac{E(X)}{n} = \frac{(n \cdot \pi)}{n} = \pi$$

$$Var(\hat{\pi}) = \sigma^2(\hat{\pi}) = Var\left(\frac{X}{n}\right) = \frac{Var(X)}{n^2} = \frac{n \cdot \pi \cdot (1-\pi)}{n^2} = \frac{\pi \cdot (1-\pi)}{n}$$

Approximation Note that the sample proportion $\hat{\pi} = \sum X_i/n$ is a mean of n independent Bernoulli random variables. Thus, one may use the central limit theorem, to conclude that (for a sufficiently large sample size n) its distribution and the distribution of X (which is Binomial) can be approximated by the normal distribution:

$$X \approx N(\mu, \sigma^2), \text{ with } \mu = E(X) = n \cdot \pi \text{ and } \sigma^2 = \sigma^2(X) = n \cdot \pi \cdot (1-\pi)$$

and

$$\hat{\pi} \approx N(\mu, \sigma^2), \text{ with } \mu = E(\hat{\pi}) = \pi \text{ and } \sigma^2 = \sigma^2(\hat{\pi}) = \pi \cdot (1-\pi)/n,$$

respectively. The sample size is considered to be large enough for a sufficiently good approximation if $n \cdot \pi \geq 5$ and $n \cdot (1-\pi) \geq 5$.

To obtain an improved approximation, the continuity correction may be used, i.e., for calculating $P(x_1 \leq X \leq x_2)$ using the standard normal distribution one applies

$$z_1 = \frac{x_1 - 0.5 - np}{\sqrt{np(1-p)}} \qquad z_2 = \frac{x_2 + 0.5 - np}{\sqrt{np(1-p)}}$$

and for the probability $P(p_1 \leq \hat{\pi} \leq p_2)$

$$z_1 = \frac{\frac{np_1 - 0.5}{n} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{p_1 - \frac{1}{2n} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \qquad z_2 = \frac{\frac{np_2 - 0.5}{n} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{p_2 - \frac{1}{2n} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \; .$$

2. random sampling without replacement

The distinction between sampling with and without replacement is only relevant for finitely sized populations. Let N denote the size of the population, M denote the number of elements in the population with property A, and n denote the sample size. Then $\pi = M/N$ is the proportion of elements in the population with property A. The statistics X and $\hat{\pi}$ are defined as before.

Under sampling without replacement X follows the hypergeometric distribution with parameters N, M and n: $X \sim H(N, M, n)$:

$$f_H(x; N, M, n) = \begin{cases} \dfrac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} & \text{if} \quad x \in [|a, b|] \\ \\ 0 & \text{otherwise} \end{cases}$$

where $a = max[0, n - (N - M)]$ and $b = min[n, M]$.

The expected value and variance of the hypergeometric variable X are given by:

$$E(X) = n \cdot \frac{M}{N} = n\pi$$

$$Var(X) = \sigma^2(X) = n\pi(1-\pi)\frac{N-n}{N-1} = n \cdot \frac{M}{N} \cdot \frac{N-M}{N} \cdot \frac{N-n}{N-1}$$

The statistic $\hat{\pi}$ has a distribution function closely related to that of $X = n \cdot \hat{\pi}$. The expectation and variance of $\hat{\pi}$ are :

$$E(\hat{\pi}) = \frac{1}{n}E(X) = \pi$$

$$Var(\hat{\pi}) = \sigma^2(\hat{\pi}) = \frac{1}{n^2}\sigma^2(X) = \frac{\pi(1-\pi)}{n} \cdot \frac{N-n}{N-1}$$

Approximations For large N and M and small n/N the hypergeometric distribution can be approximated reasonably well by the binomial distribution with $\pi = M/N$. A rule of thumb is: $n/N \leq 0.05$.

According to the central limit theorem, for sufficiently large sample sizes, the hypergeometric distribution may be approximated by the normal distribution even under sampling without replacement.

$$X \approx N(\mu, \sigma^2), \text{ with } \mu = E(X) = n \cdot \pi \text{ and } \sigma^2 = \sigma^2(X)$$

and

$$\hat{\pi} \approx N(\mu, \sigma^2), \text{ with } \mu = E(\hat{\pi}) = \pi \text{ and } \sigma^2 = \sigma^2(\hat{\pi}),$$

respectively. The sample size is considered to be sufficiently large if $nM/N \geq 5$, $n(1 - M/N) \geq 5$ and $n/N \leq 0.05$. For a more accurate approximation the continuity correction may be used.

Explained: Distribution of the Sample Proportion

According to Germany's Bureau of Statistics there were 37.3 million private households in Germany in April 1996, 35 % of which were one-person households.

Problem 1

From this population, $n = 10$ households are randomly selected without replacement.

- What is the distribution of X (number of one-person households in the sample) and $\hat{\pi}$ (proportion of one-person households in the sample)?
- Obtain the expectation, variance, and standard deviation of this distribution.
- What is the probability that the proportion of one-person households in the sample is larger than 0.2 but smaller than 0.5?

Of the $N = 37.3$ million private households in the (finitely sized) population $M = 13.055$ million are one-person households. Randomly selecting $n = 10$ households gives rise to 10 random variables $X_i, i = 1, \ldots, 10$ which take on the value $X_i = 1$ if the i-th household selected is a one-person household and $X_i = 0$ otherwise. The random variable X, being the sum of the 10 sample variables, gives the number of one-person households in the sample while $\hat{\pi} = X/n$ gives their proportion in the sample. Under sampling without replacement X is hypergeometrically distributed: $X \sim H(N; M; n) = H(37.3 \text{ million}; 13,055 \text{ million}; 10)$.

The statistic $\hat{\pi}$ has a probability function closely related to that of $X = n \cdot \hat{\pi}$. Since the population size N is very large and since $n/N = 10/(37.3 \cdot 10^6) < 0.05$ is very small, the finiteness of the population can be ignored and the binomial distribution with $\pi = M/N = 0.35$ can be used as an approximation: $X \approx B(n; \pi) = B(10; 0.35)$.

Expectation	Variance	St. deviation
$E(X) = 10 \cdot 0.35 = 3.5$	$Var(X) = 10 \cdot 0.35 \cdot 0.65 = 2.275$	$\sigma(X) = 1.5083$
$E(\hat{\pi}) = 0.35$	$Var(\hat{\pi}) = 0.35 \cdot 0.65/10 = 0.02275$	$\sigma(\hat{\pi}) = 0.1508$

The desired probability $P(0.2 < \hat{\pi} < 0.5)$ is found as follows: since $X = n \cdot \hat{\pi}$, it follows that $x_1 = 10 \cdot 0.2 = 2$ and $x_2 = 10 \cdot 0.5 = 5$, the desired probability is equal to $P(2 < X < 5)$.

$$\begin{aligned} P(2 < X < 5) &= P(X \leq 4) - P(X \leq 2) = F_B(4) - F_B(2) \\ &= 0.7515 - 0.2616 = 0.4899\,, \end{aligned}$$

where $F_B(4)$ and $F_B(2)$ can be obtained from a table of the binomial distribution $B(10; 0.35)$.

Problem 2

From the population described above a sample of size $n = 2000$ is drawn without replacement.

- What is the distribution of the number and the proportion of one-person households, respectively?
- Give their expectation, variance, and standard deviation.
- What is the probability that the number of one-person households in the sample is greater than or equal to 700 but less than or equal to 725, i.e., $P(700 \leq X \leq 725)$?

The statistics X and $\hat{\pi}$ are defined as in problem 1. Since the population is very large and the sample small relative to the population, it is irrelevant whether the sample has been drawn with or without replacement and the binomial distribution can be used as an approximation.

Expectation	Variance	St. deviation
$E(X) = 2000 \cdot 0.35 = 700$	$Var(X) = 2000 \cdot 0.35 \cdot 0.65 = 455$	$\sigma(X) = 21.33$
$E(\hat{\pi}) = 0.35$	$Var(\hat{\pi}) = 0.35 \cdot 0.65/2000 = 0.000114$	$\sigma(\hat{\pi}) = 0.01067$

There is no table for the distribution function of the binomial distribution $B(2000; 0.35)$ and a computer was used to calculate:

$$P(700 \leq X \leq 725) = P(X \leq 725) - P(X < 700) = F_B(725) - F_B(699)$$
$$= 0.8839 - 0.4916 = 0.3923$$

Since the sample size $n = 2000$ is very large and the criteria $n \cdot \pi = 2000 \cdot 0.35 = 700 \geq 5$ and $n(1-\pi) = 2000 \cdot 0.65 = 1300 \geq 5$ are satisfied, the normal distribution can be used to approximate the binomial distribution:

$$X \approx N(700; 21.33)\,, \qquad \hat{\pi} \approx N(0.35; 0.01067)\,.$$

With

$$z_1 = \frac{700 - 0.5 - 700}{21.33} = -0.02344\,, \qquad z_2 = \frac{725 + 0.5 - 700}{21.33} = 1.1955$$

it follows that

$$\begin{aligned} P(700 \leq X \leq 725) &\approx \Phi(1.1955) - \Phi(-0.02344) \\ &= \Phi(1.1955) - (1 - \Phi(0.02344)) \\ &= 0.884054 - (1 - 0.509351) = 0.3934 \end{aligned}$$

which is close to the exact calculation using the binomial distribution.

Enhanced: Drawing Balls from an Urn

From an urn with N balls, and a proportion of π red balls, samples of size n are drawn without replacement. Calculate the probability of obtaining samples with proportions of red balls between p_1 and p_2.

Problem 1

From a population of size $N = 5$ and $\pi = 0.4$ a sample of size $n = 3$ is drawn without replacement.

The random variable X which is a sum of the 3 random variables provides the number of red balls in the sample and the random variable $\hat{\pi} = X/n$ gives the proportion of red balls in the sample.

- What is the distribution of the number and the proportion of red balls in the sample, respectively?
- What is the probability that the proportion of red balls in the sample is between 1/3 and 2/3?

Because the population is finitely sized and the sampling is done without replacement it follows that the statistic X has a hypergeometric distribution: $X \sim H(N; M; n) = H(5; 2; 3)$, and $M = 0.4 \cdot 5 = 2$ (Fig. 7.4).

We want to calculate $P(1/3 \leq \hat{\pi} \leq 2/3)$. Since $X = n \cdot \hat{\pi}$, it follows that $x_1 = 3 \cdot 1/3 = 1$ and $x_2 = 3 \cdot 2/3 = 2$. Hence we are interested in $P(1 \leq X \leq 2)$:

$$P(1 \leq X \leq 2) = f(1) + f(2) = 0.9 \quad \text{where } f(1) = 0.6 \text{ and } f(2) = 0.3\,.$$

Problem 2

From a population of size $N = 1000$ and proportion $\pi = 0.2$ samples of size $n = 4$ are drawn without replacement. The random variable X is a sum of the 4 random

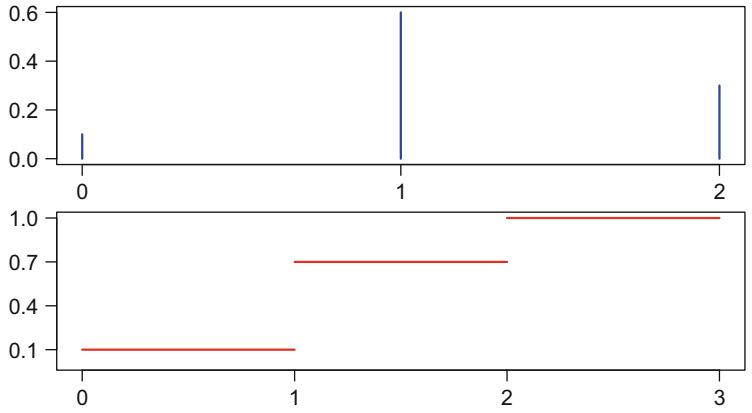


Fig. 7.4 The distribution of the number of red balls in the sample for $n = 3$

variables and gives the number of red balls in the sample. The random variable $\hat{\pi} = X/n$ gives the proportion of red balls in the sample.

- What is the distribution of the number and the proportion of red balls in the sample, respectively?
- What is the probability that the proportion of red balls in the sample is between 0.25 and 0.75?

Because the sample is drawn without replacement and the population size is finite, X follows the hypergeometric distribution: $X \sim H(1000; 200; 4)$.

Since the population is very large and since $n/N = 0.004 < 0.05$, X is approximately binomially distributed with parameter $\pi = M/N = 0.2$, i.e., $X \approx B(4; 0.2)$ (Fig. 7.5). We may use this probability distribution to calculate probabilities for $\hat{\pi}$.

We are interested in $P(0.25 \leq \hat{\pi} \leq 0.75)$. Since $X = n \cdot \hat{\pi}$ and therefore $x_1 = 4 \cdot 0.25 = 1$ and $x_2 = 4 \cdot 0.75 = 3$, the desired probability in terms of X is $P(1 \leq X \leq 3)$:

$$P(1 \leq X \leq 3) = F_B(3) - F_B(0) = 0.9984 - 0.4096 = 0.5888$$

$F_B(3)$ and $F_B(0)$ can be obtained from the table of the distribution function of the binomial distribution $B(4; 0.2)$.

Problem 3

From a population of size $N = 2500$ and with proportion $\pi = 0.2$ samples of size $n = 100$ are drawn without replacement. The random variable X is a sum of

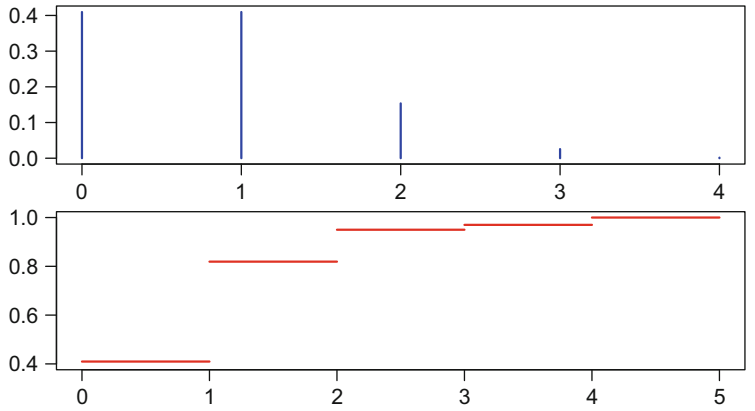


Fig. 7.5 The distribution of the number of red balls in the sample for $n = 4$

100 random variables and gives the number of red balls in the sample. The random variable $\hat{\pi} = X/n$ gives the proportion of red balls in the sample.

- What is the distribution of the number and the proportion of red balls in the sample, respectively?
- What is the probability that the proportion of red balls in the sample is between 0.14 and 0.3?

Because the sample is drawn without replacement and the population size is finite, X follows the hypergeometric distribution: $X \sim H(2500; 500; 100)$.

Since the sample size $n = 100$ is large and the criteria $n \cdot M/N = 100 \cdot 0.2 = 20 \geq 5$, $n(1 - M/N) = 80 \geq 5$ and $n/N = 0.04 < 0.05$ are satisfied, the normal distribution can be used with:

$$\begin{aligned} E(\hat{\pi}) &= \pi = 0.2, \\ Var(\hat{\pi}) &= [\pi(1-\pi)/n] \cdot [(N-n)/(N-1)] = 0.001537, \\ \sigma(\hat{\pi}) &= 0.039 \approx 0.04. \end{aligned}$$

Hence, the hypergeometric distribution is approximated by the normal distribution $N(0.2; 0.0015)$. To keep matters simple, the continuity correction is neglected (Fig. 7.6). The desired probability $P(0.14 \leq \hat{\pi} \leq 0.3)$ can be calculated by using $z_1 = (0.3 - 0.2)/0.04 = 2.5$ and. $z_2 = (0.14 - 0.2)/0.04 = -1.5$ which leads to

$$\begin{aligned} P(0.14 \leq \hat{\pi} \leq 0.3) &= \Phi(2.5) - \Phi(-1.5) = \Phi(2.5) - (1 - \Phi(1.5)) \\ &= 0.99379 - (1 - 0.933193) = 0.9269\,. \end{aligned}$$

$\Phi(2.5)$ and $\Phi(1.5)$ are obtained from a table of the standard normal distribution.

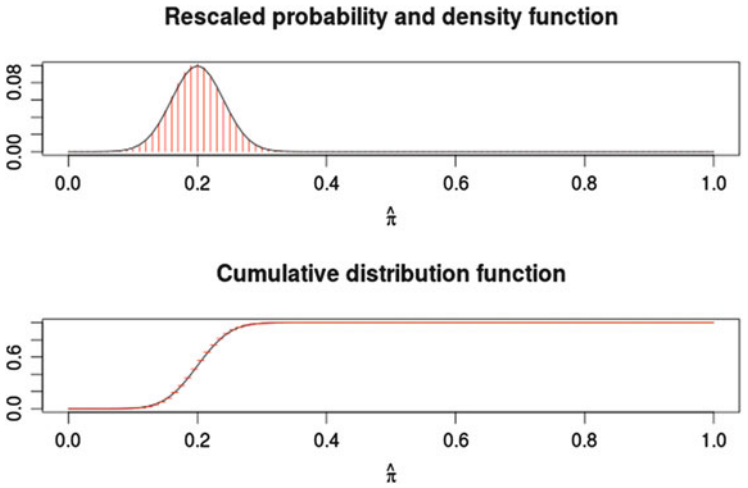


Fig. 7.6 Distribution of the number and the proportion of red balls in the sample of size $n = 100$

7.4 Distribution of the Sample Variance

Consider a population variable X with $E(X) = \mu$ and $Var(X) = \sigma^2$. From this population a random sample of size n is drawn.

The sample variance is based on the sum of squared deviations of the random variables $X_i, i = 1, \ldots, n$ from the mean. We have proposed two estimators for the variance, the *MSD* and s^2.

Since $E(X) = \mu$ is usually unknown and estimated by the sample mean $\bar{x}$, the sample variance is calculated as

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

Alternatively, the sample variance may also be calculated as

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

See the entry under "Information" for more on this version of the sample variance.

The derivation of the distribution of the sample variances s^2 will be given for the case of a normally distributed population, i.e., $X \sim N(\mu.\sigma^2)$.

Under these assumptions, the random variables $X_i, i = 1, \ldots, n$ are independently and identically normally distributed with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$:

$$X_i \sim N(\mu, \sigma) \quad i = 1, \ldots, n$$

Moreover, the sample mean $\bar{X}$ is also normally distributed with $E(\bar{x}) = \mu$ and $Var(\bar{X}) = \sigma^2(\bar{X}) = \sigma^2/n$:

$$\bar{X} \sim N(\mu, \sigma^2)\,.$$

Distribution of the Sample Variance S^2

Consider for the moment the random variable

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 .$$

It is the sum of squares of n independent standard normals, and hence has a chi-square distribution with n degrees of freedom, i.e., χ^2_n. Now consider

$$\frac{(n-1)s^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{x})^2 = \sum_{i=1}^{n} \left(\frac{X_i - \bar{x}}{\sigma} \right)^2 .$$

and note the similarity. By now using $\bar{x}$ as an estimator of μ it can be shown that we have a sum of squares of $n-1$ independent standard normals in which case $(n-1)s^2/\sigma^2$ is chi-square distributed with $n-1$ degrees of freedom. The distribution of s^2 is a simple rescaling of $(n-1)s^2/\sigma^2$. Thus we may make probability statements about s^2.

Using the properties of the chi-square distribution, the expected value and variance of S^2 are:

$$E(S^2) = \sigma^2, \qquad Var(S^2) = 2\sigma^4/(n-1)$$

Probability Statements About S^2

For known variance σ^2 and a normally distributed population one can calculate the probability that the sample variance S^2 will take on values in a central interval with prespecified probability $1 - \alpha$.

$$P\left(v_1 \leq \frac{(n-1)S^2}{\sigma^2} \leq v_2 \right) = 1 - \alpha$$

Furthermore, if we want to put equal probability mass in the tails, i.e., we impose:

$$P\left(\frac{(n-1)S^2}{\sigma^2} < v_1\right) = \frac{\alpha}{2}; \quad P\left(\frac{(n-1)S^2}{\sigma^2} > v_2\right) = \frac{\alpha}{2}$$

With $n-1$ degrees of freedom, the interval boundaries can be obtained from tables of the chi-square distribution

$$v_1 = \chi^2_{\frac{\alpha}{2};n-1}; \quad v_2 = \chi^2_{1-\frac{\alpha}{2};n-1}$$

Thus,

$$P\left(\chi^2_{\frac{\alpha}{2};n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{1-\frac{\alpha}{2};n-1}\right) = 1-\alpha$$

Rearranging yields the probability statement:

$$P\left(\frac{\sigma^2\chi^2_{\frac{\alpha}{2};n-1}}{n-1} \leq S^2 \leq \frac{\sigma^2\chi^2_{1-\frac{\alpha}{2};n-1}}{n-1}\right) = 1-\alpha$$

More Information

μ Is Known

Consider the simplifying assumption that μ is known and let us modify S^{*2} as follows:

$$S^{*2} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$$

Derivation of the expected value of S^{*2}:

$$\begin{aligned} E(S^{*2}) &= E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right] = \frac{1}{n}E\left[\sum_{i=1}^{n}(X_i-\mu)^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}E[(X_i-\mu)^2] \;= \frac{1}{n}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n}n\sigma^2 \\ &= \sigma^2 \end{aligned}$$

Note that the above argument does not assume a distribution for the X_i. It is only assumed that they are i.i.d. with common variance $Var(X_i) = E[(X_i-\mu)^2] = \sigma^2$.

Derivation of the Variance of S^{*2} In this case we assume that the X_i are i.i.d. $N(\mu, \sigma^2)$ Recall that the variance of a chi-square random variable with n degrees of freedom has mean n and variance $2n$. Since nS^{*2}/σ^2 has a chi-square distribution with n degrees of freedom, it follows that:

$$Var\left(\frac{nS^{*2}}{\sigma^2}\right) = \frac{n^2}{\sigma^2} Var(S^{*2}) = 2n$$

and therefore

$$Var(S^{*2}) = \frac{2\sigma^4}{n}\,.$$

Note also that we can derive the mean of s^{*2} using:

$$E\left(\frac{nS^{*2}}{\sigma^2}\right) = n$$

and therefore

$$E(S^{*2}) = \sigma^2\,.$$

μ Is Unknown

Since μ is typically unknown, the usual estimator of the variance is given by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Derivation of the Expectation of S^2 Recall that the variance of a random variable can be written as:

$$\begin{aligned} Var(X) &= E[(X - E(X))^2] = E[X^2 - 2XE(X) + (E(X))^2] \\ &= E(X^2) - 2E(X)E(X) + [E(X)]^2 \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

This implies that

$$E(X^2) = Var(X) + [E(X)]^2$$

Applying this result to the X_i and to $\bar{x}$ we have:

$$E(X_i^2) = Var(X_i) + [E(X_i)]^2 = \sigma^2 + \mu^2$$

$$E(\bar{x}^2) = Var(\bar{x}) + [E(\bar{x})]^2 = \frac{\sigma^2}{n} + \mu^2$$

Furthermore,

$$\begin{aligned}
E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] &= E\left[\sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2\right] \\
&= E\left[\sum_{i=1}^{n} X_i^2 - 2n\bar{X}^2 + n\bar{X}^2\right] \\
&= E\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right] = E\left[\sum_{i=1}^{n} X_i^2\right] - E\left[n\bar{X}^2\right] \\
&= \sum_{i=1}^{n} E(X_i^2) - nE(\bar{X}^2) = \sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) \\
&= n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2 \\
&= (n-1)\sigma^2
\end{aligned}$$

Therefore, the expectation of the sample variance S^2 is given by

$$\begin{aligned}
E(S^2) &= E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{1}{n-1}E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \\
&= \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2.
\end{aligned}$$

Once again, this argument does not require the assumption of normality, only that the X_i are i.i.d. with common variance σ^2.

Derivation of the Variance of S^2 In this case we assume that the X_i are i.i.d. $N(\mu, \sigma^2)$. Since $(n-1)S^2/\sigma^2$ has a chi-square distribution with $n-1$ degrees of freedom, it follows that

$$Var\left(\frac{(n-1)S^2}{\sigma^2}\right) = \frac{(n-1)^2}{\sigma^4}Var(S^2) = 2(n-1)$$

and therefore

$$Var(S^2) = \frac{2\sigma^4}{(n-1)}.$$

μ Is Unknown

In this case we use the *MSD* to estimate the variance:

$$MSD = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 .$$

Note that

$$MSD = \frac{n-1}{n}S^2 .$$

Hence

$$E(MSD) = \frac{n-1}{n}E\left[S^2\right] = \frac{n-1}{n}\sigma^2$$

and

$$Var(MSD) = \left(\frac{n-1}{n}\right)^2 Var\left[S^2\right] = \left(\frac{n-1}{n}\right)^2 \frac{2\sigma^4}{(n-1)} = \frac{n-1}{n^2}2\sigma^4$$

Note that the expectation of the *MSD* is not exactly equal to the population variance σ^2 which is the reason that the sample variance s^2 is usually used in practical applications. Nevertheless, even for moderately sized samples, the two estimates will be similar.

Explained: Distribution of the Sample Variance

To measure the variation in time needed for a certain task, the variance is often utilized. Let the time a worker needs to complete a certain task be the random variable X. Suppose X is normally distributed with $E(X) = \mu$ and $Var(X) = \sigma^2$.

A random sample of size n is drawn with replacement. The random variables X_i $(i = 1, \ldots, n)$ are therefore independent and identically normally distributed.

Problem 1

A random sample of size $n = 15$ is drawn. What is the probability that the sample variance S^2 will take on values in the interval $[0.5 \cdot \sigma^2; 1.5 \cdot \sigma^2]$? That is, the probability to be calculated is $P(0.5\sigma^2 \leq S^2 \leq 1.5\sigma^2)$.

To solve the problem, each side is multiplied by $(n-1)/\sigma^2$:

$$P(0.5\sigma^2 \leq S^2 \leq 1.5\sigma^2) = P\left(\frac{n-1}{\sigma^2}0.5\sigma^2 \leq \frac{n-1}{\sigma^2}S^2 \leq \frac{n-1}{\sigma^2}1.5\sigma^2\right)$$
$$= P\left((n-1)\cdot 1.5 \leq \frac{n-1}{\sigma^2}S^2 \leq (n-1)\cdot 1.5\right)$$

Since $n-1=14$ it follows that:

$$P(0.5\sigma^2 \leq S^2 \leq 1.5\sigma^2) = P\left(7 \leq \frac{n-1}{\sigma^2}S^2 \leq 21\right)$$

The probability that S^2 will take on values between $0.5\cdot\sigma^2$ and $1.5\cdot\sigma^2$is identical to the probability that the transformed random variable $(n-1)S^2/\sigma^2$ will take values between 7 and 21.

The random variable $(n-1)S^2/\sigma^2$ is chi-square $n-1=14$ degrees of freedom. The probability can be found by using a table of the chi-square distribution.

$$P(0.5\sigma^2 \leq S^2 \leq 1.5\sigma^2) = P\left(7 \leq \frac{n-1}{\sigma^2}S^2 \leq 21\right)$$
$$= P\left(\frac{n-1}{\sigma^2}S^2 \leq 21\right) - P\left(\frac{n-1}{\sigma^2}S^2 \leq 7\right)$$
$$= 0.8984 - 0.0653 = 0.8331$$

The probability that S^2 will lie in the interval $[0.5\cdot\sigma^2$ and $1.5\cdot\sigma^2]$ is equal to 0.8331.

Figure 7.7 shows the density function of the chi-square distribution with 14 degrees of freedom, where the symbol Y is a shorthand for $(n-1)S^2/\sigma^2$.

Problem 2

The goal is to determine a central interval of variation for the sample variance S^2 with prespecified probability $1-\alpha = 0.95$ We assume the same population as in problem 1 and use a random sample of size $n=30$. Since

$$P\left(v_1 \leq \frac{(n-1)S^2}{\sigma^2} \leq v_2\right) = 0.95$$

and we again put equal probability mass in the tails:

$$P\left(\frac{(n-1)S^2}{\sigma^2} \leq v_1\right) = 0.025\,; \qquad P\left(\frac{(n-1)S^2}{\sigma^2} \leq v_2\right) = 0.975$$

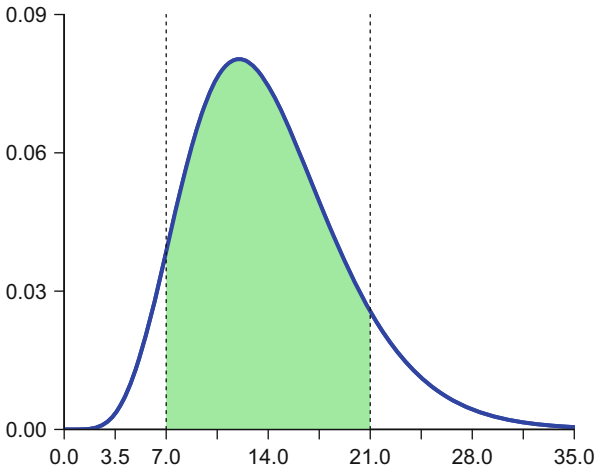


Fig. 7.7 Density function of the chi-square distribution with 14 degrees of freedom

Using tables for the chi-square distribution with 29 degrees of freedom we obtain $v_1 = 16.05$ and $v_2 = 45.72$. Thus,

$$P\left(16.05 \leq \frac{(n-1)S^2}{\sigma^2} \leq 45.72\right) = 0.95$$

With probability 0.95, the transformed random variable $(n-1)S^2/\sigma^2$ takes values in the interval [16.05; 45.72]. Rearranging gives the interval:

$$P\left(\frac{16.05\sigma^2}{n-1} < S^2 < \frac{45.72\sigma^2}{n-1}\right) = 0.95$$

$$P(0.5534\sigma^2 < S^2 < 1.5766\sigma^2) = 0.95$$

With a probability of 0.95 the sample variance S^2 will take values in the interval $[0.5534\sigma^2; 1.5766\sigma^2]$. The exact numerical boundaries of the interval can be determined only if the population variance σ^2 of the variable X is known.

Chapter 8
Estimation

8.1 Estimation Theory

Assume a given population with distribution function $F(x)$. In general, the distribution and its characteristics or parameters are not known. Suppose we are interested in say the expectation μ and the variance σ^2. (Alternatively, if the data are binary, we may be interested in the population proportion π). As outlined previously, we can learn about the population or equivalently its distribution function F, through (random) sampling. The data may then be used to infer properties of the population, hence the term indirect inference. At the outset, it is important to emphasize that the conclusions drawn may be incorrect, particularly if the sample is small, or not representative of the underlying population. The tools of probability may be used to provide measures of the accuracy or correctness of the estimates or conclusions. We will focus on the estimation of unknown parameters or characteristics. Assume θ to be the object of interest, then we differentiate two types of procedures: point estimation and interval estimation.

Point Estimation

The determination of a single estimate using a random sample is referred to as point estimation. It is desirable that the estimate provides the best possible approximation to the unknown parameter.

The Estimator or Estimating Function

We will be drawing n independent observations from the population. In that case $X_i, i = 1, .., n$ are i.i.d. random variables. The estimator is defined to be a function g

W.K. Härdle et al., *Introduction to Statistics*, DOI 10.1007/978-3-319-17704-5_8

of the X_i. We write

$$\hat{\theta} = g(\bullet)\,,$$

the estimator of θ, in which case it is a random variable. The symbol will also represent a specific estimate for a given dataset. It should be clear from the context which applies.

A point estimate thus depends on the sample size n and the realizations that have been drawn. The point estimate will rarely correspond to the true value of the unknown parameter. Indeed, repeated sampling will generally yield different estimates. If the sample size is large, we would expect these to be close to the true parameter value.

A crucial problem of point estimations is the selection of the best estimator. In some cases, the population parameter or characteristic of interest has a natural sample analogue. For example, one typically uses the sample mean to estimate the population mean, the sample proportion to estimate the population proportion and the sample variance to estimate the population variance. (See e.g., the discussion in Sect. 7.1.)

Explained: Basic Examples of Estimation Procedures

Given a supposed population of $N = 2000$ households let the random variable X be household net income (in EUR). The mean net income of this population, i.e., the expectation $E(X) = \mu$ is unknown and the subject of our estimation. The sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,.$$

is used. A random sample of size n yields the sample values $x_1, \ldots, x_n$.

a) A random sample of $n = 20$ private households yields the data given in Table 8.1

From the data we obtain $\bar{x} = 48300/20 = 2415$ EUR. As can be easily seen, the calculation is identical with the arithmetic mean, a measure which we already used in descriptive statistics. An important objective of inductive statistics is to provide a measure of the accuracy of this result as an estimate of the underlying population mean.

To illustrate the point, we obtain 24 further random samples of the size $n = 20$. Table 8.2 tabulates the sample means for the 25 samples.

In Table 8.2 the samples are reordered so that the sample means are in increasing order. Evidently, there is considerable variation in the sample means, which illustrates the random character of estimation, in particular that the estimator $\bar{x}$ is a random variable.

Table 8.1 Data on household net income

i	Households net income (EUR) x_i	i	Households net income (EUR) x_i
1	800	11	2500
2	1200	12	2500
3	1400	13	2500
4	1500	14	2700
5	1500	15	2850
6	1500	16	3300
7	1800	17	3650
8	1800	18	3700
9	2300	19	4100
10	2400	20	4300

Table 8.2 Mean household net income (EUR)

Sample	$\bar{x}$	Sample	$\bar{x}$	Sample	$\bar{x}$
1	1884.90	10	2241.15	18	2395.25
2	1915.30	11	2243.15	19	2413.40
3	2060.90	12	2267.75	20	2415.00
4	2062.15	13	2298.80	21	2567.50
5	2110.30	14	2317.00	22	2607.25
6	2126.50	15	2319.55	23	2635.00
7	2163.10	16	2361.25	24	2659.00
8	2168.50	17	2363.50	25	2774.30
9	2203.85				

Consequently point estimates need to be supplemented with a measure of their precision (e.g., by giving the standard deviation of the estimator).

Figure 8.1 displays the estimated values $\bar{x}$ of the 25 samples. In order to depict the deviation of the estimated values from the true mean of the population, the actual value μ is illustrated as a dashed line.

b) From the same population 100 random samples of size $n = 100$ were drawn and mean household net incomes were calculated. The results are provided in Fig. 8.2. The actual value μ appears as a dashed line.

8.2 Properties of Estimators

When estimating a specific parameter or characteristic of a population, several possible estimators $\hat{\theta}$ exist.

Example 1 Suppose that the underlying population distribution is symmetric. In this case the population expectation equals the population's median. Thus the unknown expectation can be estimated using either the sample mean or the sample median.

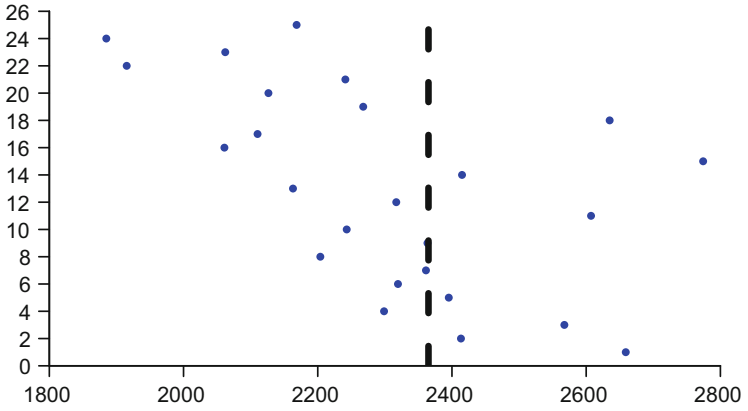


Fig. 8.1 Estimated values of $\bar{x}$ from 25 random samples of size $n = 20$

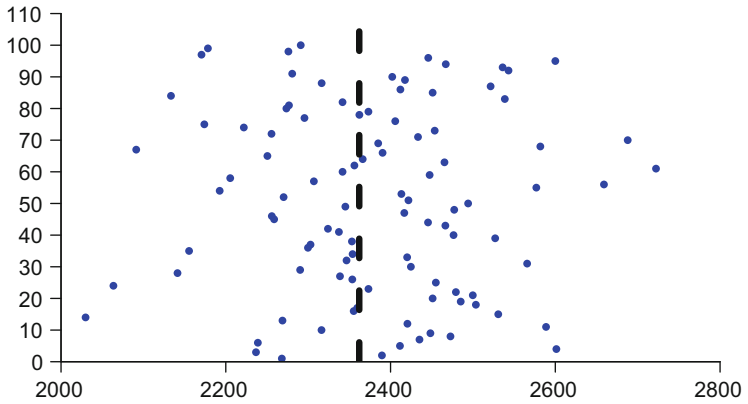


Fig. 8.2 Estimated values of $\bar{x}$ from 100 random samples of size $n = 100$

In general, the two estimators will provide different estimates. Which estimator should be used?

Example 2 To estimate the variance σ^2 we may use either of the following:

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

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

Which estimator should be used?

Example 3 Suppose that the underlying population distribution is Poisson. For the Poisson distribution $E(X) = Var(X) = \lambda$. Therefore, the unknown parameter λ could be estimated using the sample mean or the sample variance. Again in this case the two estimators will in general yield different estimates.

In order to obtain an objective comparison, we need to examine the properties of the estimators.

Mean Squared Error

A general measure of the accuracy of an estimator is the Mean Squared Deviation, or Mean Squared Error (MSE). The *MSE* measures the average squared distance between the estimator $\hat{\theta}$ and the true parameter θ:

$$MSE = E[(\hat{\theta} - \theta)^2] .$$

It is straightforward to show that the *MSE* can be separated into two components:

$$MSE = E[(\hat{\theta} - \theta)^2] = E[(\hat{\theta} - E(\hat{\theta}))^2] + [E(\hat{\theta}) - \theta]^2 .$$

The first term on the right side is the variance of $\hat{\theta}$:

$$E[(\hat{\theta} - E(\hat{\theta}))^2] = Var(\hat{\theta}) ,$$

The second term is the square of the bias $E(\hat{\theta}) - \theta$. Hence, the MSE is the sum of the variance and the squared bias of the estimator:

$$MSE = Var(\hat{\theta}) + [bias\,(\hat{\theta})]^2 .$$

If several estimators are available for an unknown parameter of the population, one would thus select that one with the smallest *MSE*.

Starting with the *MSE* three important properties of estimators are described, which should facilitate the search for the "best" estimator.

Unbiasedness

An estimator $\hat{\theta}$ of the unknown parameter θ is unbiased, if the expectation of the estimator matches the true parameter value:

$$E(\hat{\theta}) = \theta .$$

That is, the mean of the sampling distribution of $\hat{\theta}$ equals the true parameter value θ.

For an unbiased estimator the *MSE* equals the variance of the estimator:

$$MSE = Var(\hat{\theta}).$$

Thus the variance of the estimator provides a good measure of the precision of the estimator.

If the estimator is biased, then the expectation of the estimator is different from the true parameter value. That is,

$$bias(\hat{\theta}) = E(\hat{\theta}) - \theta \neq 0\,.$$

Asymptotic Unbiasedness

An estimator $\hat{\theta}$ is called asymptotically unbiased, if

$$\lim_{n\to\infty} E(\hat{\theta}) = \theta\,,$$

i.e., the bias converges to zero with increasing sample size n.

Efficiency

Often there are several unbiased estimators available for the same parameter. In this case, one would like to select the one with the smallest variance (which in this case is equal to the *MSE*).

Let $\hat{\theta}_n$ and $\hat{\theta}_n^\star$ be two unbiased estimators of θ using a sample of size n. The estimator $\hat{\theta}_n$ is called relatively efficient in comparison to $\hat{\theta}_n^\star$, if the variance of $\hat{\theta}_n$ is smaller than the variance of $\hat{\theta}_n^\star$, i.e.,

$$Var(\hat{\theta}_n) < Var(\hat{\theta}_n^\star)\,.$$

The estimator $\hat{\theta}_n$ is called efficient if its variance is smaller than that of any other unbiased estimator.

Consistency

The consistency of an estimator is a property which focuses on the behavior of the estimator in large samples. In particular consistency requires that the estimator be close to the true parameter value with high probability in large samples. It is sufficient if the bias and variance of the estimator converge to zero. Formally, suppose

$$\lim_{n\to\infty}[E(\hat{\theta}_n) - \theta] = 0$$

and

$$\lim_{n\to\infty} Var(\hat{\theta}_n) = 0$$

Then the estimator is consistent. Equivalently, the two conditions may be summarized using:

$$\lim_{n\to\infty} MSE(\hat{\theta}_n) = 0\,.$$

This notion of consistency is also referred to as "squared mean consistency."

An alternative version known as weak consistency is defined as follows:

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \epsilon) = 1$$

That is, the probability that the estimator $\hat{\theta}_n$ yields values within an arbitrarily small interval around the true parameter value θ converges to one with increasing sample size n. The probability that the estimator $\hat{\theta}_n$ differs from the true parameter value by more than ϵ converges to zero with increasing sample size n. That is,

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| \geq \epsilon) = 0$$

More Information

Mean Squared Error

Recall the *MSE* is defined as

$$MSE = E[(\hat{\theta} - \theta)^2]$$

Expanding the expression one obtains:

$$\begin{aligned} MSE &= E[(\hat{\theta} - \theta)^2] \\ &= E[(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2] \\ &= E[(\hat{\theta} - E(\hat{\theta}))^2 + 2(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta) + (E(\hat{\theta}) - \theta)^2] \\ &= E[(\hat{\theta} - E(\hat{\theta}))^2] + 2E[(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta)] + [E(\hat{\theta}) - \theta)^2] \,. \end{aligned}$$

For the middle term we have:

$$2E[(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta)] = 2[E(\hat{\theta}) - E(\hat{\theta})][E(\hat{\theta}) - \theta)] = 0$$

and consequently we have

$$\begin{aligned} MSE &= E[(\hat{\theta} - E(\hat{\theta}))^2] + [E(\hat{\theta}) - \theta]^2 \\ &= Var(\hat{\theta}) + [bias(\hat{\theta})]^2 \,. \end{aligned}$$

The *MSE* does not measure the actual estimation error that has occurred in a particular sample. It measures the average squared error that would occur in repeated sample.

Unbiasedness

Figure 8.3 displays three estimators of a parameter θ. The estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased since their expectation coincides with the true parameter θ (denoted by

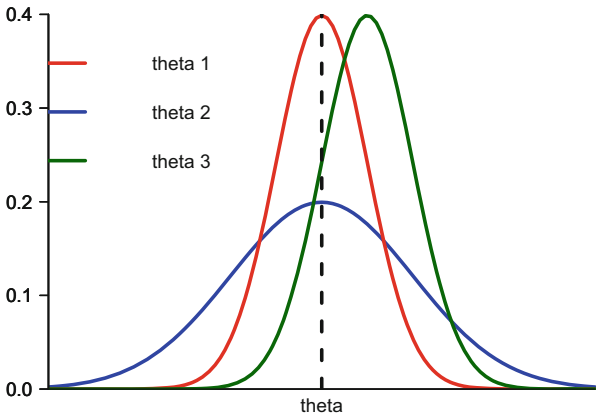


Fig. 8.3 Illustration of unbiasedness

the vertical dashed line). In contrast, the estimator $\hat{\theta}_3$ is biased. For both unbiased estimators

$$MSE = Var(\hat{\theta}) ,$$

holds, as the bias equals zero. However $\hat{\theta}_1$ has lower variance and is therefore preferred to $\hat{\theta}_2$. It is also preferred to $\hat{\theta}_3$ which has the same variance but exhibits substantial positive bias. Each of the following widely used estimators are unbiased.

Sample Mean $\bar{x}$

The sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is an unbiased estimator of unknown expectation $E(X) = \mu$ since

$$E(\bar{X}) = \mu$$

See section *Distribution of the Sample Mean*.

Sample Proportion $\hat{\pi}$

The sample proportion

$$\hat{\pi} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is an unbiased estimator for the population proportion π since

$$E(\hat{\pi}) = \pi ,$$

See section *Distribution of the Sample Fraction*.

Sample Variance

Assume a random sample of size n.

1. If the expectation $E(X) = \mu$ of the population is unknown and estimated using the sample mean, the estimator

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

is an unbiased estimator of σ^2, since

$$E(s^2) = \sigma^2 ,$$

See section *Distribution of the Sample Variance*. The standard deviation which is the square root of the sample variance s^2 is not an unbiased estimator of σ, as it tends to underestimate the population standard deviation. This result can be proven by means of Jensen's inequality.

The estimator:

$$MSD = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2$$

is not unbiased, since

$$E(MSD) = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2\right] = \frac{1}{n}E\left[\sum_{i=1}^{n}(X_i - \bar{x})^2\right] = \frac{n-1}{n}\sigma^2 .$$

See section *Distribution of the Sample Variance*. The bias is given by:

$$E(MSD) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n} .$$

Using the estimator *MSD* one will tend to underestimate the unknown variance. The estimator, however, is asymptotically unbiased as with increasing sample size n the bias converges to zero. Division by $n-1$, (as in s^2) rather than by n (as in the *MSD*) assures unbiasedness.

Efficiency

- The sample mean $\bar{x}$ is an efficient estimator of the unknown population expectation μ. This is true for any distribution.
- Suppose data are drawn from a $N(\mu; \sigma^2)$ distribution. The sample mean $\bar{x}$ is an efficient estimator of μ. It can be shown that no unbiased estimator of μ exists which has a smaller variance.
- The sample mean $\bar{x}$ is an efficient estimator for the unknown parameter λ of a Poisson distribution.

- The sample proportion $\hat{\pi}$ is an efficient estimator of the unknown population proportion π for a dichotomous population, i.e., the underlying random variables have a common Bernoulli distribution.
- For a normally distributed population the sample mean $\bar{x}$ and the sample median *md* are unbiased estimators of the unknown expectation μ. For random samples (with replacement) we have:

$$\sigma^2(\bar{x}) = \frac{\sigma^2}{n}$$

Furthermore one can show that

$$\sigma^2(md) = \frac{\pi}{2}\frac{\sigma^2}{n} = 1.571\ \sigma^2(\bar{x})$$

and hence

$$\sigma^2(\bar{x}) < \sigma^2(md)\,.$$

The sample mean $\bar{x}$ is relatively efficient in contrast to the sample median *md*.
- The relative efficiency of various estimators of the same parameter in general depends on the distribution from which one is drawing observations.

Consistency

- Consistency is usually considered to be a minimum requirement of an estimator. Of course, consistency does not preclude the estimator having large bias and variance in small or moderately sized samples. Consistency only guarantees that bias and variance go to zero for sufficiently large samples. On the other hand, since sample size cannot usually be increased at will, consistency may provide a poor guide to the finite sample properties of the estimator.
- For random samples, the sample mean $\bar{x}_n$ is a consistent estimator of the population expectation μ since $bias\ \bar{x}_n = 0$ and the variance $Var(\bar{x}_n) = \sigma^2/n$ converge to zero, i.e.,

$$\lim_{n\to\infty} \frac{\sigma^2}{n} = 0\,.$$

- For random samples the sample proportion $\hat{\pi}_n$ is a consistent estimator for the population proportion π as the estimator is unbiased $bias\ \hat{\pi}_n = 0$ and the variance $Var(\hat{\pi}_n) = \pi(1-\pi)/n$ converges to zero, i.e.,

$$\lim_{n\to\infty} \frac{\pi(1-\pi)}{n} = 0\,.$$

- For a Gaussian distributed population the sample median *md* is a consistent estimator for the unknown parameter μ.
- For a Gaussian distribution, the estimator

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

is consistent for the unknown variance σ^2, since the estimator is unbiased *bias* $s^2 = 0$ and the variance $Var(s^2) = 2\sigma^4/(n-1)$ converges to zero:

$$\lim_{n\to\infty} \frac{2\sigma^4}{n-1} = 0\,.$$

The sample variance is also a consistent estimator of the population variance for arbitrary distributions which have a finite mean and variance.

Explained: Properties of Estimators

Assume a population with mean μ and variance σ^2. Let (X_1, X_2, X_3) be a random sample drawn from the population. Each random variable X_i, $i = 1, 2, 3$ has $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Consider the following three estimators of the population mean:

1. $\hat{\mu}_1 = \frac{1}{3}(X_1 + X_2 + X_3)$
2. $\hat{\mu}_2 = \frac{1}{4}(2X_1 + 2X_3)$
3. $\hat{\mu}_3 = \frac{1}{3}(2X_1 + X_2)$

- Which estimators are unbiased?
- Which estimator is most efficient?

All of them are unbiased, since $E(X_i) = \mu$:

$$\begin{aligned}
E(\hat{\mu}_1) &= E\big[\frac{1}{3}(X_1 + X_2 + X_3)\big] = \frac{1}{3}[E(X_1) + E(X_2) + E(X_3)] \\
&= \frac{1}{3}(\mu + \mu + \mu) = \mu \\
E(\hat{\mu}_2) &= E\big[\frac{1}{4}(2X_1 + 2X_3)\big] = \frac{1}{4}[2E(X_1) + 2E(X_3)] \\
&= \frac{1}{4}(2\mu + 2\mu) = \mu \\
E(\hat{\mu}_3) &= E\big[\frac{1}{3}(2X_1 + X_2)\big] = \frac{1}{3}[2E(X_1) + E(X_2)] \\
&= \frac{1}{3}(2\mu + \mu) = \mu
\end{aligned}$$

The variance of each estimator is given by:

$$\begin{aligned}
Var(\hat{\mu}_1) &= Var\Big[\frac{1}{3}(X_1 + X_2 + X_3)\Big] = \frac{1}{9}Var(X_1 + X_2 + X_3) \\
&= \frac{1}{9}[Var(X_1) + Var(X_2) + Var(X_3)] = \frac{1}{9}(\sigma^2 + \sigma^2 + \sigma^2) = \frac{1}{3}\sigma^2 \\
Var(\hat{\mu}_2) &= Var\Big[\frac{1}{4}(2X_1 + 2X_3)\Big] = \frac{1}{16}Var(2X_1 + 2X_3) \\
&= \frac{1}{16}[4Var(X_1) + 4Var(X_3)] = \frac{1}{16}(4\sigma^2 + 4\sigma^2) = \frac{1}{2}\sigma^2 \\
Var(\hat{\mu}_3) &= Var\Big[\frac{1}{3}(2X_1 + X_2)\Big] = \frac{1}{9}Var(2X_1 + X_2) \\
&= \frac{1}{9}[4Var(X_1) + Var(X_2)] = \frac{1}{9}(4\sigma^2 + \sigma^2) = \frac{5}{9}\sigma^2
\end{aligned}$$

Since we use all the data, the first estimator is the most efficient. This estimator is of course the sample mean. Note that even though the second and third estimators each use two observations, the third is less efficient than the second because it does not weight the observations equally.

Enhanced: Properties of Estimation Functions

The unknown mean $E(X) = \mu$ and variance σ^2 will be estimated. A random sample of size $n = 12$ was drawn from a population yielding the following data: {1; 5; 3; 8; 7; 2; 1; 4; 3; 5; 3; 6}.

Estimation of the Expectation

The sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i\,,$$

is an unbiased and efficient estimator. Substituting the sample values yields

$$\bar{x} = \frac{1}{12}(1 + 5 + 3 + 8 + 7 + 2 + 1 + 4 + 3 + 5 + 3 + 6) = \frac{48}{12} = 4\,.$$

This result constitutes a point estimate of μ.

Estimation of the Variance

The estimator is given by:

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

Substituting the sample values yields the point estimate

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{12}(x_i - \bar{x})^2 \\ &= \frac{1}{11}[(1-4)^2 + (5-4)^2 + \cdots + (3-4)^2 + (6-4)^2] = \frac{1}{11}\cdot 56 = 5.09\,. \end{aligned}$$

8.3 Construction of Estimators

In this section we will discuss principles for constructing estimators of an unknown parameter. We have already seen how sample moments can be used to estimate their population counterparts (e.g., the sample mean, variance, or proportion are used to estimate the corresponding population mean, variance, or proportion). This principle is known as the method of moments. We now consider two other principles: maximum likelihood and least squares.

Maximum Likelihood

The maximum likelihood approach is one of the most important estimation procedures. The essential idea is to find the probability law—within a prespecified family—which is most likely to have generated the observed data.

Assume a discrete resp. continuous random variable X having the probability resp. density function $f(x|\theta)$ in the population. An important prerequisite of the maximum likelihood method is that the type of distribution must be known prior to estimation. The distribution depends on an unknown parameter θ.

Example 1 Suppose we are drawing from a binomial distribution. Hence, the probability function $f(x|\theta)$ is $B(n;\pi)$, which depends on the unknown parameter π.

Example 2 Suppose we are drawing from a normal distribution. Then, the probability density function $f(x|\theta)$ is $N(\mu;\sigma^2)$ which depends on the unknown parameters μ and σ^2

A random sample of size n is drawn from a distribution $f(x|\theta)$. Thus, the random variables X_i $i = 1, .., n$ are independent and identically distributed with probability law: $f(x_i|\theta)$ $\forall$ $i = 1, \ldots, n$. Since the observations are independent, the joint distribution of all the random variables equals the product of their individual distributions, i.e.,

$$P(\{X_1 = x_1\} \cap \cdots \cap \{X_n = x_n\}|\theta) = f(x_1, \ldots, x_n|\theta) = f(x_1|\theta) \cdot \cdots \cdot f(x_n|\theta) .$$

Given the sample, we may now ask the question, what is the probability (or probability density) of having drawn this sample for different values of the unknown parameter θ. Mathematically, we define the likelihood function $L(\theta)$ to be a function of θ conditional on the data $(x_1, \ldots, x_n)$. That is,

$$L(\theta|x_1, \ldots, x_n) = f(x_1|\theta) \cdot \cdots \cdot f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) .$$

$L(\theta)$ gives the probability (or probability density) for the realized sample $(x_1, \ldots, x_n)$ at each value of θ.

The maximum likelihood principle states that one should select the value $\hat{\theta}$, which maximizes the likelihood function:

$$L(\hat{\theta}) = \max_{\theta} L(\theta) .$$

Under general conditions $L(\bullet)$ has a maximum. A necessary condition is that the first derivative equals zero:

$$\frac{\partial L(\hat{\theta})}{\partial \theta} = 0 .$$

For simplicity, it is common to take the logarithm yielding the log-likelihood function $\log L(\hat{\theta})$. Since the logarithm constitutes a monotone transformation, the maximum of $\log(L(\hat{\theta}))$ occurs at the same value of $\hat{\theta}$ as for the original likelihood function. The first order condition becomes:

$$\frac{\partial \log L(\hat{\theta})}{\partial \theta} = 0 .$$

The resulting maximum likelihood estimator $\hat{\theta}$ has been studied widely and is known to have many favorable properties under general conditions. Among them, it is consistent, asymptotically normal, and efficient in large samples.

Least Squares Estimation

Suppose that the expectations of the random variables $X_1, \ldots, X_n$ depend on the unknown parameter θ through known functions g_i:

$$E(X_i) = g_i(\theta) \qquad i = 1, \ldots, n$$

In the simplest case $g_i(\theta) = \theta \ \forall \ i$.

Given data $x_1, \ldots, x_n$, then an estimator $\hat{\theta}$ may be chosen by minimizing the sum of squared deviation of the data from $g_i(\hat{\theta})$; i.e.,

$$Q(\theta) = \sum_{i=1}^{n} (x_i - g_i(\theta))^2$$

has to be minimized. The solution may be found by differentiating with respect to θ and setting the first derivative equal to zero. The resulting minimizer $\hat{\theta}$ is called the least squares estimator. Least squares estimators have favorable properties. They are typically consistent, asymptotically normal, and efficient in large samples.

More Information

Applications of ML

ML Estimation of μ and σ^2 for the Gaussian Distribution

Assume a Gaussian distributed random variable X with the unknown parameters μ and σ^2. Assume further that $X_1, \ldots, X_n$ is a random sample drawn from this distribution. Then for each $X_i, i = 1, \ldots, n$ we have:

$$f(x_i|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

The likelihood function is given by:

$$\begin{aligned} L(\mu, \sigma^2|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\mu, \sigma) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \cdot exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right). \end{aligned}$$

Taking logarithms yields:

$$\log L(\mu, \sigma^2 | x_1, \ldots, x_n) = -\frac{n}{2} \cdot \log(2\pi) - \frac{n}{2} \cdot \log \sigma^2 - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2 .$$

ML Estimation of μ and σ^2

To maximize $L(\mu; \sigma^2)$ for given $(x_1, \ldots, x_n)$, $\hat{\mu}$ and $\hat{\sigma}^2$ are chosen to maximize the log-likelihood function. Taking the partial derivatives with respect to μ and σ^2, and setting the resulting equations equal to zero yields:

$$\frac{\partial \log L}{\partial \mu} = -\frac{2 \cdot \sum_{i=1}^{n} (x_i - \hat{\mu}) \cdot (-1)}{2\sigma^2} = 0$$

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2} \cdot \frac{1}{\hat{\sigma}^2} + \frac{1}{2} \cdot \frac{1}{\hat{\sigma}^4} \cdot \sum_{i=1}^{n} (x_i - \mu)^2 = 0 .$$

The first equation may be solved to produce the ML estimator $\hat{\mu}$ of μ :

$$\sum_{i=1}^{n} (x_i - \hat{\mu}) = 0$$

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x} .$$

The result may be substituted in the second equation which may then be solved for $\hat{\sigma}^2$:

$$\frac{n}{2\hat{\sigma}^2} = \frac{1}{2\hat{\sigma}^4} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 ,$$

Thus the ML estimator of μ is the sample mean, which we know to be consistent, unbiased, and efficient. The ML estimator of σ^2 is the *MSD*, which though biased in finite samples is consistent and asymptotically efficient. Second order conditions for a maximum can be readily verified.

ML Estimation of π for a Binomial Distribution

Suppose one is drawing a random sample (with replacement) of size n from a dichotomous population with unknown parameter π. Let X be the number of successes. Then X is binomially distributed $B(n, \pi)$. The likelihood function is given by:

$$L(\pi|x) = \binom{n}{x} \cdot \pi^x \cdot (1-\pi)^{n-x}$$

and the log-likelihood is given by

$$\log L(\pi|x) = \log \binom{n}{x} + x \log \pi + (n-x) \log(1-\pi) .$$

Differentiating with respect to π and setting to zero one obtains:

$$\begin{aligned} \frac{\partial \log L(\pi|x)}{\partial \pi} = \frac{x}{\hat{\pi}} - \frac{n-x}{1-\hat{\pi}} &= 0 \\ x(1-\hat{\pi}) - (n-x)\hat{\pi} &= 0 \\ \hat{\pi} &= \frac{x}{n} \end{aligned}$$

To verify that this is a maximum, we check that the second derivative is negative:

$$\frac{\partial^2 \log L(\pi|x)}{\partial \pi^2} = -\frac{x}{\pi^2} - \frac{n-x}{(1-\pi)^2}$$

The ML estimator is the sample proportion $\hat{\pi}$ which is unbiased, consistent, and efficient.

ML Estimation of λ in a Poisson Distribution

Let $X_1, \ldots, X_n$ be a random sample (with replacement) of size n from a Poisson distribution with unknown parameter $\lambda > 0$. Then, for each $X_i, i = 1, \ldots, n$ we have

$$f_{PO}(x_i; \lambda) = \frac{\lambda^{x_i}}{x_i !} e^{-\lambda}.$$

The likelihood function for the realized sample $x_1, \ldots, x_n$ is given by

$$L(\lambda|x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i !} e^{-\lambda} = \frac{\lambda^{x_1 + \cdots + x_n}}{x_1 ! \cdot \ldots \cdot x_n !} e^{-n\lambda}$$

and the log-likelihood becomes:

$$\log L(\lambda|x_1,\ldots,x_n) = \sum_{i=1}^{n} \log\left(\frac{\lambda^{x_i}}{x_i\,!}e^{-\lambda}\right) = \sum_{i=1}^{n}(x_i \log\lambda - \log(x_i\,!) - \lambda)\,.$$

Differentiating with respect to λ and setting the expression equal to zero yields

$$\frac{\partial \log L}{\partial \lambda} = \sum_{i=1}^{n}\left(\frac{x_i}{\hat{\lambda}} - 1\right) = 0,$$

and hence

$$\frac{1}{\hat{\lambda}}\sum_{i=1}^{n} x_i - n = 0,$$

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}\,.$$

The ML estimation of λ is the arithmetic mean of the sample values.

A sufficient condition for a maximum is fulfilled if:

$$\frac{\partial^2 \log L}{\partial \lambda^2} = -\frac{1}{\lambda^2}\sum_{i=1}^{n} x_i < 0\,,$$

Since $\lambda > 0$ and a Poisson distributed random sample cannot have negative realizations, the condition is satisfied.

ML Estimation of λ in an Exponential Distribution

Let $X_1,\ldots,X_n$ be a random sample (with replacement) of size n from an exponential distribution with unknown parameter $\lambda > 0$. Then for each $X_i, i = 1,\ldots,n$ we have:

$$f_{EX}(x_i|\lambda) = \begin{array}{ll} \lambda e^{-\lambda x_i} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{array}$$

The likelihood function for the realized sample $x_1,\ldots,x_n$ is given by:

$$L(\lambda|x_1,\ldots,x_n) = \prod_{i=1}^{n} \lambda e^{\lambda x_i} = \lambda^n \prod_{i=1}^{n} e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

and the corresponding log-likelihood function is:

$$\log L(\lambda|x_1,\ldots,x_n) = n\log\lambda - \lambda\sum_{i=1}^{n} x_i\,.$$

Setting the first derivative to zero yields:

$$\frac{\partial \log L(\lambda)}{\partial\lambda} = \frac{n}{\hat{\lambda}} - \sum_{i=1}^{n} x_i = 0$$

Solving for the ML estimator $\hat{\lambda}$ of λ one obtains:

$$\frac{n}{\hat{\lambda}} = \sum_{i=1}^{n} x_i$$

$$\hat{\lambda} = \frac{n}{\sum\limits_{i=1}^{n} x_i} = 1\ /\ \bar{x}.$$

Taking the second derivative with respect to λ one obtains:

$$\frac{\partial^2 \log L(\lambda)}{\partial\lambda^2} = -\frac{n}{\lambda^2}\,,$$

whereby the sufficient condition for a maximum is fulfilled since $n > 0$ and $\lambda > 0$.

Application of Least Squares

A random sample of size n is drawn from a population with unknown expectation $E(X) = \mu$. The $X_i, i = 1,\ldots,n$ are identically and independently distributed with $E(X_i) = \mu$ so that $g_i(\mu) = \mu$ for each i. The unknown parameter μ is estimated using least squares which minimizes the sum of squared deviations of the observations from the estimator $\hat{\mu}$. That is,

$$Q(\mu) = \sum_{i=1}^{n}(x_i - \mu)^2$$

is minimized. Differentiating and setting to zero yields:

$$\frac{\partial Q(\mu)}{\partial\mu} = -2\sum_{i=1}^{n}(x_i - \mu) = 0\,.$$

A little algebra yields the LS estimator:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

To verify that this is a minimum we check that the second derivative is positive at $\mu = \hat{\mu}$:

$$\frac{\partial^2 Q(\hat{\mu})}{\partial \mu^2} = 2n > 0\,.$$

If the data are drawn from a $N(\mu; \sigma^2)$ distribution, then the ML estimator of μ is also the sample mean. However, note that under ML, normality has been assumed while the LS estimator does not require such assumptions.

Explained: ML Estimation of an Exponential Distribution

While waiting for his flight, Mr. Businessman amuses himself by measuring the time (in minutes) between landings on a particular runway. He records the following observations: {3, 6, 6, 4, 8, 2, 4, 5, 9, 3}.

The random variable X which is the time interval between touch-downs is assumed to be exponentially distributed with unknown parameter $\lambda > 0$. He proceeds to estimate this parameter using ML. The likelihood function for the sample $(x_1, \ldots, x_{10})$ is given by

$$\begin{aligned} L(\lambda|3,6,6,4,8,2,4,5,9,3) =& \lambda e^{-3\lambda} \cdot \lambda e^{-6\lambda} \cdot \lambda e^{-6\lambda} \cdot \lambda e^{-4\lambda} \cdot \lambda e^{-8\lambda} \cdot \\ & \cdot \lambda e^{-2\lambda} \cdot \lambda e^{-4\lambda} \cdot \lambda e^{-5\lambda} \cdot \lambda e^{-9\lambda} \cdot \lambda e^{-3\lambda} \\ =& \lambda^{10} e^{-50\lambda} \end{aligned}$$

and the log-likelihood is given by:

$$\log L(\lambda) = 10 \log \lambda - 50\lambda\,.$$

Differentiating with respect to λ and setting to zero, one obtains:

$$\frac{\partial \log L(\lambda)}{\partial \lambda} = \frac{10}{\hat{\lambda}} - 50 = 0\,.$$

Solving this linear equation produces the ML estimator $\hat{\lambda}$ of λ:

$$\hat{\lambda} = \frac{10}{50} = 0.2 = 1/\bar{x}$$

Table 8.3 Number of car accidents for 50 days

No. of car accidents per day	No. of days
0	21
1	18
2	7
3	3
4	1

The second derivative (which is used to ensure that the result is a maximum) yields:

$$\frac{\partial^2 \log L(\lambda)}{\partial \lambda^2} = -\frac{10}{\lambda^2}$$

whereby the sufficient condition for a maximum is fulfilled.

Explained: ML Estimation of a Poisson Distribution

Data are collected on the number of car accidents for each of 50 days. Table 8.3 summarizes the observations.

The Poisson is used to model the number of occurrences of an event in a given time period. In this case, it is the number of accidents per day. A critical assumption is that the events are independent. Let X be the number of car accidents per day, then $X \sim PO(\lambda)$. The parameter λ is unknown and will be estimated using ML.

The likelihood function of the realized sample $x_1, \ldots, x_n$ is given by:

$$L(\lambda | x_1, \ldots, x_n) = \frac{\lambda^{x_1 + \cdots + x_{50}}}{x_1 ! \cdot \cdots \cdot x_{50} !} e^{-50\lambda} = \frac{\lambda^{45}}{0 ! \cdot 0 ! \cdot \cdots \cdot 3 ! \cdot 4 !} e^{-50\lambda} .$$

and the log-likelihood is given by:

$$\log L(\lambda | x_1, \ldots, x_{50}) = 45 \log \lambda - [\log(0 !) + \log(0 !) + \cdots + \log(3 !) + \log(4 !)] - 50\lambda .$$

Differentiating with respect to λ and setting to zero yields

$$\frac{\partial \log L}{\partial \lambda} = \frac{45}{\hat{\lambda}} - 50 = 0$$

and hence

$$\hat{\lambda} = \frac{45}{50} = 0.9 = \bar{x} .$$

We verify the second order (sufficient) condition for a maximum:

$$\frac{\partial^2 \log L}{\partial \lambda^2} = -\frac{1}{\lambda^2} 45 < 0\,.$$

8.4 Interval Estimation

Recall that an estimator $\hat{\theta}$ of a parameter θ is a random variable. Even if the estimator has desirable properties (such as consistency and efficiency), one cannot determine from the point estimate alone whether it is likely to be a good approximation to the true parameter. To provide information about the accuracy of the estimation process one usually applies interval estimation.

An interval estimator for an unknown parameter θ produces an interval such that:

- the probability that the estimation process results in an interval which contains the true parameter value θ, equals a given probability $1 - \alpha$.

Such an interval is called **confidence interval** and the corresponding probability the confidence level. The evaluation of the interval is based:

- on a random sample (with replacement) of size n, $X_1, \ldots, X_n$,
- the determination of two (random) values

$$V_L = g_L(X_1, \ldots, X_n) \;\; \textit{and} \;\; V_U = g_U(X_1, \ldots, X_n)\,,$$

for the lower and upper limits of the interval.

If these functions satisfy the condition

$$P(V_L \leq \theta \leq V_U) = 1 - \alpha\,,$$

then $[V_L; V_U]$ yields an interval for θ with confidence level $1 - \alpha$. That is, the probability that the interval contains the true value θ in repeated samples equals $1 - \alpha$. Commonly, α is chosen so that the confidence level $1 - \alpha$ is high (e.g., $0,90$, $0,95$ or $0,99$).

For a specific sample $x_1, \ldots, x_n$ we will denote a realized confidence interval with lowercase letters $[v_L; v_U]$.

It is essential to understand the interpretation of the confidence interval:

Before drawing a sample the limits of the confidence level are random variables. Since V_L and V_U are functions of the random variables $X_1, \ldots, X_n$, they are themselves random variables. Hence $[V_L; V_U]$ is a random interval, for which probability statements can be made.

$1 - \alpha$ is the probability that the *estimation procedure* produces intervals which contain the true value of the parameter θ. Put differently, if interval estimation

were repeated many times, then $(1-\alpha)\cdot 100\,\%$ of the intervals will contain θ and conversely $\alpha\cdot 100\,\%$ of the intervals will not contain θ.

Once the data have been drawn, the realizations $x_1,\ldots,x_n$ are substituted into V_L and V_U which leads to the realized confidence interval $[v_L; v_U]$. The limits v_L and v_U are fixed values. Either the unknown parameter θ lies within the estimation interval or it does not.

If one of the two limits is unrestricted, then one obtains one-sided confidence intervals:

- if $V_L = -\infty$ then one obtains an upper confidence interval $(-\infty; V_U]$, with $P(\theta \le V_U) = 1-\alpha$.
- if $V_U = +\infty$ then one obtains a lower confidence interval $[V_L; +\infty)$, with $P(V_L \le \theta) = 1-\alpha$.

For example, a one-sided upper confidence interval is of interest, if we want to have some assurance that the underlying parameter does not exceed a certain value.

Two-sided confidence intervals $[V_L; V_U]$ are used when both upper and lower bounds on the unknown parameter are of interest.

With two-sided confidence intervals the difference $V_U - V_L$ is referred to as the length (or width) of the interval. The length generally depends on the confidence level $1-\alpha$ and on the sample size n. Holding the sample size n constant, an increase in the confidence level $1-\alpha$ usually increases the length of the confidence interval. Hence, greater certainty that the unknown parameter θ will lie within the interval results in less precision about its position. On the other hand, an increase in sample size n while holding the confidence level $1-\alpha$ constant in general shortens the confidence interval.

There is a variety of ways of constructing a confidence interval or region which—in repeated samples—will contain the true parameter with probability $1-\alpha$. A convenient method for specifying a two-sided confidence interval is to require that each of the tails contain probability $\alpha/2$. That is the limits V_L, V_U are constructed so that:

$$P(\theta < V_L) = \alpha/2 \;\; \textit{and} \;\; P(V_U < \theta) = \alpha/2$$

Consequently

$$P(\theta < V_L) + P(V_U < \theta) = \alpha/2 + \alpha/2 = \alpha.$$

Our discussion of confidence intervals will focus on these equal tailed intervals. We will see that when the sampling distribution of the estimator is (approximately) symmetric, the resulting confidence intervals will be symmetric around the estimated value.

To determine the limits of the confidence interval the estimator $\hat{\theta}$ is typically used. Furthermore, its estimated standard error $\sigma(\hat{\theta})$ usually plays a direct role in

determining the width of the confidence interval. Indeed, confidence intervals are often of the form:

$$[V_L, V_U] = [\hat{\theta} - c \cdot \sigma(\hat{\theta}), \quad \hat{\theta} + c \cdot \sigma(\hat{\theta})]$$

where c is determined from the sampling distribution of $\hat{\theta}$ and depends on α. The corresponding confidence level is given by:

$$P(V_L \leq \theta \leq V_U) = P(\hat{\theta} - c \cdot \sigma(\hat{\theta}) \leq \theta \leq \hat{\theta} + c \cdot \sigma(\hat{\theta})) = 1 - \alpha\,.$$

8.5 Confidence Interval for the Mean

Assume a random variable X with unknown expectation $E(X) = \mu$. We wish to perform interval estimation for μ. Let $X_1, \ldots, X_n$ represent n i.i.d. draws from this population. It is known that the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is an unbiased, consistent and asymptotically efficient estimator of μ. The variance and standard deviation of $\bar{X}$ are (in the case of random sampling with replacement, see Chap. 7) given by:

$$Var(\bar{X}) = \sigma^2(\bar{X}) = \frac{\sigma^2}{n}$$

$$\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}\,.$$

For the construction of a two-sided confidence interval for μ:

- we start with the estimator $\bar{x}$,
- the standard deviation $\sigma(\bar{x})$ is used as the measure of accuracy
- a factor c will be required to multiply the standard deviation of $\bar{X}$ to achieve the given confidence level.

To construct the interval

$$[V_L; V_U] = [\bar{x} - c \cdot \sigma(\bar{X}); \bar{x} + c \cdot \sigma(\bar{X})]$$

we substitute $\sigma(\bar{X})$

$$[V_L; V_U] = \left[\bar{x} - c \cdot \frac{\sigma}{\sqrt{n}}; \bar{x} + c \cdot \frac{\sigma}{\sqrt{n}}\right]$$

and write the probability statement

$$P\left(\bar{x} - c \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + c \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha\,,$$

In order to determine c, we will need to make assumptions on the underlying distribution of the data which induces the distribution of $\bar{X}$. Below we will assume that the underlying data are normally distributed. Alternatively, in large samples, we know by the central limit theorem that $\bar{X}$ is approximately normally distributed. We also will need to distinguish between two cases: σ is known, and σ is unknown.

Confidence Interval for the Mean with Known Variance

Normally Distributed Population

Suppose X is normally distributed with $E(X) = \mu$ and $Var(X) = \sigma^2$:

$$X \sim N(\mu; \sigma^2)$$

For expositional purposes we assume σ^2 is known and the expectation μ is unknown. Suppose one draws a random sample of size n.

The random variables $X_1, \ldots, X_n$ are i.i.d. normally distributed with $E(X) = \mu$ and $Var(X) = \sigma^2$:

$$X_i \sim N(\mu; \sigma^2) \text{ for each } i\,.$$

From this it follows, that the estimator $\bar{X}$ is also normally distributed with $E(\bar{X}) = \mu$ and $Var(\bar{X}) = \sigma^2(\bar{X}) = \sigma^2/n$:

$$\bar{X} \sim N(\mu, \sigma^2(\bar{X}))\,.$$

The standardized random variable

$$Z = \frac{\bar{X} - \mu}{\sigma(\bar{X})} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \sqrt{n}\frac{\bar{X} - \mu}{\sigma}$$

is standard normal: $Z \sim N(0, 1)$.

Let $z_{\alpha/2}$ be the $\alpha/2$-quantile and $z_{1-\alpha/2}$ the $(1 - \alpha/2)$-quantile of the standard normal distribution. Then,

$$P(z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha.$$

The symmetry of the standard normal distribution implies that

$$z_{\alpha/2} = -z_{1-\alpha/2}$$

Hence,

$$P(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha.$$

For the probability $1 - \alpha/2$ the corresponding quantile $z_{1-\alpha/2}$ is found in standard normal tables.

After substitution for Z, we will isolate μ in the middle of the probability statement as follows:

$$\begin{aligned}
& P\left(-z_{1-\frac{\alpha}{2}} \leq Z \leq z_{1-\frac{\alpha}{2}}\right) \\
&= P\left(-z_{1-\frac{\alpha}{2}} \leq \sqrt{n}\frac{\bar{X}-\mu}{\sigma} \leq z_{1-\frac{\alpha}{2}}\right) \\
&= P\left(-z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \bar{X}-\mu \leq z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(\bar{X}-z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}+z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha\,.
\end{aligned}$$

The last probability statement yields the confidence interval for μ

$$\left[\bar{X}-z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\,,\quad \bar{X}+z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

The constant c which multiplies the standard deviation of $\bar{X}$ is given by $c = z_{1-\alpha/2}$. For a given sample, $x_1, \ldots, x_n$, substitution in the above expression yields a realization of the confidence interval.

Properties of the Confidence Interval

- The confidence interval constructed above assigns equal probabilities to each of the tails:

$$P\left(\mu < \bar{X}-z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = \frac{\alpha}{2}\,,\; P\left(\bar{X}+z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu\right) = \frac{\alpha}{2}\,.$$

- Symmetry of the distribution of Z results in a symmetric confidence interval around $\bar{X}$.

- The length of the confidence interval

$$\left(\bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) - \left(\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 2z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

  does not depend on the realized values $x_1, \ldots, x_n$. For a given σ, n, and $1 - \alpha$ one obtains different estimation intervals from sample to sample. However, these estimation intervals all have the same fixed length.
- The width of the confidence interval depends on the standard deviation σ of the population, the sample size n, and via $z_{1-\alpha/2}$ on the given confidence level. Increases in the standard deviation σ will ceteris paribus cause the interval to become wider. Increases in the confidence level $1 - \alpha$ will also cause the interval to become wider. Increases in the sample size n result in increased precision in estimation and hence in narrowing of the confidence interval.

  If the population distribution is unknown but the variance is known, then by the central limit theorem, $\bar{x}$ is approximately normally distributed given a sufficiently large normal sample size. In this case, the above confidence interval may be viewed as approximate.

Confidence Interval for the Mean with Unknown Variance

Normal Distribution in the Population

Suppose as before that

$$X \sim N(\mu; \sigma^2)\,,\ X_i \sim N(\mu; \sigma^2) \text{ for all } i \text{ and } \bar{X} \sim N(\mu; \sigma^2(\bar{X}))$$

We will again need a random variable which depends only on the unknown parameter μ. The standardized random variable Z will not work, because it requires us to know σ^2. Suppose the variance σ^2 is estimated using

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2\,.$$

and replace σ with the standard deviation S in the Z statistic to obtain

$$T = \sqrt{n} \frac{\bar{X} - \mu}{S}\,.$$

The random variable T follows a **t-distribution** with $n - 1$ degrees of freedom:

$$T \sim t(n-1)$$

Let $t_{n-1;\alpha/2}$ be the $\alpha/2$-quantile and $t_{n-1;1-\alpha/2}$ be the $(1-\alpha/2)$-quantile of the t-distribution. Due to the symmetry of the t-distribution

$$t_{n-1;\alpha/2} = -t_{n-1;1-\alpha/2}\,,$$

Hence it follows

$$P(-t_{n-1;1-\alpha/2} \leq T \leq t_{n-1;1-\alpha/2}) = 1-\alpha\,.$$

For the probability $1-\alpha/2$ one then obtains $t_{n-1;1-\alpha/2}$ from the table of the t-distribution.

Substituting T and after some algebraic manipulation we have a confidence interval.

$$\left[\bar{X} - t_{n-1;1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\,,\;\; \bar{X} + t_{n-1;1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\right]$$

for the unknown parameter μ with corresponding probability statement and confidence level given by:

$$P\left(\bar{X} - t_{n-1;1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1;1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\right) = 1-\alpha\,.$$

Since the t-distribution converges to $N(0;1)$ as sample size n increases, the standard normal may be used instead of the t-distribution if the sample size is sufficiently large. As a rule of thumb, this is the case for $n > 30$.

Properties of the Confidence Interval

- The properties are similar to those in the previous case except that the length of the confidence interval is no longer fixed but a random variable since it depends on the estimate s of σ:

  $$2t_{n-1;1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

  With given sample size n and confidence level $1-\alpha$ one obtains different estimated intervals from sample to sample, which also may display different lengths.
- As before, the length of the confidence interval depends on the sample size n and via $t_{n-1;1-\alpha/2}$ on the given confidence level $1-\alpha$.
- As the quantiles $t_{n-1;1-\alpha/2}$ from the t-distribution are larger than the quantiles $z_{1-\alpha/2}$ from the standard normal the confidence intervals are wider when the variance is unknown. This additional absence of knowledge about the underlying data generating mechanism is "embedded" in the t-distribution.

If the population distribution is unknown, then again using the central limit theorem, $\bar{X}$ is approximately normally distributed, and the above procedure yields approximate confidence intervals.

Explained: Confidence Intervals for the Average Household Net Income

For a population of $N = 2000$ households let X be a random variable representing a household's net income. Expected household net-income, i.e., $E(X) = \mu$, is unknown and must be estimated. We are interested in a point estimate and a confidence interval with a confidence level of $1 - \alpha = 0.95$.

To estimate μ we use the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \,.$$

The specification of the confidence interval is determined by the information that is available about the population.

Normally Distributed Population

1.) Confidence interval for μ with known standard deviation σ

Assume the random variable X (household net-income) to be normally distributed with standard deviation $\sigma = 1012.8$. Using this information we may calculate a two-sided confidence interval:

$$\left[\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \,,\;\; \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

for μ which has confidence level

$$P\left(\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha$$

To the given level of $1 - \alpha = 0.95$ one obtains $z_{1-\alpha/2} = z_{0.975} = 1.96$. Substituting σ and $z_{1-\alpha/2}$ yields

$$P\left(\bar{x} - 1.96 \frac{1012.8}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{1012.8}{\sqrt{n}} \right) = 0.95$$

Table 8.4 Observations on household net income, sample size $n = 20$ (data have been reordered)

i	Households net-income x_i	i	Households net-income x_i
1	800	11	2500
2	1200	12	2500
3	1400	13	2500
4	1500	14	2700
5	1500	15	2850
6	1500	16	3300
7	1800	17	3650
8	1800	18	3700
9	2300	19	4100
10	2400	20	4300

and

$$\left[\bar{x} - 1.96\frac{1012.8}{\sqrt{n}}, \quad \bar{x} + 1.96\frac{1012.8}{\sqrt{n}}\right]$$

A random sample (with replacement) of size $n = 20$ households from the above population yielded the results provided in Table 8.4

Mean household net income is $\bar{x} = 48,300/20 = 2,415$.

The estimated confidence interval is given by:

$$\left[2415 - 1.96\frac{1012.8}{\sqrt{20}}, \quad 2415 + 1.96\frac{1012.8}{\sqrt{20}}\right] = [2415 - 443.88\,, \quad 2415 + 443.88]$$
$$= [1971.12, \quad 2858.88]$$

To illustrate some issues related to confidence intervals, 24 further samples of size $n = 20$ are drawn. Mean households net income $\bar{x}$ and the corresponding confidence interval are computed for each sample. They are given in Table 8.5.

Figure 8.4 shows the 25 point estimates and confidence intervals. The true mean μ of the population is depicted as a dotted line. Note the following points.

- The limits V_L and V_U of a confidence interval are random variables and as such differ from sample to sample.
- Of the 25 intervals, 23 intervals (92 %) contain the true value μ and 2 intervals (samples no. 9 and no. 24) do not. Does this contradict the fixed confidence level 0.95?

  The answer is NO, since the confidence level refers to a very large number of samples (much larger than 25).
- All 25 intervals have the same width 887.76, since the standard deviation σ of the population has been assumed to be known.

Table 8.5 Mean household net income and confidence interval for 25 random samples of size $n = 20$

i	$\bar{x}$	v_L	v_U	i	$\bar{x}$	v_L	v_U
1	2413.40	1969.52	2857.28	14	2126.50	1682.62	2570.38
2	2317.00	1873.12	2760.88	15	2243.15	1799.27	2687.03
3	2567.50	2123.62	3011.38	16	2361.25	1917.37	2805.13
4	2060.90	1617.02	2504.78	17	2607.25	2163.37	3051.13
5	2363.50	1919.62	2807.38	18	2319.55	1875.67	2763.43
6	2774.30	2330.42	3218.18	19	2203.85	1759.97	2647.73
7	2298.80	1854.92	2742.68	20	2395.25	1951.37	2839.13
8	72241.15	1797.27	2685.03	21	2659.00	2215.12	3102.88
9	1915.30	1471.42	2359.18	22	2168.50	1724.62	2612.38
10	2062.15	1618.27	2506.03	23	2110.30	1666.42	2554.18
11	2267.75	1823.87	2711.63	24	1884.90	1441.02	2328.78
12	2163.10	1719.22	2606.98	25	2415.00	1971.12	2858.88
13	2635.00	2191.12	3078.88				

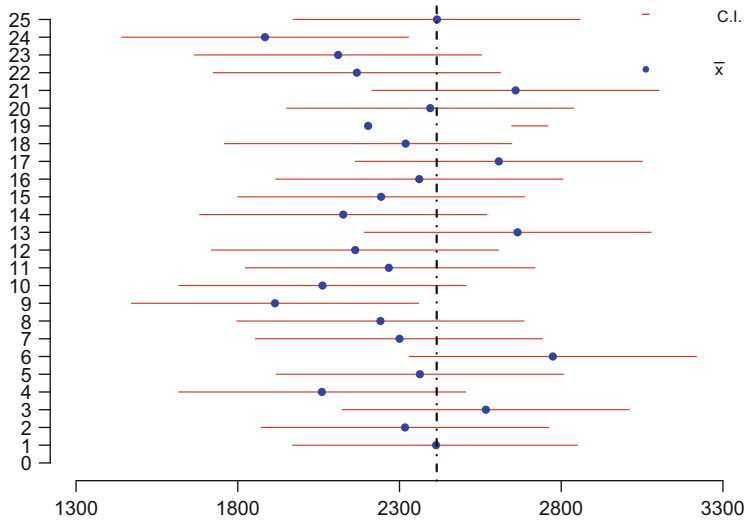


Fig. 8.4 Point estimates and confidence intervals for 25 random samples of size $n = 20$

2.) Confidence Interval for μ with unknown Standard Deviation σ

Again assume a normally distributed random variable X (household net-income) where the standard deviation is unknown : $X \sim N(\mu; \sigma^2)$. We will draw random samples of size $n = 20$. To determine the confidence interval for μ the variance σ^2

is estimated using s^2. The confidence interval is given by

$$\left[\bar{x} - t_{n-1;1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\,,\;\; \bar{x} + t_{n-1;1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right]$$

with confidence level

$$P\left(\bar{X} - t_{n-1;1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1;1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

For the given confidence level $1 - \alpha = 0.95$ tables for the t-distribution yield $t_{n-1;1-\alpha/2} = t_{19;0.975} = 2.093$. Substituting in the above, one obtains

$$\left[\bar{x} - 2.093 \frac{s}{\sqrt{n}}\,,\;\; \bar{x} + 2.093 \frac{s}{\sqrt{n}}\right]$$

Referring to the data provided in Table 8.1, we calculate the mean and standard deviation to be $\bar{x} = 48\,300/20 = 2\,415$, $s = 1001.065$ and the confidence interval to be

$$\begin{aligned}
&\left[2415 - 2,093 \frac{1001.065}{\sqrt{20}},\;\; 2415 + 2,093 \frac{1001.065}{\sqrt{20}}\right] \\
&= [2415 - 468.51\,,\;\; 2415 + 468.51] \\
&= [1946.49,\;\; 2883.51]\,.
\end{aligned}$$

Table 8.6 contains mean household net-income $\bar{x}$, the standard deviation s and the confidence interval as well as the width of the confidence interval for the 25 samples.

Figure 8.5 shows the 25 point estimates and confidence intervals. For illustrative purposes the true mean μ of the population is depicted as a dashed line.

In this case only one interval does not cover the true value of the parameter μ (sample no. 24). From Table 8.6 and Fig. 8.5 it is evident that the lengths of the intervals vary from sample to sample and are hence random variables. The cause is the unknown standard deviation σ of the population, which has to be estimated for each sample.

Unknown Population Distribution and Unknown Standard Deviation

The case most frequently occurring in practice is now considered. In this case the distribution of the random variable X and the standard deviation σ are unknown. In order to use the procedures we have proposed, it is necessary that the sample size n be sufficiently large, so that the central limit theorem can be applied. We select $n = 100$.

Table 8.6 Mean household net income $\bar{x}$, standard deviation s, confidence interval and interval width for 25 samples of size $n = 20$

i	$\bar{x}$	s	v_L	v_U	*Width*
1	2413.40	1032.150	1930.34	2896.46	966.12
2	2317.00	872.325	1908.74	2825.26	816.52
3	2567.50	1002.008	2098.55	3036.45	937.90
4	2060.90	812.365	1680.71	2441.09	760.38
5	2363.50	1376.648	1719.22	3007.78	1288.56
6	2774.30	1213.779	2206.24	3342.63	1136.12
7	2298.80	843.736	1903.92	2693.68	789.76
8	2241.15	1116.827	1718.46	2763.84	1045.38
9	1915.30	1113.122	1394.35	2436.25	1041.90
10	2062.15	856.069	1661.50	2462.80	801.30
11	2267.75	1065.227	1769.21	2766.29	997.08
12	2163.10	1040.966	1675.92	2650.28	974.36
13	2635.00	1154.294	2094.78	3175.22	1080.44
14	2126.50	1103.508	1610.05	2642.95	1032.90
15	2243.15	1126.913	1715.74	2770.56	1054.82
16	2361.25	1166.260	1815.43	2907.07	1091.64
17	2607.25	848.019	2210.37	3004.13	793.76
18	2319.55	941.236	1879.04	2760.06	881.02
19	2203.85	974.980	1747.55	2660.15	912.60
20	2395.25	899.461	1974.29	2816.21	841.92
21	2659.00	969.720	2205.16	3112.84	907.68
22	2168.50	763.222	1811.31	2525.69	714.38
23	2110.30	1127.608	1582.57	2638.03	1055.46
24	1884.90	928.420	1450.39	2319.41	869.02
25	2415.00	1001.065	1946.49	2883.51	937.02

Then,

$$\left[\bar{x} - z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\,,\ \ \bar{x} + z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right]$$

is a confidence interval for the unknown parameter μ at the approximative confidence level

$$P\left(\bar{X} - z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right) \approx 1 - \alpha$$

Again, if $1 - \alpha = 0.95$ tables for the standard normal distribution yield $z_{1-\alpha/2} = z_{0.975} = 1.96$.

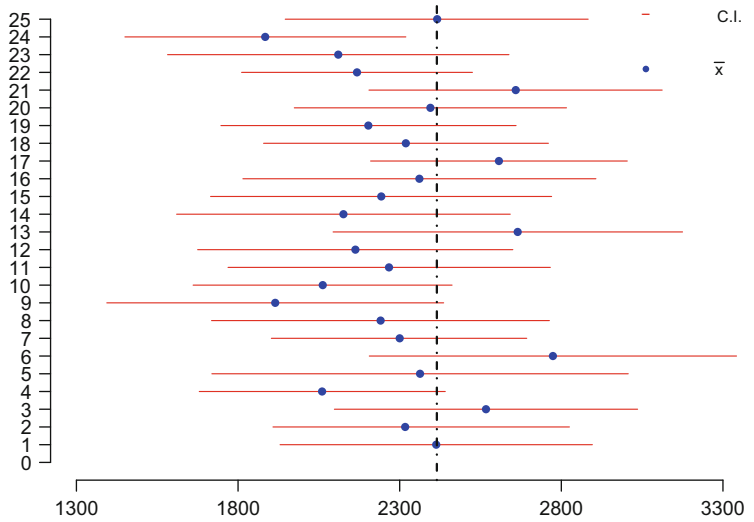


Fig. 8.5 Estimation intervals for 25 random samples of size $n = 20$

Figure 8.6 depicts the point estimates and confidence intervals for 50 random samples (with replacement). For illustrative purposes the true mean μ of the population is depicted using a dashed line. Numerical results are not provided.

We again observe that the width of the intervals varies from sample to sample and are hence random variables. This is due to the unknown standard deviation of the population. Of the 50 estimation intervals 2 (4 %) do not cover the true parameter value μ.

Enhanced: Confidence Intervals for the Lifetime of a Bulb

The marketing department of a lamp manufacturer needs values for the average lifetime of a particular type of bulb.

- It is of course not possible to sample the entire population since it consists of bulbs that are yet to be produced. Furthermore, in determining its lifetime, the bulb is destroyed. Hence, a sample from the population needs to be drawn.
- To ensure that the sample is representative, a random sample is drawn.
- Drawing a random sample with replacement is not feasible since once its lifetime is measured, the bulb is destroyed. Since the total production is large, however, the fact that one is sampling without replacement does not significantly effect the distribution from which one is sampling.
- We are interested in a point estimate for the unknown average life time μ as well as a symmetric confidence interval with confidence level $1 - \alpha = 0.95$.

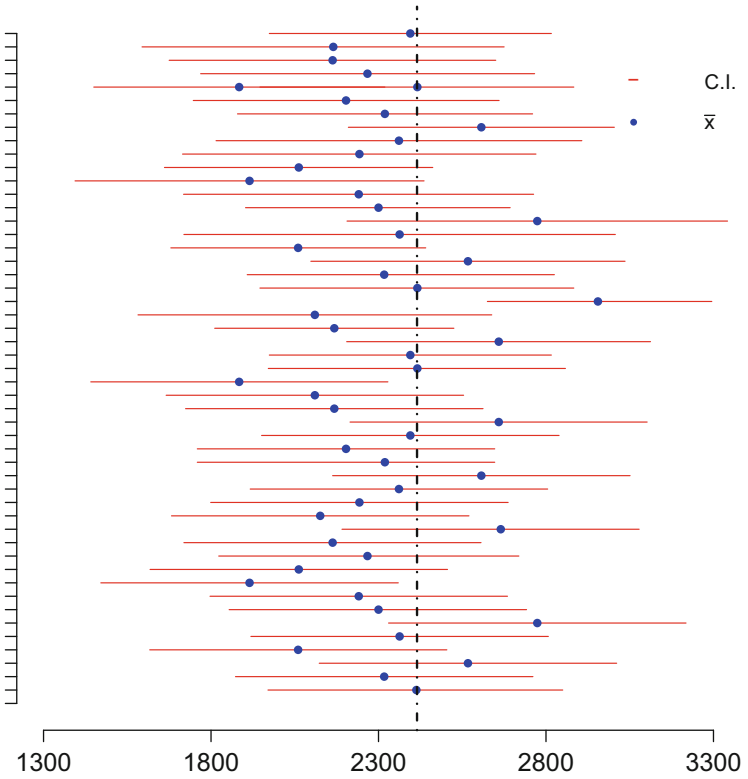


Fig. 8.6 Confidence intervals of 50 random samples of size $n = 100$

- Neither the variance σ^2 nor the distribution of the random variable $X = \{\text{life time}\}$ is known. We assume the sample size n is large enough so that we can use the approximate confidence interval

$$\left[\bar{X} - z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\,,\ \bar{X} + z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right]$$

  with corresponding approximate confidence level

$$P\left(\bar{X} - z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \le \mu \le \bar{X} + z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right) \approx 1 - \alpha$$

  At the given confidence level $1 - \alpha = 0.95$ one may consult a table for the standard normal distribution to obtain $z_{1-\alpha/2} = z_{0.975} = 1.96$.
- There is a trade-off between the accuracy of the approximation and the cost of sampling. We select $n = 50$.

The specific sample yields the following point estimates:

Mean life time in sample $\bar{x}$:	1600	hours
Variance s^2 in sample:	8100	hours^2
Standard deviation s in the sample:	90	hours

The confidence interval is given by:

$$\left[1600 - 1.96\frac{90}{\sqrt{50}}\,,\;\; 1600 + 1.96\frac{90}{\sqrt{50}}\right] = [1600 - 24.95\,,\;\; 1600 + 24.95]$$
$$= [1575.05,\;\; 1624.95]$$

A high confidence level (in this case 0.95) is selected to ensure that the resulting confidence intervals contain the true parameter value with high probability.

From the point of view of marketing and quality control, it is important that the advertised lifetime is met or exceeded with high probability. Thus, one is interested in a one-sided confidence interval of the form

$$P\left(\bar{X} - z_{1-\alpha}\frac{S}{\sqrt{n}} \le \mu\right) = 1 - \alpha = 0.95$$

where one obtains $z_{1-\alpha} = z_{0.95} = 1.645$ from tables of the standard normal. Using the previous data one obtains:

$$v_L = 1600 - 1.645 \cdot \frac{90}{\sqrt{50}} = 1600 - 20.94 = 1579.06 \;\text{ h}$$

and the corresponding one-sided estimation interval

$$[1579.06\,;\, +\infty)\,.$$

Interactive: Confidence Intervals for the Mean

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please select

- the confidence level $1 - \alpha$
- the sample size n
- if you expect the true population variance to be known

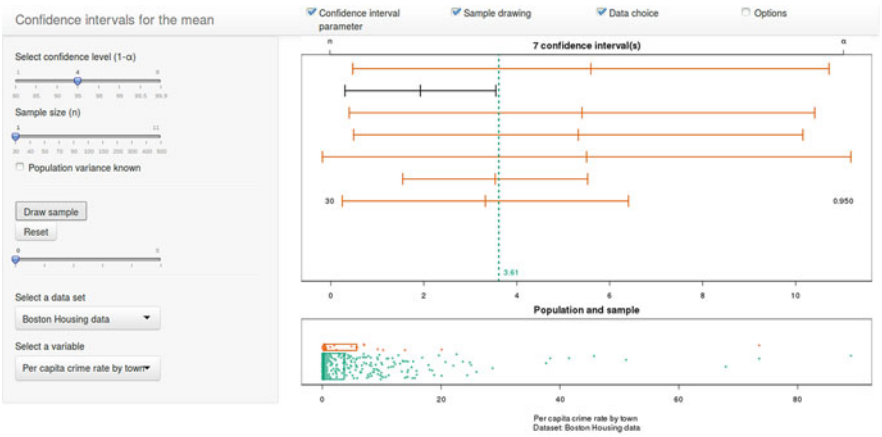


Fig. 8.7 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_ci1

Use

- "Draw sample" to manually construct a confidence interval
- "Reset" to reset the graphic
- the slider to cause an automated drawing of samples

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

Output

The upper graphic in Fig. 8.7 displays the resulting confidence intervals. The dashed line represents the population mean. After drawing an appropriate amount of samples we may see that $1 - \alpha$ of the observed intervals capture the population mean.

The graphic below is a scatterplot including the population (green) and sample (orange).

8.6 Confidence Interval for Proportion

Suppose we are drawing from a dichotomous population, where π denotes the proportion of elements with a given property. We want to estimate a confidence interval for the unknown parameter π.

We draw a random sample of size n in such a manner that $X_1, \ldots, X_n$ are independently and identically Bernoulli distributed (see Sect. *Binomial Distribution*).

The sample proportion is the number of "successes" in the sample divided by the sample size n, that is, the mean of the Bernoulli variables $X_1, \ldots, X_n$. It is worth emphasizing that the sample proportion is a sample mean and as such inherits the properties and behavior of a sample mean. Thus,

$$\hat{\pi} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

with expectation and variance

$$E(\hat{\pi}) = \pi \,, \quad Var(\hat{\pi}) = \frac{\pi(1-\pi)}{n}$$

It is an unbiased and consistent estimator of π (see Sect. *Properties of Estimators*).

Since it is quite difficult to construct confidence intervals for small samples, we will restrict ourselves to the case where the sample size n is sufficiently large so that we may use the Central Limit Theorem to obtain the distribution of the estimator. In particular,

$$Z = \frac{\hat{\pi} - \pi}{\sigma(\hat{\pi})} = \frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

is approximately normal: $Z \sim N(0; 1)$. Hence, we conclude that

$$P\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\hat{\pi} - \pi}{\sigma(\hat{\pi})} \leq z_{1-\frac{\alpha}{2}}\right) \approx 1 - \alpha \,,$$

where $z_{1-\alpha/2}$ is obtained from standard normal tables. Still we cannot construct a confidence interval for π, since the variance of $\hat{\pi}$ depends on π which is unknown. We simply replace $\sigma(\hat{\pi})$ with a consistent estimate

$$\hat{\sigma}(\hat{\pi}) = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

This is a consistent estimator of $\sqrt{\frac{\pi(1-\pi)}{n}}$ since $\hat{\pi}$ is a consistent estimator of π. The above probability statement becomes

$$P\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\hat{\pi} - \pi}{\hat{\sigma}(\hat{\pi})} \leq z_{1-\frac{\alpha}{2}}\right) \approx 1 - \alpha$$

We isolate π in the middle of the probability statement to obtain:

$$P\left(\hat{\pi} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \;\leq\; \pi \;\leq\; \hat{\pi} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right) \approx 1-\alpha\,.$$

Hence for large sample sizes an approximate confidence interval is given by:

$$\left[\hat{\pi} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}},\quad \hat{\pi} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

The normal distribution provides a reasonable approximation so long as π is not too close to zero or one. Typically, sample size should be no smaller than 30, and preferably substantially larger, e.g., $n \geq 100$.

Properties of Confidence Intervals

- The two-sided confidence intervals we have constructed assign roughly equal probabilities to the tails:

$$P\left(\pi < \hat{\pi} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right) \approx \frac{\alpha}{2}\,,$$

$$P\left(\hat{\pi} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} < \pi\right) \approx \frac{\alpha}{2}\,.$$

- By construction the confidence interval is symmetric around the point estimate $\hat{\pi}$.
- The length of the interval

$$2z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

  is a random variable, since it depends through $\hat{\pi}$ on the random sample.
- The length of the confidence interval also depends on the confidence level $1-\alpha$ and on n.

Explained: Confidence Intervals for the Percentage of Votes

The leader of a political party 'F' is interested in knowing what fraction of citizens would vote for it if an election were held. A survey of $n = 2000$ citizens is performed which asks the question:

> If there were an election tomorrow which party would receive your vote?

According to the survey 103 citizens declared that they would vote for "F." We wish to estimate a 95 % confidence interval for π, the proportion of voters who would vote for "F."

Note the following:

- In order to insure that a citizen that has been already asked is not sampled a second time, we sample without replacement (though the probability of replication is low given the sample size).
- Since interest is focused on party "F," the event A is defined as "the individual votes for F" and the complementary event $\bar{A}$ as "the individual does not vote for F." Thus for our purposes the population is dichotomous. The proportion of votes for party F is $\pi = P(A)$.
- The sample size is sufficiently large ($n = 2000$), so that one may construct an approximate confidence interval using the normal approximation:

$$\left[\hat{\pi} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}, \quad \hat{\pi} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

  which has an approximate confidence level of 95 %. We obtain $z_{0.975} = 1.96$

The results of the survey yield $\hat{\pi} = 103/2000 = 0.0515$ and a 95 % confidence interval:

$$\left[0.0515 - 1.96 \cdot \sqrt{\frac{0.0515 \cdot 0.9485}{2000}}, \quad 0.0515 + 1.96 \cdot \sqrt{\frac{0.0515 \cdot 0.9485}{2000}}\right]$$

$$= [0.0418\,;\, 0.0612]\,.$$

Interactive: Confidence Intervals for the Proportion

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please select

- the confidence level $1 - \alpha$
- the sample size n

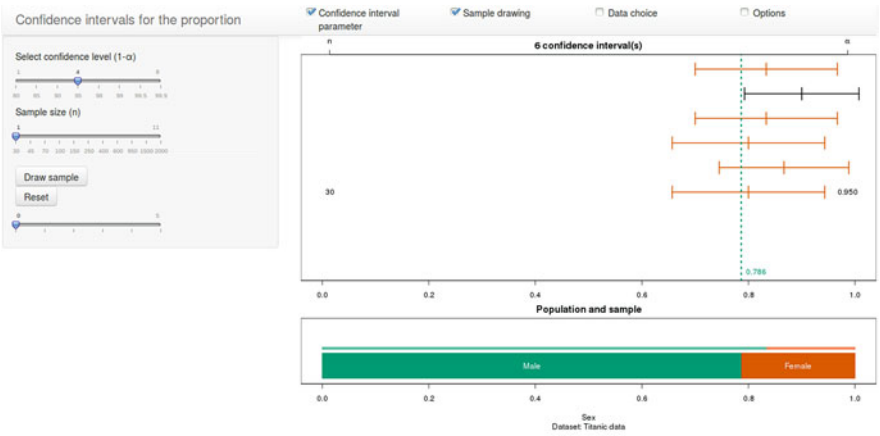


Fig. 8.8 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_cipi

Use

- "Draw sample" to manually construct a confidence interval
- "Reset" to reset the graphic
- the slider to cause an automated drawing of samples

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

Output

The upper graphic in Fig. 8.8 displays the resulting confidence intervals. The dashed line represents the population proportion. After drawing an appropriate amount of samples we may see that $1 - \alpha$ of the observed intervals capture the population proportion.

The graphic below compares the proportion of the population (thick bar) and sample (thin bar).

8.7 Confidence Interval for the Variance

We want to derive a confidence interval for the unknown variance σ^2 of a population under the following assumption:

1. The population is normally distributed: $X \sim N(\mu; \sigma^2)$.
2. The expectation $E(X) = \mu$ is unknown.

3. A random sample of size n is drawn, the random variables $X_1, \ldots, X_n$ are independently and identically normally distributed.

As we have seen above, an unbiased estimator of the unknown variance σ^2 is given by:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 .$$

It has been shown (see Sect. *Distribution of the Sample Variance*) that

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2$$

follows a chi-square distribution with $n-1$ degrees of freedom. We can now make probability statements of the following form:

$$P\left(\frac{\sigma^2 \chi^2_{\frac{\alpha}{2};n-1}}{n-1} \leq S^2 \leq \frac{\sigma^2 \chi^2_{1-\frac{\alpha}{2};n-1}}{n-1} \right) = 1 - \alpha$$

Here, $\chi^2_{\frac{\alpha}{2};n-1}$ is the $\alpha/2$-quantile and $\chi^2_{1-\frac{\alpha}{2};n-1}$ the $(1-\alpha/2)$-quantile of the chi-square distribution with $n-1$ degrees of freedom. By algebraic manipulation we may isolate σ^2 in the middle of the probability statement:

$$P\left(\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2};n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2};n-1}} \right) = 1 - \alpha .$$

The corresponding estimate of the confidence interval is given by

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2};n-1}} , \quad \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2};n-1}} \right] .$$

The interpretation is the same as before: a proportion $1-\alpha$ of confidence intervals constructed in this fashion will contain the true parameter value σ^2.

Properties of the Confidence Interval

- By construction, these confidence intervals assign equal probability mass to the tails:

$$P\left(\sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2};n-1}} \right) = \frac{\alpha}{2} , \quad P\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2};n-1}} < \sigma^2 \right) = \frac{\alpha}{2} .$$

Table 8.7 Realizations of random samples of size $n = 20$ (ordered by size)

i	Household net income (EUR) x_i	i	Household net income (EUR) x_i
1	800	11	2500
2	1200	12	2500
3	1400	13	2500
4	1500	14	2700
5	1500	15	2850
6	1500	16	3300
7	1800	17	3650
8	1800	18	3700
9	2300	19	4100
10	2400	20	4300

- The confidence interval is not symmetric around the point estimate s^2, since the chi-square distribution is not a symmetric distribution.
- The length of the confidence interval

$$(n-1)S^2\left(\frac{1}{\chi^2_{\frac{\alpha}{2};n-1}}-\frac{1}{\chi^2_{1-\frac{\alpha}{2};n-1}}\right)$$

depends on the sampled values $x_1, \ldots, x_n$ and is a random variable. The length of the interval also depends on the sample size n and on the confidence level $1-\alpha$.

Explained: Confidence Intervals for the Variance of Household Net Income

For a population of $N = 2000$ households let X denote net household income. We assume that X is approximately normally distributed $X \sim N(\mu; \sigma^2)$; the two parameters μ and the variance σ^2 are unknown.

Construction of confidence intervals for the unknown mean μ has been studied in the section *Confidence Intervals for the Expectation*.

Here, we want to focus on the unknown variance σ^2, for which we will construct a confidence interval with confidence level $1-\alpha = 0.95$.

Mean household income of the sample is (Table 8.7)

$$\bar{x} = 48\,300/20 = 2\,415\,.$$

Our point estimate for the unknown variance σ^2 is given by

$$s^2 = 1\,002\,131.58$$

Using chi-square tables we find

$$\chi^2_{\alpha/2;n-1} = \chi^2_{0.025;19} = 8.91 \quad \text{and} \quad \chi^2_{1-\alpha/2;n-1} = \chi^2_{0.975;19} = 32.85$$

Hence the confidence interval is given by

$$\left[\frac{19 \cdot 1002131.58}{32.85}, \quad \frac{19 \cdot 1002131.58}{8.91}\right] = [579619.48, \quad 2136980.92]\,.$$

Interactive: Confidence Intervals for the Variance

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please select

- the confidence level $1 - \alpha$
- the sample size n

Use

- "Draw sample" to manually construct a confidence interval
- "Reset" to reset the graphic
- the slider to cause an automated drawing of samples

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

Output

The upper graphic in Fig. 8.9 displays the resulting confidence intervals. The dashed line represents the population variance. After drawing an appropriate amount of samples we may see that $1 - \alpha$ of the observed intervals capture the population variance.

The graphic below compares the variance of the population (green) and sample (orange).

8.8 Confidence Interval for the Difference of Two Means

There are various ways to construct a confidence interval for the difference of two means $\mu_1 - \mu_2$ depending on the assumptions one makes. Our assumptions are as follows:

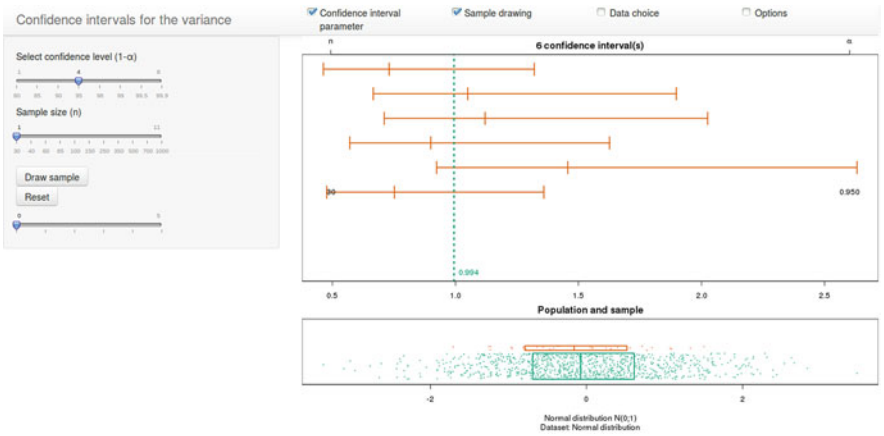


Fig. 8.9 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_cisig

- In the two populations the random variables X_1 and X_2 are normally distributed with parameters $E(X_1) = \mu_1$, $E(X_2) = \mu_2$, $Var(X_1) = \sigma_1^2$ and $Var(X_2) = \sigma_2^2$, i.e., $X_1 \sim N(\mu_1; \sigma_1^2)$ and $X_2 \sim N(\mu_2; \sigma_2^2)$.
- From each population a random sample is drawn (with replacement), The sample sizes are denoted by n_1 and n_2, respectively.
- The random samples are independent of each other.

When constructing confidence intervals for the difference $\mu_1 - \mu_2$ of two means one is often interested in seeing whether the value 0 is covered by the interval. If $\mu_1 - \mu_2 = 0$ is not an element of the interval, then the two populations are different at least with respect to their means.

Since X_1 and X_2 are normally distributed, $\bar{x_1}$ and $\bar{x_2}$ are also normal (see Sect. *Distribution of the Sample Mean*). Moreover we have:

$$E(\bar{X_1}) = \mu_1 \quad | \quad Var(\bar{X_1}) = \sigma^2(\bar{X_1}) = \frac{\sigma_1^2}{n_1},$$

$$E(\bar{X_2}) = \mu_2 \quad | \quad Var(\bar{X_2}) = \sigma^2(\bar{X_2}) = \frac{\sigma_2^2}{n_2}.$$

In summary

$$\bar{X_1} \sim N\left(\mu_1; \frac{\sigma_1^2}{n_1}\right) \qquad \bar{X_2} \sim N\left(\mu_2; \frac{\sigma_2^2}{n_2}\right)$$

Since linear combinations of independent normally distributed random variables are also normally distributed, we also have that the difference of the two sample means

$$D = \bar{X_1} - \bar{X_2}$$

is normally distributed with expectation

$$E(D) = E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

and variance

$$Var(D) = \sigma_D^2 = Var(\bar{X}_1 - \bar{X}_2) = Var(\bar{X}_1) + Var(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\,.$$

The standardized random variable

$$Z = \frac{D - E(D)}{\sigma_D} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

is therefore $N(0;1)$.

We distinguish between two cases:

- the variances of the two populations σ_1^2 and σ_2^2 are known
- the variances of the two populations σ_1^2 and σ_2^2 are unknown

1. Case: The Variances σ_1^2 and σ_2^2 of the Two Populations Are Known

If both variances σ_1^2 and σ_2^2 known, we have the confidence interval

$$\left[(\bar{X}_1 - \bar{X}_2) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\quad (\bar{X}_1 - \bar{X}_2) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

for the difference $\mu_1 - \mu_2$ at confidence level $1-\alpha$; i.e.,

$$P\left((\bar{X}_1 - \bar{X}_2) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le (\bar{X}_1 - \bar{X}_2) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1 - \alpha\,.$$

Properties of the Confidence Interval

- By construction these confidence intervals assign equal probability mass to the tails:

$$P\left(\mu_1 - \mu_2 < D - z_{1-\frac{\alpha}{2}}\sigma_D\right) = \frac{\alpha}{2}\,,\quad P\left(D + z_{1-\frac{\alpha}{2}}\sigma_D < \mu_1 - \mu_2\right) = \frac{\alpha}{2}\,.$$

- The confidence interval is symmetric around the estimated difference D,
- The length of the interval is constant given n_1 and n_2, the variances σ_1^2 and σ_2^2 and the confidence level $1-\alpha$.

Hint: If we cannot assume the populations to be normally distributed, but the two sample sizes $n_1 \geq 30$ and $n_2 \geq 30$, the Central Limit Theorem may be used to justify the same confidence interval procedure. In this case, the confidence level is approximately $1-\alpha$.

2. Case: The Variances σ_1^2 and σ_2^2 of the Two Populations Are Unknown

In this case σ_1^2 and σ_2^2 are estimated using the unbiased and consistent estimators

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2 \qquad S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(X_{2i} - \bar{X}_2)^2.$$

If we can assume ***variance homogeneity***, i.e., $\sigma_1^2 = \sigma_2^2$, one may produce an estimate s^2 for the joint variance σ^2. This is the weighted arithmetic mean of the two sample variances:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

s^2 is also called a pooled variance. The estimator s_D^2 for σ_D^2 is hence:

$$s_D^2 = s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = \frac{n_1 + n_2}{n_1 n_2}\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

The standard deviation s_D—the square root of s_D^2—is used to standardize. The resulting random variable

$$T = \frac{D - E(D)}{S_D} = \frac{(\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{n_1+n_2}{n_1 n_2}\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}}$$

is t-distributed with $n_1 + n_2 - 2$ degrees of freedom. We may now construct a confidence interval for the difference $\mu_1 - \mu_2$:

$$\left[(\bar{x}_1 - \bar{x}_2) - t_{n_1+n_2-2;1-\frac{\alpha}{2}}s_D\,,\quad (\bar{x}_1 - \bar{x}_2) + t_{n_1+n_2-2;1-\frac{\alpha}{2}}s_D\right]$$

at a confidence level $(1-\alpha)$:

$$P\left((\bar{X}_1-\bar{X}_2)-t_{n_1+n_2-2;1-\frac{\alpha}{2}}S_D \;\leq\; \mu_1-\mu_2 \;\leq\; (\bar{X}_1-\bar{X}_2)+t_{n_1+n_2-2;1-\frac{\alpha}{2}}S_D\right) \approx 1-\alpha\,.$$

If one has ***variance heterogeneity***, i.e., $\sigma_1^2 \neq \sigma_2^2$, we use the estimator S_D^2 of σ_D^2 given by:

$$S_D^2 = \frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}$$

If the two sample sizes are sufficiently large ($n_1 > 30$ and $n_2 > 30$), then we may use

$$\left[(\bar{x}_1-\bar{x}_2)-z_{1-\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}},\quad (\bar{x}_1-\bar{x}_2)+z_{1-\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right]$$

for the confidence interval at level $(1-\alpha)$, i.e.,

$$P\left((\bar{X}_1-\bar{X}_2)-z_{1-\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} \;\leq\; \mu_1-\mu_2 \;\leq\; (\bar{X}_1-\bar{X}_2)+z_{1-\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right) = 1-\alpha\,.$$

Properties of Confidence Intervals When Variances Are Unknown

- By construction these confidence intervals assign equal probability mass to the tails.
- The confidence interval is symmetric around the point estimate $\bar{x}_1-\bar{x}_2$.
- The length of the confidence interval is random since it depends on s_1^2 and s_2^2.
- The confidence interval also depends on the sample sizes n_1 and n_2 and on the confidence level $1-\alpha$.

Explained: Confidence Interval for the Difference of Car Gas Consumptions

An automobile club wants to compare (highway) gas consumption of two similar types of cars produced by company A and B. To assist the club, we will construct a confidence interval for the difference of the two means $\mu_1 - \mu_2$ at a confidence level $1 - \alpha = 0.95$. We make the following assumptions:

- It is assumed that the random variables:

  $X_1 =$ "gas consumption per 100 km of A type cars," and

  $X_2 =$ "gas consumption per 100 km of B type cars"

  are normally distributed with unknown means $E(X_1) = \mu_1$ and $E(X_2) = \mu_2$ and unknown variances $Var(X_1) = \sigma_1^2$ and $Var(X_2) = \sigma_2^2$. We do not assume that variances are equal.
- We assume $n_1 \geq 30$ and $n_2 \geq 30$.
- The populations are large so we will perform sampling with replacement.
- We will assume all observations are independent.

The confidence interval for the difference $\mu_1 - \mu_2$ can be constructed using

$$\left[(\bar{x}_1 - \bar{x}_2) - z_{1-\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \quad (\bar{x}_1 - \bar{x}_2) + z_{1-\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

with approximate confidence level

$$P\left((\bar{X}_1 - \bar{X}_2) - z_{1-\frac{\alpha}{2}} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{X}_1 - \bar{X}_2) + z_{1-\frac{\alpha}{2}} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right) = 0.95\,.$$

where $z_{1-\alpha/2} = 1.96$. The automobile club tests 36 cars of company A and 40 cars of company B. The following quantities are calculated (in liters per 100 km, l/100 km) :

$$\begin{array}{l|l} \bar{x}_1 = 9.2 \text{ l/100 km}, & s = 0.6 \text{ l/100 km} \\ \bar{x}_2 = 8.4 \text{ l/100 km}, & s = 0.4 \text{ l/100 km} \end{array}$$

The confidence interval is:

$$\left[(9.2 - 8.4) - 1.96 \cdot \sqrt{\frac{0.6^2}{36} + \frac{0.4^2}{40}} \;\;,\;\; (9.2 - 8.4) + 1.96 \cdot \sqrt{\frac{0.6^2}{36} + \frac{0.4^2}{40}} \right] = [0.586 \;\;,\;\; 1.032]\,.$$

This interval does not cover 0. We will see later that this implies a statistically significant difference in mean gas consumption between the two populations.

Enhanced: Confidence Intervals of the Difference of Two Mean Stock Prices

Company X wants to analyze its share performance on two stock exchanges using the spot price which is observed daily at 12.00 p.m. The company is particularly interested in the difference of mean spot prices. We will construct both a point estimate and a confidence interval at level $1 - \alpha = 0.95$.

The random variables are:

$$X_1 = \text{“the spot price on the first stock exchange”},$$

$$X_2 = \text{“the spot price on the second stock exchange”}.$$

The means $E(X_1) = \mu_1$, $E(X_2) = \mu_2$ and variances $Var(X_1) = \sigma_1^2$, $Var(X_2) = \sigma_2^2$ are unknown. We assume that

- prices are independent at the two stock exchanges
- the variances are equal (variance homogeneity)

We draw a random sample from each population. The sample sizes are $n_1 = 10$ and $n_2 = 10$. Since the company X has been traded at the two stock exchanges for a long time, both populations are large. Hence we can assume that we are sampling with replacement. Moreover we assume independence of the two samples.

In demonstrating the construction of confidence intervals for the difference $\mu_1 - \mu_2$, consider the following two cases:

- X_1 and X_2 are normally distributed
- the distributions of X_1 and X_2 are unknown

1. Case

We have $X_1 \sim N(\mu_1; \sigma^2)$ and $X_2 \sim N(\mu_2; \sigma^2)$. The standardized random variable

$$T = \frac{D - E(D)}{S_D} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{n_1+n_2}{n_1 n_2} \cdot \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}},$$

is t-distributed with $n_1 + n_2 - 2 = 18$ degrees of freedom. Under these assumptions

$$\left[(\bar{x}_1 - \bar{x}_2) - t_{n_1+n_2-2;1-\frac{\alpha}{2}} s_D, \ ; (\bar{x}_1 - \bar{x}_2) + t_{n_1+n_2-2;1-\frac{\alpha}{2}} s_D\right]$$

Table 8.8 Spot prices for $n = 10$ randomly selected days

i	x_{1i}	x_{2i}	$(x_{1i} - \bar{x}_1)^2$	$(x_{2i} - \bar{x}_2)^2$
1	18.50	18.45	0.0841	0.1296
2	19.00	18.90	0.0441	0.0081
3	18.70	18.80	0.0081	0.0001
4	19.30	19.50	0.2601	0.4761
5	17.10	17.30	2.8561	2.2801
6	18.30	18.10	0.2401	0.5041
7	18.60	18.80	0.0361	00001
8	19.00	18.85	0.0441	0.0016
9	19.40	19.50	0.3721	0.4761
10	20.00	19.90	1.4641	1.1881

is a confidence interval for the difference of the two spot price means $\mu_1 - \mu_2$ at a confidence level

$$P\left((\bar{X}_1 - \bar{X}_2) - t_{n_1+n_2-2;1-\frac{\alpha}{2}} s_D \;\leq\; \mu_1 - \mu_2 \;\leq\; (\bar{X}_1 - \bar{X}_2) + t_{n_1+n_2-2;1-\frac{\alpha}{2}} s_D\right)$$
$$= 1 - \alpha = 0.95\,.$$

For $1 - \alpha = 0.95$, we find $t_{n_1+n_2-2;1-\alpha/2} = t_{18;0.975} = 2.101$.

For $n = 10$ randomly selected days, we record spot prices on each of the two exchanges, given in Table 8.8 in column 2 and 3. Columns 4 and 5 contain squared deviations from the estimated means which are used to calculate the individual variances.

We obtain:

$$\bar{x}_1 = 18.79 \qquad \bar{x}_2 = 18.81$$
$$s_1^2 = 0.601 \qquad s_2^2 = 0.563\,.$$

Since we have assumed homogeneity of variances, the point estimate s^2 for the joint or pooled variance σ^2 is given by the weighted arithmetic mean of the sample variances:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{9 \cdot 0.601 + 9 \cdot 0.563}{18} = 0.582\,.$$

The variance of the difference of the sample means is

$$s_D^2 = s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = 0.582 \cdot \frac{1}{5} = 0.1164\,.$$

and the standard deviation is $s_D = 0.3412$. A confidence interval for the difference is given by:

$$[(18.79 - 18.81) - 2.101 \cdot 0.3412 \ , \ (18.79 - 18.81) + 2.101 \cdot 0.3412] \\ = [-0.7369 \ , \ 0.6969] \, .$$

which is small relative to the levels of the individual spot prices. The confidence interval includes the value 0. Hence there does not appear to be an appreciable difference between the two mean spot prices μ_1 and μ_2. In a later chapter we will see how this implies that there is no statistically significant difference between the two prices.

2. Case

We will now drop the assumption of normality of X_1 and X_2. We will require larger sample sizes in order that we may rely upon the central limit theorem as an approximation to the distributions of $\bar{X}_1$ and $\bar{X}_2$ (and their difference $\bar{X}_1 - \bar{X}_2$). We will draw samples of size $n_1, n_2 = 50$. The standardized random variable

$$\frac{D - E(D)}{S_D} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{n_1+n_2}{n_1 n_2} \cdot \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}}$$

is approximately normally distributed. Under the above assumptions

$$\left[(\bar{x}_1 - \bar{x}_2) - z_{1-\frac{\alpha}{2}} s_D, \quad (\bar{x}_1 - \bar{x}_2) + z_{1-\frac{\alpha}{2}} s_D \right]$$

is an approximate confidence interval for the difference $\mu_1 - \mu_2$ at confidence level .95:

$$P\left((\bar{X}_1 - \bar{X}_2) - z_{1-\frac{\alpha}{2}} S_D \ \leq \ \mu_1 - \mu_2 \ \leq \ (\bar{X}_1 - \bar{X}_2) + z_{1-\frac{\alpha}{2}} S_D \right) \approx 1 - \alpha = 0.95 \, .$$

where $z_{1-\alpha/2} = z_{0.975} = 1.96$. Using our two samples of 50 observations each we obtain:

$$\bar{x}_1 = 18.80 \qquad s_1^2 = 0.5967$$
$$\bar{x}_2 = 18.83 \qquad s_2^2 = 0.6188 \, .$$

Since we assumed homogeneous variances, we estimate σ^2 using

$$s^2 = \frac{49 \cdot 0.5967 + 49 \cdot 0.6188}{98} = 0.6078 \, .$$

and

$$s_D^2 = s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) = 0.6078 \cdot \frac{1}{25} = 0.0243 \,.$$

The standard deviation is $s_D = 0.1559$.

The confidence interval is given by:

$$\begin{aligned} &[(18.80 - 18.83) - 1.96 \cdot 0.1559 \;\;,\;\; (18.80 - 18.83) + 1.96 \cdot 0.1559] \\ &= [-0.3356 \;\;,\;\; 0.2756] \,. \end{aligned}$$

The interpretation follows as in case 1 above.

Comparing the Two Approaches

- In case 1 we had more information about the population than in case 2.
- The difference of the two sample means and the joint variances are approximately of the same size in both cases.
- The variance s_D^2 and standard deviation s_D of the difference are much smaller in case 2 due to the larger sample size.
- The length of the confidence interval in case 2 is much smaller than in case 1.
- The confidence interval in case 2 is approximate because of the absence of exact knowledge of the underlying distributions.

Interactive: Confidence Intervals for the Difference of Two Means

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please select

- the confidence level $1 - \alpha$
- the sample sizes n_1 and n_2

Use

- "Draw sample" to manually construct a confidence interval
- "Reset" to reset the graphic
- the slider to cause an automated drawing of samples

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

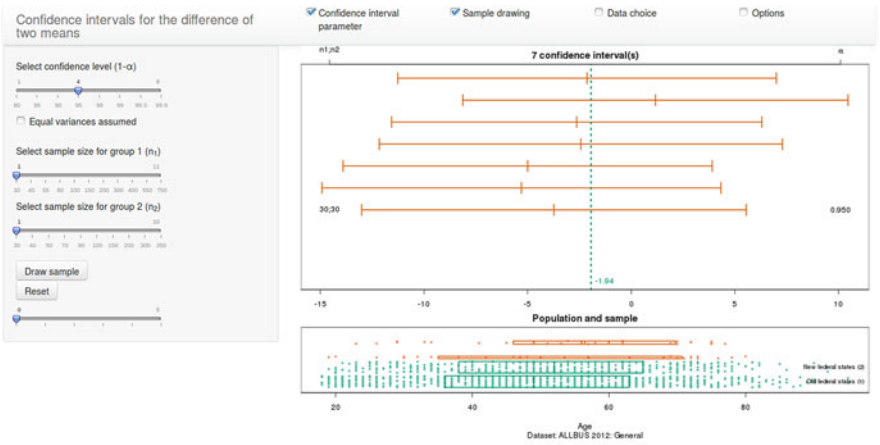


Fig. 8.10 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_ci2

Output

The upper graphic in Fig. 8.10 displays the resulting confidence intervals. The dashed line represents the difference between the two means in the population. After drawing an appropriate amount of samples we may see that $1 - \alpha$ of the observed intervals capture the population difference.

The graphic below shows a scatterplot of the population (green) and sample (orange).

8.9 Confidence Interval Length

The length of a confidence interval generally depends on the confidence level and on the sample size n. An increase in the confidence level $1 - \alpha$ (keeping the sample size n constant) yields a broader confidence interval. Increasing the sample size n (while keeping the confidence level $1 - \alpha$ constant), enhances precision and yields a smaller interval. Hence by adjusting the confidence level and sample size, we may control the width of the confidence interval.

Until now we have assumed that confidence level and sample size are given. In some applications, however, it is necessary to find the sample size which yields a confidence interval of prespecified width at a confidence level $1 - \alpha$.

The problem will be illustrated using confidence intervals for a mean μ and a proportion π. We will assume sampling with replacement from a large population.

(a) Confidence Interval for μ

We assume that the population is normally distributed. The exact sample size can be found, if the length of the sample size is not random, i.e., does not depend on the data. This is true if the variance σ^2 of the population is known. The length of the confidence interval for μ is given by:

$$L = 2 \cdot z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

and depends on the confidence level $1-\alpha$ as well as the sample size n. If the length L and the confidence level $1-\alpha$ are given, we may solve the above equation for n. More precisely, the required sample size is the smallest integer for which the condition holds:

$$n \geq \frac{4\sigma^2 z_{1-\frac{\alpha}{2}}^2}{L^2}.$$

In order to obtain a confidence interval not exceeding length L and confidence level $1-\alpha$, n has to be *at least* as large as this integer.
Hint: If the variance σ^2 is unknown, the length of the interval for μ

$$L = 2 \cdot t_{n-1;1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

is random since it depends on the standard deviation s which is a function of the sample. There are procedures which ensure that the **expected** length of the confidence interval equals some value, but these will not be considered here.

(b) Confidence Interval for π

Suppose we have a sufficiently large sample so that the sample proportion $\hat{\pi}$ is approximately normally distributed. The length of the confidence interval for π is given by

$$L = 2 \cdot z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}.$$

Given a prespecified value L, and confidence level $1-\alpha$, we could solve the above equation for the required sample size. More precisely, the required sample size

would be the smallest integer for which the following condition holds:

$$n \geq \frac{4 \cdot z^2_{1-\frac{\alpha}{2}} \hat{\pi} \cdot (1 - \hat{\pi})}{L^2}.$$

However, note that $\hat{\pi}$ is random, in which case the required sample size would vary from sample to sample. Fortunately, we can arrive at a conservative minimum sample size as follows. Note first that $\pi(1 - \pi)$ is maximal when $\pi = 0.5$ and $1 - \pi = 0.5$. This is the situation which requires, ceteris paribus, the largest sample size. Thus if we select

$$n \geq \frac{4 \cdot z^2_{1-\frac{\alpha}{2}} \cdot \frac{1}{2} \cdot \frac{1}{2}}{L^2} = \frac{z^2_{1-\frac{\alpha}{2}}}{L^2}.$$

then it will also be the case that

$$n \geq \frac{4 \cdot z^2_{1-\frac{\alpha}{2}} \pi \cdot (1 - \pi)}{L^2}.$$

for any other π. We need to take some care to ensure that sample size n is sufficiently large so that the normal distribution applies.

Explained: Finding a Required Sample Size

The Bimmelbahn Corporation would like to make a statement about the timeliness of its trains, in particular, the average delay and the proportion of timely trains. Confidence interval based on a random sample will be used.

1. Question

What should be the sample size in order to find a confidence interval for mean delay at a confidence level $1 - \alpha = 0.90$ and width 60 min? We assume that the random variable X = "duration of delays" is normally distributed with mean $E(X) = \mu$ and known variance $Var(X) = \sigma^2 = 68.8$. We want a confidence interval for μ. Note that $z_{1-\alpha/2} = z_{0.95} = 1.645$. Hence the required sample size is

$$n \geq \frac{4\sigma^2 z^2_{1-\frac{\alpha}{2}}}{L^2} = \frac{4 \cdot 68.8^2 \cdot 1.645^2}{60^2} = 14.23\,.$$

Thus if $n \geq 15$, the confidence interval will have the desired precision and width.

2. Question

What should be the sample size so that the confidence interval for π (the proportion of timely trains) be of length not exceeding 0.1 at a confidence level $1-\alpha = 0.95$? We assume the normal approximation holds for the distribution of $\hat{\pi}$ (rule of thumb: $n \geq 100$). Note that $z_{1-\alpha/2} = z_{0.975} = 1.96$.

We need to find n to satisfy:

$$n \geq \frac{4 \cdot z^2_{1-\frac{\alpha}{2}} \cdot \hat{\pi} \cdot (1-\hat{\pi})}{L^2}$$

A conservative bound for minimum sample size may be obtained by setting $\pi = 0.5$.

We obtain:

$$n \geq \frac{z^2}{L^2} = \frac{1.96^2}{0.1^2} = 384.16\,.$$

Thus to achieve the desired width and confidence level, we need $n \geq 385$.

Enhanced: Finding the Sample Size for an Election Threshold

The leader of a small political party would like to know whether the party will receive more than 5 % of the vote if the election were held tomorrow. He has appointed a statistician to perform the analysis. During their conversation the statistician highlights the following issues:

- In order to find the exact proportion of supporters, one would have to ask all the voters (i.e., the whole population).
- The proportion of votes in the sample is but an estimate or approximation of the true proportion.
- The confidence interval provides a measure of the uncertainty associated with the estimate.
- The length and level of confidence may be chosen by the politician.
- The shorter the required interval and the higher the confidence level, the larger the sample size.

The statistician calculates the required sample size using

$$n \geq \frac{4 \cdot z^2_{1-\frac{\alpha}{2}} \cdot \hat{\pi} \cdot (1-\hat{\pi})}{L^2}\,.$$

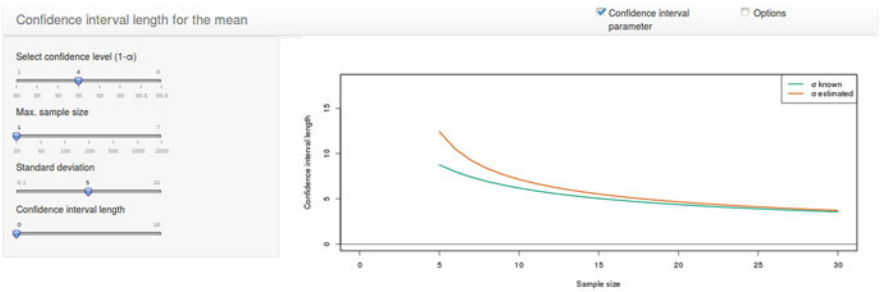


Fig. 8.11 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_cilen

Since $\hat{\pi}$ is unknown, the statistician uses the largest imaginable proportion of votes for his party. That proportion is 10 %. (This is because $\pi \cdot (1 - \pi)$ increases with π.) This yields a conservative value for minimum sample size.

Interactive: Confidence Interval Length for the Mean

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please select

- the confidence level $1 - \alpha$
- the maximum sample size n
- the standard deviation σ
- the confidence interval length L

Output

The plot in Fig. 8.11 displays a graph showing the changes of the confidence interval length L if the sample size n varies.

Chapter 9
Statistical Tests

9.1 Key Concepts

Statistical tests are tools for the analysis of hypotheses about the characteristics of unknown probability distributions or relationships between random variables. If the probability distribution is specified up to a finite set of parameters, testing for the fully specified probability density amounts to testing whether the parameters take on specific values. As the mathematical specification of a class of probability distributions involves writing down a function that contains parameters whose values aren't known a priori, tests based on postulated parameters that determine the characteristics of a probability distribution are dubbed "parametric" tests. Statistical estimation procedures can be used to obtain estimates of the specific parameter(s) of interest. Statistical test theory provides a means of quantifying the significance of such estimates. Closely related to the choice of the parameter value(s) is the choice of the class of probability distributions. Such a fully specified distribution has to describe reality as accurately and reliably as possible. In practice, the choice of a functional class such as the Normal (or Gaussian) distribution and estimating and testing parameters is an iterative process. Empirical researchers will have to consider various models (alternative distributions) at the explorative stage of the investigation into the nature of the phenomena of interest. However, very often certain probability models are chosen a priori for their tractability rather than on theoretical grounds. When the postulated class of distribution functions is theory-driven in that it is the result of logical deduction from accepted premises, testing for the significance of parameters forms an important part of the verification of scientific theory. Much of empirical research is, however, data-driven in that there is no a priori distribution function.

The objective of a parametric statistical hypothesis test procedure can be summarized as follows. Given a certain population with parametric distribution function $F(x)$ (with parameters such as expected value μ and variance σ_2 in the class of

W.K. Härdle et al., *Introduction to Statistics*, DOI 10.1007/978-3-319-17704-5_9

normal distributions or the proportion π in a repeated Bernoulli experiment), some "guess" (hypothesis) about the true parameter value(s) has to be tested on the basis of an observed sample of finite size. Clearly, this would not be necessary if one could observe the random variable under consideration for all members of the population of interest (which of course is not theoretically possible for continuous random variables, as a continuous distribution is comprised of infinitely many possible outcomes). In general, (sub-)samples cannot convey all the information necessary to precisely describe the underlying distribution (even if they are representative in terms of some suitable concept) and are the result of a (random) sampling process, therefore their implied parameter values (as determined by statistical estimation procedures conducted with the sample data) are themselves random variables. Often these estimates will only equal the correct population parameter value on average. Fortunately, statistical tests provide an appropriate yardstick that allows us to quantify and assess whether the difference between the sample-specific (i.e., statistically estimated) and hypothesized parameter values is statistically significant. In short, we evaluate whether our hypothesized parameter value is close enough to the estimated parameter value for the sampling process to have caused the difference or whether the two numbers (respectively vectors) cannot be reconciled even after having allowed for sampling noise (i.e., whether the noise created by observing on only a finite sample of elements can account for the difference).

In order to put the above verification problem on an objective decision theoretical basis, statistical tests have been devised to tackle the problems that may otherwise lead us to rely on subjective assessments. Questions that we will need to address include:

- What is the "correct" formulation of the actual hypothesis in mathematical terms?
- How is the data to be condensed? (i.e., which statistics or estimators are to be used)
- How is the difference of the condensed collected data from the structure implied by the hypothesis to be quantified? (i.e., what expression will we use for our test statistic)
- How is the quantified difference to be evaluated in decision-theoretical terms? When is the difference *statistically significant*? (i.e., what is the distribution of our test statistic and what is acceptable sampling noise)

To provide an objective rational for verification of hypotheses (given certain assumptions about the functional class of the distribution, etc.), statistical tests must satisfactorily address *all* of the above issues.

We can get a grasp of the key concepts and terms of statistical tests by considering an example of a parametric test. Let θ be a parameter of the distribution function of random variable X. Its true value is unknown, but we can specify the *parameter space*, which is the set of possible values it can assume.

Table 9.1 One-sided and two-sided hypothesis tests

	Null hypothesis	Alternative hypothesis
a) Two-sided test	$H_0 : \theta = \theta_0$	$H_1 : \theta \neq \theta_0$
b) One-sided tests		
Right-sided test	$H_0 : \theta \leq \theta_0$	$H_1 : \theta > \theta_0$
Left-sided test	$H_0 : \theta \geq \theta_0$	$H_1 : \theta < \theta_0$

Formulating the Hypothesis

The hypothesis states a relation between the true parameter θ and the hypothetical value θ_0. Usually a pair of connected hypotheses is formulated, the *null hypothesis* H_0 and the *alternative hypothesis* H_1.

The null hypothesis is the statistical statement to be tested; thus it has to be formulated in such a way that statistical tests can be performed upon it. Sometimes the underlying scientific hypothesis can be tested directly, but in general the scientific statement has to be translated into a statistically tractable null hypothesis. In many cases the null hypothesis will be the converse of the conjecture to be tested. This is due to certain properties of parametric statistical tests, which we will be dealing with later on.

The asserted relation between the true parameter θ and the hypothetical value θ_0 is stated so that the combined feasible parameter values of both null and alternative hypothesis capture the entire parameter space. Clearly the alternative hypothesis can be thought of as a converse of the null hypothesis. The possible variants are given in Table 9.1

The two-sided hypothesis in a) is a so-called simple hypothesis, because the parameter set of the null hypothesis contains exactly one value. As the alternative hypothesis highlights, deviations from the hypothetical value θ_0 in both directions are relevant to the validity of the hypothesis. That's why it is referred to as two-sided.

The hypotheses of the one-sided tests under b) belong to the class of composite hypotheses. "Composite" refers to the parameter set of the null hypothesis being composed of more than one value. Consequently, not rejecting the null hypothesis wouldn't completely specify the distribution function, as there is a set of (in above cases infinitely many) parameter values that have not been rejected. The hypotheses are one-sided, because deviation from the hypothetical parameter value in only one direction can negate the null hypothesis—depending on that direction these tests are further characterized as left- or right-sided.

Clearly, the scientific problem to be formulated in statistical terms determines which test will be of interest (applied).

Note some important principles of hypothesis formulation:

- Statistical test procedures "test" (i.e., reject or do not reject) the null hypothesis.
- Null and alternative hypothesis are disjoint, that is, their respective parameter spaces don't contain the same value.
- Parameter sets encompassing exactly one value will always belong to the null hypothesis.

Test Statistic

In order to follow the above procedure, we need a quantity to base our decision rule on. We need a suitable estimator in order to extract the information required to properly compare the hypothetical with the sample-specific parameter value(s).

If an estimator is used as a verification quantity within a statistical test procedure, we call it a test statistic, or simply a statistic. We will denote the statistic by $V = V(X_1, \ldots, X_n)$.

The statistic V is a function of the sample variables $X_1, \ldots, X_n$ and hence itself a random variable with some distribution $F_V(v)$. In order to conduct a statistical test, the distribution of V for a valid null hypothesis has to be known (at least approximately). Thus, we consider F_V conditional on (given) the null hypothesis: $F_V = F_V(v|\mathrm{H}_0)$.

So in the case of a parametric test this means that the distribution of the test statistic depends on the (unknown) parameter θ: $F(v \mid \theta)$. In order to determine this distribution, the parameter θ has to be specified numerically. But the only a priori information about θ at hand is the hypothetical boundary value θ_0. Thus we will now (at least for the time being) assume that θ_0 is the true parameter value prevailing in the population, i.e., $\theta = \theta_0$. In a two-sided test, this assumption accurately reflects the null hypothesis. In a one-sided test, the boundary value θ_0 must belong to the null hypothesis—one reason, why "equality," i.e., $\theta = \theta_0$ always belongs to the parameter space of the null hypothesis. For all three possible test scenarios we are thus assuming that the test statistic V has a distribution with parameter θ_0 under the null hypothesis.

Observing the random variable under consideration on n statistical observations yields a sample $x_1, \ldots, x_n$. Plugging these realizations into the test statistic gives a realization of the test statistic: $v = v(x_1, \ldots, x_n)$.

Decision Regions and Significance Level

Being a random variable, the test statistic can take on one of several possible values. If the test statistic for a given sample is sufficiently close to the hypothetical parameter value, the difference may be considered "random." In this case the null hypothesis won't be rejected. Yet this doesn't mean that the null hypothesis is correct (or has been 'accepted') and hence that θ_0 is the true parameter value. The only permissible statement is that, given the particular sample, it cannot be ruled out for a certain degree of confidence, that the underlying population follows a distribution specified by the parameter value θ_0.

Large deviations of the test statistic from the hypothetical parameter value make the null hypothesis appear implausible. In this situation the sample may "as well"

have been generated by a population distributed according to parameter values suggested in the alternative hypothesis. We can then assume that a parameter value other than θ_0 specifies the true population distribution. Still that doesn't mean θ_0 is wrong with certainty. We can only say that it is very unlikely that a population following the thus specified probability distribution has generated the sample we have observed.

Following these considerations, the set of possible test statistic realizations is partitioned into two disjoint regions, reflecting whether the observed sample can be reconciled with the null hypothesis for a given level of "plausibility" (non-rejection region) or not (rejection region).

Non-rejection Region of Null Hypothesis

The non-rejection region for H_0 is the set of possible outcomes of the test statistic leading to a decision in favor for H_0, i.e., non-rejection for H_0.

Rejection Region of Null Hypothesis

The rejection region (or critical region) for H_0 encompasses all possible outcomes of the test statistic that lead to a rejection for H_0.

Rejection and non-rejection regions for H_0 form a disjoint and exhaustive decomposition of all possible outcomes of the test statistic. If the outcomes are real-valued, there are boundary values termed "critical values" that partition the real line into rejection and non-rejection regions. The critical value itself belongs to the non-rejection region.

In order to obtain a usable, decision rule, these critical values have to be computed. This is accomplished using probability theory.

The probability, that, any sample induces the test to reject H_0 given the null hypothesis is actually true (i.e., the true parameter value falls into the region stated in the null hypothesis) must not be greater than the significance level α:

$$P\,(V \text{ is element of rejection region for } \mathrm{H}_0 \mid \theta_0) \leq \alpha.$$

Accordingly, the probability of V assuming a value in the non-rejection region, when V is computed from a sample drawn from a population with parameter θ_0, is at least $(1 - \alpha)$:

$$P\,(V \text{ is element of non-rejection region associated with } \mathrm{H}_0 \mid \theta_0) \geq 1 - \alpha.$$

Given the probability α, critical values can be derived from the test statistics' conditional probability distribution $\mathrm{F}(v|\mathrm{H}_0)$. This helps us to understand why the distribution of the test statistic given H_0 is true must be known (at least approximately).

As the probability α determines whether any given sample deviates significantly from the value implied by the hypothesized parameter set, it is termed the *level of significance*. For heuristic reasons (mainly historical in nature), the significance level is chosen to be small such that the null hypothesis is only rejected if the sample is very unlikely to stem from the hypothesized distribution—usually either 0.01, 0.05 or 0.10.

We will now derive decision regions for the three test scenarios we have introduced earlier for a given significance level α and validity for H_0. For convenience's sake in what follows below we assume V to be normally distributed.

Two-Sided Test

$$\mathrm{H}_0 : \theta = \theta_0 \quad \text{versus} \quad \mathrm{H}_1 : \theta \neq \theta_0$$

Rejection Region for H_0
In a two-sided test, the rejection region is composed of two sets (areas), as deviations of the sample statistic from the hypothesized parameter value θ_0 in two directions matter. The non-rejection region is separated from these two rejection regions by two critical values c_l and c_u (it actually resides between the two portions of the rejection region—this helps explain why two-sided tests are also often referred to as two-tailed tests, the two rejection regions reside in the tails of the probability distribution of V).

The rejection region consists of all realizations v of the test statistic V smaller than the *lower* critical value c_l or greater than the *upper* critical value c_u:

$$\{v|v < c_l \text{ or } v > c_u\}.$$

The combined probability of sampling a value from the rejection region, given H_0 (i.e., θ_0) is true, equals the given significance level α:

$$\mathrm{P}(V < c_l|\theta_0) + \mathrm{P}(V > c_u|\theta_0) = \alpha/2 + \alpha/2 = \alpha$$

Non-rejection Region for H_0
The non-rejection region for H_0 encompasses all possible values v of the test statistic V smaller than (or equal to) the upper critical value c_u and greater than (or equal to) the lower critical value c_l:

$$\{c_l \leq v \leq c_u\}.$$

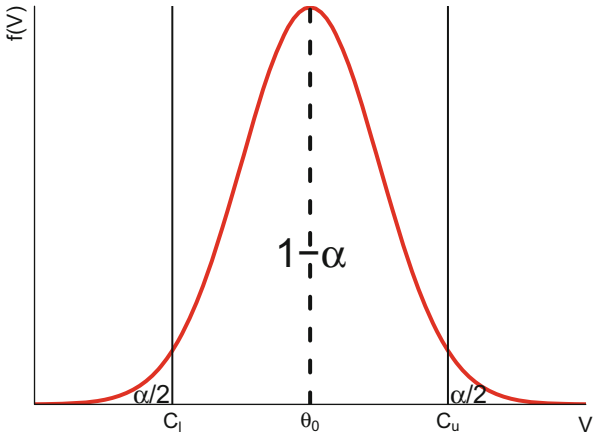


Fig. 9.1 Distribution of test statistic V for two-sided test

The probability of encountering a test statistic realization within the non-rejection region, given θ_0 is true, is $(1-\alpha)$ (Fig. 9.1):

$$\mathrm{P}\{c_l \leq V \leq c_u \,|\, \theta_0\} = (1-\alpha)\,.$$

One-Sided Tests

By design, there is exactly one critical region associated with one-sided tests: Deviations of the test statistics from the hypothetical parameter value are "significant" in only one direction. The critical value splitting non-rejection and rejection region is denoted by c.

1. Left-sided test:

$$\mathrm{H}_0 : \theta \geq \theta_0 \quad \text{versus} \quad \mathrm{H}_1 : \theta < \theta_0$$

Rejection region for H_0 The critical or rejection region for H_0 consists of realizations v of the test statistic V smaller than c:

$$\{v \,|\, v < c\}\,.$$

The probability that the test statistic assumes a value from the rejection region, given H_0 is true, is less than or equal to the significance level α:

$$\mathrm{P}\{V < c \,|\, \theta_0\} \leq \alpha.$$

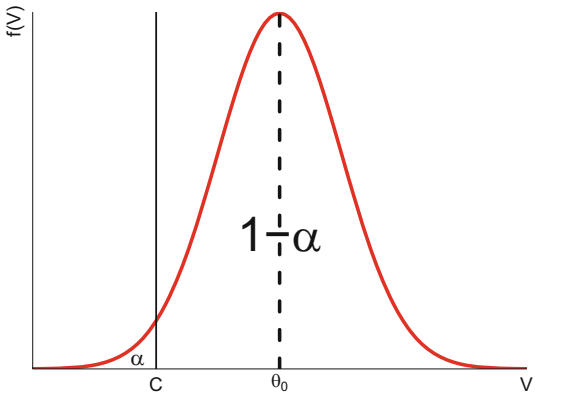


Fig. 9.2 Distribution of test statistic V for left-sided test

Non-rejection Region for H_0 The non-rejection region for H_0 encompasses all realizations v of the test statistic V greater than or equal to c:

$$\{v \mid v \geq c\}.$$

The probability of the test statistic assuming a value within the non-rejection region, given H_0 is true, is at least $(1 - \alpha)$ (Fig. 9.2):

$$P\{V \geq c \mid \theta_0\} \geq 1 - \alpha.$$

2. Right-sided test:

$$H_0 : \theta \leq \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0$$

Rejection region for H_0 The rejection region for H_0 consists of all realizations v of the test statistic V greater than c:

$$\{v \mid v > c\}.$$

The probability of v falling into the rejection region, given H_0 is true, is less than or equal to the given (chosen) significance level α:

$$P\{V > c \mid \theta_0\} \leq \alpha.$$

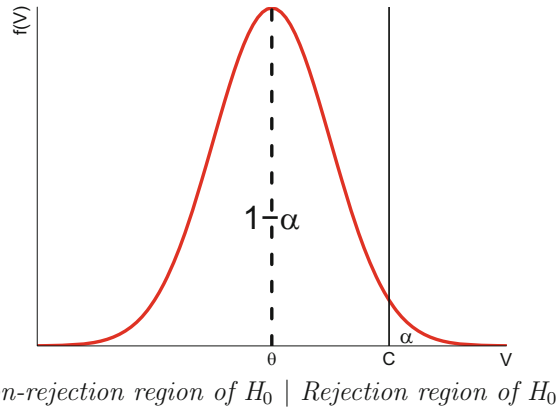


Fig. 9.3 Distribution of test statistic V for right-sided test

Non-rejection Region for H_0 The non-rejection region for H_0 is the set of test statistic values v less than or equal to c:

$$\{v \mid v \leq c\}.$$

The probability of v assuming a value from the non-rejection region, given H_0 is true, is greater than or equal to $(1 - \alpha)$ (Fig. 9.3):

$$P\{V \leq c \mid \theta_0\} \geq 1 - \alpha.$$

As statistical tests are based on finite samples from the (theoretically infinitely large) population, wrong decisions concerning the parameter values specifying the underlying distribution cannot be ruled out.

Depending on the actual value of the test statistic v, the null hypothesis will either be not-rejected or rejected. We will symbolize this as follows:

'H_0' : Test does not-reject the null hypothesis.

'H_1' : Test rejects the null hypothesis.

Irrespective of the decisions made on the basis of particular samples, there are two possible "true" states of the world, only one of which can be true at any point in time:

H_0 : Null hypothesis is "really" true.

H_1 : Null hypothesis is wrong, i.e., the alternative hypothesis is true.

Table 9.2 (2×2)-table of test decision and true situation

Sample-based decision	True parameter in population distribution	
	H_0	H_1
'H_0' (i.e., Test does not-reject H_0)	*Right Decision* 'H_0'$\vert H_0$ $P(\text{'}H_0\text{'}\vert H_0) = 1-\alpha$	*Type II error* 'H_0'$\vert H_1$ $P(\text{'}H_0\text{'}\vert H_1) = \beta$
'H_1' (i.e., Test rejects H_0)	*Type I error* 'H_1'$\vert H_0$ $P(\text{'}H_1\text{'}\vert H_0) = \alpha$	*Right Decision* 'H_1'$\vert H_1$ $P(\text{'}H_1\text{'}\vert H_1) = 1-\beta$

Joining the categorizations of the sample-induced test decision and true situation together yields a (2×2)-table of possible combinations shown in Table 9.2.

The Null Hypothesis Is True

Let us first examine the nature of the wrong and right decision to be made given the null hypothesis H_0 is true "in reality."

Suppose, a test statistic computed using an observed sample deviates substantially from the proposed boundary parameter value θ_0. It is in fact the scope of statistical tests to rationally assess these deviations in terms of significance, i.e., evaluate whether the deviation *is* substantial in statistical terms. But for the moment assume that the deviation is substantial in that the test statistic realization v falls into the rejection region. Following the decision rule created for the test, the null hypothesis will be rejected. Yet our decision doesn't affect the true data generation process, and consequently we may have made an error which we expect to make with probability α (when our null hypothesis is true). This error is dubbed *type I error* or α-error, and its (probabilistic) magnitude is what we control when we set up the test procedure. By fixing (choosing) α we set the probability

$$P(\text{'}H_1\text{'}\vert H_0) = P(\text{Test rejects null given the null is true}) = \alpha$$

as a parameter—the significance level. Even though we can vary the significance level α, we cannot completely prevent the occurrence of a Type I Error (which will occur with probability α). Setting α to zero amounts to never rejecting the null hypothesis, consequently never rejecting given the null hypothesis describes reality correctly. The probability of making the right decision, given the null hypothesis is true is computed as

$$P(\text{'}H_0\text{'}\vert H_0) = P(\text{Test does not-reject the null given the null is true}) = 1-\alpha,$$

which equals one, if we set α to zero. As tempting as setting α to zero sounds there is a down side which we will see occurs when the alternative, rather than the null, hypothesis is true.

The Alternative Hypothesis Is True

What are the right and wrong decisions that can be made when the *alternative* hypothesis states the true parameter range?

If the test statistic computed from an observed sample indicates a relatively small deviation from the parameter value θ proposed in the null hypothesis, the decision rule will induce us to not-reject the null hypothesis H_0. Since we are presently postulating H_1 to be true we know that this is an error. This outcome 'H_0'$|H_1$ (non-rejection of a false null) is commonly known as the *type II error* or β-error.

As is the case in the situation called α-error, we cannot rule out the β-error either: Even though it is "unlikely" that a sample drawn from a population that does not belong to the null hypothesis gives a test statistic value "close" to the null hypothesis value, it is still possible—and this will happen with probability

$$P\left('H_0'|H_1\right) = \beta\left(\theta_1\right),$$

given the alternative hypothesis correctly describes reality.

Note that β depends on the true parameter value θ_1. As this still hasn't been disclosed to us (and never will), we cannot compute this probability.

There is, of course, also the possibility of the decision rule inducing us to make a right decision, i.e., reject H_0 when the alternative hypothesis is true: 'H_1'$|H_1$. The conditional probability of this happening is (conditional on the alternative hypothesis being true):

$$P\left('H_1'|H_1\right) = 1 - \beta\left(\theta_1\right).$$

The probability $\beta\left(\theta_1\right)$ of making a type II error depends on the given significance level α. Decreasing α for a constant sample size n will result in an increased probability of the β-error, and vice versa. This "error trade-off" cannot be overcome, that is, it is not possible to reduce α whilst also reducing βThis dilemma is depicted in Figs. 9.4 and 9.5.

As already mentioned, the probability of making a type II error also depends on the true value of the parameter to be tested. Given a fixed sample size n and significance level α, the distance between θ_1 and θ_0 is inversely related to $\beta\left(\theta_1\right)$: The greater the distance, the smaller is the probability of making a type II error when the alternative hypothesis is true.

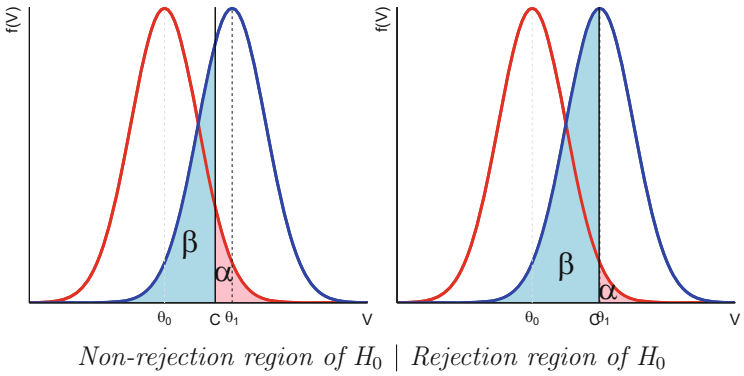


Fig. 9.4 Relationship between significance level and type II error

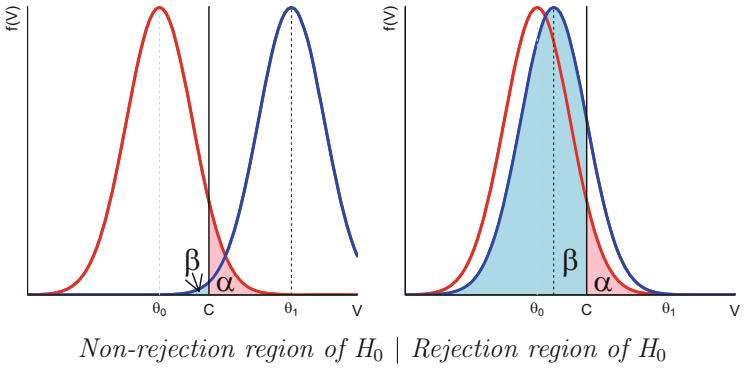


Fig. 9.5 Distribution of test statistic V under null and alternative hypothesis

The following two diagrams show this for our normally distributed test statistic V.

Interpretation of Test "Results": Reasoning Behind Controlling the Type I Error

Statistical inference is a means of inferring probability distributions (or their characteristics, e.g., parameters like the expected value) from samples with limited size for either practical or economical reasons. As these subsets of the population don't convey the complete information about the distribution of the variable under consideration, making errors is inevitable. All we try to achieve is to quantify and control them in the sense that in a repeated sampling context they occur with a certain probability. As already pointed out: Rejecting a hypothesis doesn't prove it wrong—the probability of the hypothesis actually being right (i.e., of making a type

I error) merely doesn't exceed a small threshold that is set by the researcher. Not-rejecting the null hypothesis exposes the researcher to the risk of making an error of type II which occurs with a probability that cannot be quantified statistically. As we have seen, depending on the true parameter, the corresponding probability β can be "significantly" greater than the controlled α-probability. For this reason, the scientific conjecture to be tested statistically is usually chosen as the null, rather than the alternative, hypothesis so that the probability of rejecting it in error (a type I error) can be controlled. The possibility of a reject H_0 decision being wrong can then be quantified to be no more than α. The same logic applies, if the decision object is of high ethical or moral importance, e.g., human health when it comes to testing a new drug or the assumption of innocence until guilt is proven in the case of suspected criminals.

Power of a Test

The probability of rejecting the null hypothesis as a function of all possible parameter values (i.e., those θ of the null *and* alternative hypothesis) is called the power of a test, denoted by $P(\theta)$:

$$P(\theta) = P(V \text{ is element of the rejection region for } H_0 \mid \theta) = P(\text{`}H_1\text{'} \mid \theta).$$

If the true parameter θ is element of the subset of the parameter space stated in the alternative hypothesis, a right decision has been made: (`H_1'$|H_1$). Hence, for all true parameter values θ that agree with the alternative hypothesis, the power measures the probability of correctly rejecting the null hypothesis (respectively correctly not-rejecting the alternative hypothesis):

$$P(\theta) = P(\text{`}H_1\text{'} \mid H_1) = 1 - \beta(\theta)\ ; \quad \forall \theta \in \theta_1,$$

where θ_1 is the subset of the parameter space (the parameter space is the set of all parameters that θ can equal) specified by the alternative hypothesis.

If the true parameter equals θ_0, the set of values under the null hypothesis, the power returns the probability of making a wrong decision, i.e., the probability of the situation (`H_1'$|H_0$). This is a familiar quantity, namely, the probability of making a type I, or α error:

$$P(\theta) = P(\text{`}H_1\text{'} \mid H_0) \leq \alpha(\theta)\ ; \quad \forall \theta \in \theta_0,$$

where θ_0 is the subset of the parameter space specified by the null hypothesis.

The power measures the reliability of a test procedure in correctly rejecting a false null hypothesis.

OC-Curve

The operating characteristic (OC-curve) is equal to $1 - P(\theta)$, it provides the probability of not rejecting the null hypothesis as a function of all possible θ:

$$1 - P(\theta) = P(V \text{ is element of the non-rejection region for } \mathrm{H}_0|\theta) = P(\text{`}\mathrm{H}_0\text{'}|\theta).$$

If the true parameter θ is a member of the subset of the parameter space associated with the alternative hypothesis, the operating characteristic assigns a probability of making the wrong decision ($\text{`}\mathrm{H}_0\text{'}|\mathrm{H}_1$), that is, the probability of making a type II error:

$$1 - P(\theta) = P(\text{`}\mathrm{H}_0\text{'}|\mathrm{H}_1) = \beta(\theta)\ ; \quad \forall\theta \in \theta_1,$$

where θ_1 is the subset of parameters specified by the alternative hypothesis.

If, on the other hand, the true parameter is in the subset of values specified by the null hypothesis, the operating characteristic measures the probability of the situation,($\text{`}\mathrm{H}_0\text{'}|\mathrm{H}_0$), i.e., making the right decision in not rejecting the null hypothesis:

$$1 - P(\theta) = P(\text{`}\mathrm{H}_0\text{'}|\mathrm{H}_0) \geq 1 - \alpha(\theta)\ ; \quad \forall\theta \in \theta_0,$$

where θ_0 is the parameter set of the null hypothesis.

The shape of the graph of the operating characteristic curve (similarly the power curve) depends on the:

- test statistic and its distribution, which must be determined not only for the boundary parameter value delineated by the null hypothesis θ_0, but also for all admissible parameter values;
- given significance level α; and
- sample size n.

A Decision-Theoretical View on Statistical Hypothesis Testing

In the absence of a consistent rational for the conduct of empirical research (and the economic trade-offs involved in deciding what proportion of resources to allocate to different competing lines of research/thought), the scientific community has more or less explicitly agreed that certain significance levels (most notably 0.05 and 0.01) are adequate. Of course, use of these vary with the degree to which various measured variables or impacts can be accurately quantified. If making errors can be tackled within a cost benefit decision-making approach, an approximate

collective preference order can be assumed that strikes a balance between the long-term scientific success of a society, economic success and short-term costs. As it is impossible to predict the future value of undertaking a particular scientific effort, the economics of science as an allocation tool itself has to deal with uncertainty in the level of generated knowledge for each feasible research environment. For these reasons, significance levels chosen for empirical research not closely linked to a specific application will always be conventions based on some human perception of how frequent "infrequent" should be. But even on the more applied level, significance levels aren't usually the result of a systematic analysis of the relevant preference system and the experimental conditions. Consider some crucial problems of public choice. In deciding how many ambulances to fund for a particular area, a community actively caps the number of patients to be catered for at the same time. If you wanted to test whether three ambulances are sufficient, i.e., not more than three citizens become critically ill at any time, where would you fix the significance level? Setting it to zero would imply the decision is to buy as many ambulances and employ as many staff as there are citizens, if one cannot rule out the occurrence of an epidemic possibility of all citizens coincidentally becoming ill at the same time. Clearly, this is not feasible in any society. No matter which significance level the decision-maker chooses—she will always have to accept the possibility of (rather) unlikely events causing unfortunate outcomes for society (i.e., deaths in the community in the case of choice of how many ambulances).

As noted above, the choice of a suitable significance level is—more or less—arbitrary, because at least one important component of the specification of the decision problem cannot be observed or formalized: on the general level of fundamental research, the future benefits are unknown or they cannot be compared to today's resource spending as their pecuniary value cannot be determined. On the more applied level, research into health or other issues related to the well-being of humans cannot be rationalized for the intangibility of the involved "commodities" (i.e., health). But there are certain applications that can be reduced to cost benefit analysis. Carrying out sample-based quality control in a manufacturing company, for example, requires inspectors to accurately quantify the impact of given choices on the proportion of defective output. She can estimate the expected number of returned items and resulting currency value of related lost sales, etc. as market prices (values) already exist for such items. The preference order applied could for example be the appetite of shareholders to face a particular risk-return profile implied by the choice of alternative work practices.

More Information: Examples

Statistical tests are procedures for the analysis of assumptions about unknown probability distributions or their characteristics. The "behavior" of random variables from a population is inferred from a sample of limited size, constrained by

either practical or economical parameters. This inductive character makes them an important pillar of inferential statistics, the second branch being statistical estimation procedures. We will now illustrate the theory introduced in this chapter with some practical examples.

Example 1 A large software company is subject to a court trial, media coverage bringing it to the forefront of public debate. The management wants to assess the impact of the legal action on revenues. Average monthly sales before the lawsuit are known, serving as the hypothetical value to be tested. The average of a randomly selected sample of monthly revenues from the time after the trial firstly hit the news is calculated. The directors are particularly interested whether the revenues have fallen and ask their in-house statistician to test the hypothesis that the average monthly revenue has fallen since the beginning of the lawsuit. Hence, the monthly revenue is treated as a random variable, and the test is based on its mean.

Example 2 An environmental organization claims that the proportion of citizens opposed to nuclear power is 60 %. The operators of the nuclear power plants dismiss this figure as overstated and commission a statistical analysis based on a random sample. The variable "attitude to nuclear energy" is measured by only two outcomes, e.g., "support" and "opposed." Hence, the statistician tests the mean of the population distribution of a dichotomous variable: Can the hypothetical value of 0.6 be reconciled by the sample?

In both examples an unknown parameter of the probability distribution in the population is tested. The test procedures employed are known as *parametric tests*. Furthermore, as they are based on one single sample, they are called *one-sample tests*.

Example 3 Two producers of mobile phones launch separate advertising campaigns claiming to build the phones with the longest stand-by time. Clearly, one of them must be wrong—given, that stand-by time is sufficiently precisely measured such that the average stand-by time doesn't coincide *and* standby time varies across individual phones, i.e., is a random variable. The managers of a consumer organization are concerned and want to assess whether the cellular phones manufactured by the two companies differ significantly with respect to stand-by time. The statistical investigation has to be based on an average to account for the fluctuations of stand-by time across output. Samples are drawn independently from both producers' output in order to compare the average location of the duration as measured by the sample means. An inductive statement is sought whether or not the mean stand-by times in the overall outputs are (significantly) different or not.

The test procedure applied is a *parametric test*, as one tests for equality of the two *means*. This can only be done on the basis of two samples: This is an example of a *two-sample test procedure*.

Example 4 Someone claims that a specific die (single dice) is what statisticians call a *fair die*: The probability of any outcome is equal. The hypothesis to be tested is that the outcomes of the die rolling process has a discrete uniform distribution.

This test doesn't refer to a parameter of the underlying population distribution, i.e., doesn't take a particular distribution class as given. Consequently, it is classified as *nonparametric test* or *distribution-free test*. This particular type belongs to the class of *goodness-of-fit tests*, as one wants to verify how good a given sample can be explained to be generated by a particular, completely specified, theoretical probability distribution.

More Information: Hypothesis Testing Using Statistical Software

Let's assume you want to carry out a right-sided statistical test about a parameter θ: $H_0 : \theta \leq \theta_0$ and $H_1 : \theta > \theta_0$ For simplicity, we also assume that the test statistic V follows a standard normal distribution (i.e., a normal distribution with mean 0 and variance 1).

The rejection region for H_0 is the set of all test statistic realizations v greater than the critical value c: $\{v|V > c\}$. The probability of the test statistic assuming a value within the rejection region equals the given (chosen) significance level, $\alpha = P(V > c|\theta_0)$, and is given by the green area in Fig. 9.6.

The test decision is made by comparing the realized test statistic value with the critical value: If the realized test statistic value, computed from a particular sample of size n, is greater than the critical value, then the null hypothesis is rejected. The critical value splits the distribution of all possible test statistic values into two sets with probabilities α and $1 - \alpha$.

Popular statistical software packages (e.g., SAS, SPSS, Statistica, Systat, XploRe, R) not only compute the test statistic value v, but additionally return a so-called *p-value*. This is the theoretical probability that V assumes a value greater

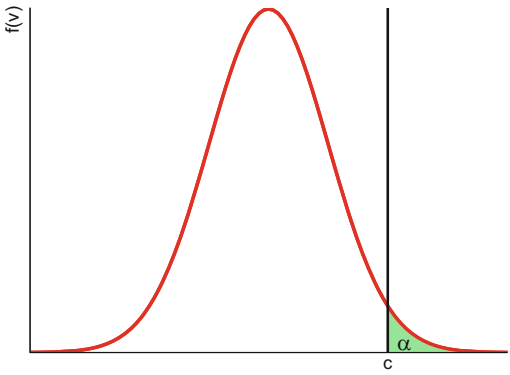


Fig. 9.6 The rejection region of test statistic V

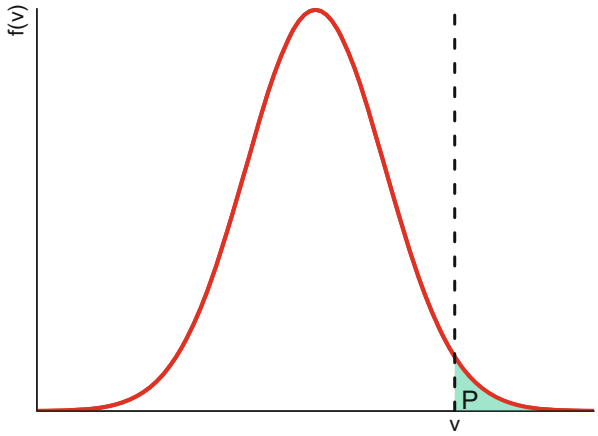


Fig. 9.7 Illustration of the *p*-value

than that computed from the given sample: $P(V > v|\theta_0)$. The *p*-value is sometimes called *significance* or *1-tailed P*, and we will denote it by $p = P(V > v|\theta_0)$. The crucial assumption underlying its computation is that the distribution of V is the one that follows from assuming that θ_0 is the true parameter value. In Fig. 9.7, p is depicted by the blue area.

As the *p*-value represents the minimum significance level for not rejecting the null hypothesis, the user doesn't need to look up the critical value corresponding to the given significance level in a table. She merely needs to compare α with the size of the *p*-value as follows:

Rejecting the Null Hypothesis

If the parameter estimate is "substantially" larger than the hypothetical parameter value θ_0, the *p*-value will be relatively small. Recall that the null hypothesis is one sided with values less than or equal to θ_0, consequently estimates that are greater than θ_0 are less easily reconciled with the null hypothesis than those within the postulated parameter range. The "farther" away the estimate lies from the null hypothesis, the less probable it is to have been generated by sampling from a population distribution with θ less than or equal to θ_0. The *p*-value is the probability that v will be observed given a true parameter θ_0. In our example, this becomes decreasingly likely with rising parameter estimate, and a sufficiently large parameter

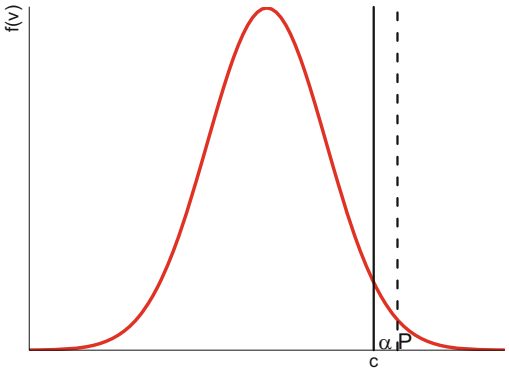


Fig. 9.8 Case of rejection

estimate will induce us to infer that θ_0 and θ differ significantly. Given a parameter estimate, the p-value tells us how likely the observed distance to θ_0 is to occur. When this probability is small, the risk of being wrong in rejecting the null hypothesis is small. That is, we conclude that the null hypothesis is false rather than conclude that the null is not false and that a highly unlikely outcome (under the null) has occurred.

Let's translate these considerations into a decision rule:

A p-value smaller than α is a reflection of the test statistic value v falling into the rejection region for H_0 for the given significance level α. Thus, the null hypothesis is rejected.

This is true for both left- and right-sided tests, as we did not specify how p was computed. In our example, it's $p = P(V > v|\theta_0)$, but for a left-sided test it would be $p = P(V < v|\theta_0)$. Figure 9.8 shows the right-sided test case.

Not Rejecting the Null Hypothesis

If the parameter estimate value is close to the hypothetical parameter value θ_0, then the validity of the null hypothesis appears "relatively" plausible. The probability of the estimate assuming values greater than v, is relatively high. In other words, θ_0 and the estimate are close enough to interpret their distance as the result of the noise created by the sampling process. Consequently, H_0 won't be rejected. Hence the following decision rule:

For $p > \alpha$ the test statistic realization v is an element of the non-rejection region for H_0, and the null hypothesis isn't rejected. Once again this rule holds for all single- and two-sided tests, p suitably computed (Fig. 9.9).

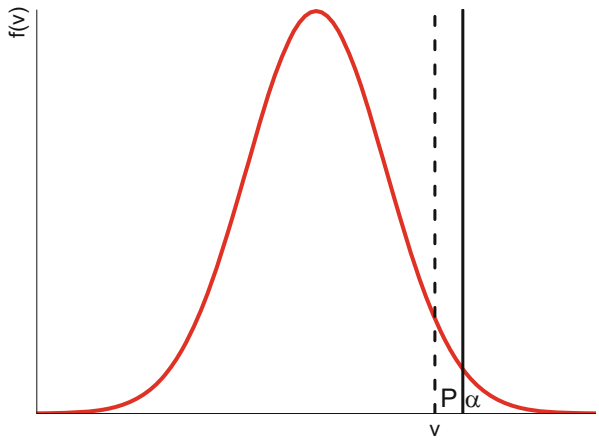


Fig. 9.9 Case of non-rejection

9.2 Testing Normal Means

In many applications one is interested in the mean of the population distribution of a particular attribute (random variable). Statistical estimation theory "tells" us how to best estimate the expectation for a given distribution shape, yet doesn't help us in assessing the uncertainty of the estimated average: an average computed from a sample of size $n = 5$ will be a single number as will be the one based on a sample size of $n = 5,000$. Intuition (and the law of large numbers) leads us to believe that the latter estimate is "probably" more representative than the former in that on the average the sample mean (e.g., the arithmetic mean) of large samples is closer to the population than that of small samples. That is, sample means computed from large samples are statistically more reliable. A method of quantifying the average closeness to the population parameter is to compute the standard error of the statistic under consideration (here: the mean), i.e., the square root of the estimated average squared deviation of the estimator from the population parameter. The actual sample mean for a given sample in conjunction with its standard deviation would specify an interval (i.e., the sample mean plus/minus one or more standard errors) in which the sample mean isn't "unlikely" to fall into, given the theoretical mean equals the one estimated from the observed sample. Now suppose a scientist proposes a value for the theoretical mean derived from some theory or prior data analysis. If the hypothetical value turns out to be close to the sample mean and, in particular, within a certain range around the sample mean like the one specified by the standard error, he is more likely to propose it to be the true population mean then if he had initially proposed a more distant value. But how can the distance of the sample mean from the hypothetical population mean be assessed in probabilistic terms suitable

for decision making based on the α error concept? In other words: How can we construct a statistical test for the mean of a random variable?

Our goal is to test for a specific value of the expectation $\mu = \mathrm{E}(X)$ of a population distribution. Our data are a randomly drawn sample of size n, theoretically represented by the sample variables $X_1, \ldots, X_n$, and we want to base the test decision at a significance level of α.

Hypotheses

We can construct one- and two-sided tests.

1. *Two-sided test*

$$\mathrm{H}_0 : \mu = \mu_0, \quad \mathrm{H}_1 : \mu \neq \mu_0.$$

2. *Right-sided test*

$$\mathrm{H}_0 : \mu \leq \mu_0, \quad \mathrm{H}_1 : \mu > \mu_0.$$

3. *Left-sided test*

$$\mathrm{H}_0 : \mu \geq \mu_0, \quad \mathrm{H}_1 : \mu < \mu_0.$$

In a one-sided statistical hypothesis testing problem the scientific conjecture to be validated is usually stated as alternative hypothesis H_1 rather than the null hypothesis H_0. That is, the researcher tries to statistically verify that the negation of the hypothesis to be tested does *not* hold for a certain significance level α. This is due to the "nature" of the significance level we have mentioned earlier: Rejecting the null hypothesis at a given significance level only means that the probability of it not being false is no greater than α. Yet, it is chosen to be small (most commonly 0.05 or 0.01), as one tries to control the α error in order to be "reasonably certain" that an "unwanted" proposition is *not* true. This makes sense if one thinks of some critical applications that rely on this approach. In testing a new drug for harmful side effects, for example, one wants to have a rational for rejecting their systematic occurrence. In doing so one accepts the converse claim that side effects are 'negligible'. Underlying this approach is the (unknown) relationship between α and β: Whereas we can control the former, the latter is a function of not only the former but also other test conditions such as the underlying distribution.

For these reasons it is common to speak of *not rejecting* a hypothesis instead of *accepting* it.

Test Statistic, Its Distribution, and Derived Decision Regions

We need a quantity to condense the information in the random sample that is required to make a probabilistic statement about the unknown distribution characteristic (in the present case the population mean). For parametric tests, this is an estimator of the parameter. We have already shown that the arithmetic mean

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is a statistically "reasonable" point estimator of the unknown population mean, i.e., the unknown expectation $E(X)$, in particular it's unbiased and consistent. The variance and standard deviation of $\overline{X}$ computed from a random sample (i.e., independent and identically distributed—i.i.d.) are given by

$$Var\left(\overline{X}\right) = \sigma^2\left(\overline{X}\right) = \sigma_{\overline{X}}^2 = \frac{\sigma_X^2}{n}$$

$$\sigma\left(\overline{X}\right) = \frac{\sigma_X}{\sqrt{n}}$$

We will construct our test statistic around the sample mean $\overline{X}$. In order to derive the (rejection/non-rejection) regions corresponding to a given significance level, we need to make an assumption concerning the distribution of the sample mean. Either

- The random variable under investigation X is normally distributed, implying normal distribution of $\overline{X}$; *or*
- n is sufficiently large to justify the application of the Central Limit Theorem: If the sample variables X_i are i.i.d. with finite mean and variance, $\overline{X}$ is approximately normally distributed regardless of the underlying (continuous or discrete, symmetric or skewed) distribution. In this case, our test will in turn be an approximate one, i.e., has additional imprecision.

We thus postulate:

$\overline{X}$ is (at least approximately) normally distributed with expectation $E\left(\overline{X}\right) = \mu$ and variance $Var\left(\overline{X}\right) = \sigma_X^2/n$.

Thus, the distribution of the estimator of the population mean μ depends on exactly the unknown parameter we are seeking to test μ. The only way to overcome this circular reference is to assign a numerical value to μ. The least arbitrary value to take is the boundary value in the null hypothesis, i.e., the value that separates the parameter ranges for H_0 and H_1: μ_0. This approach does in fact make sense, if you recall the principle of rejecting the null hypothesis in order to not-reject the alternative hypothesis: Basing the decision on a postulated distribution of our test statistic with parameter μ_0 enables us to test this particular μ, by removing the

uncertainty in the distribution function. Note that in the two-sided test this μ_0 makes up the entire parameter space of the null hypothesis. In one-sided tests, it is the boundary value.

Let's put our assumption into practice and set the expectation of X, i.e., μ, to μ_0: Given the null hypothesis $\mathrm{H}_0 : \mu = \mu_0$ is true, respectively μ equals the boundary value of the null hypothesis for single-sided test, we can write $\overline{X}$ is (at least approximately) normally distributed with expectation $E\left(\overline{X}\right) = \mu_0$ and variance $Var\left(\overline{X}\right) = \sigma_X^2/n$, or, using common notation for normal distribution functions:

$$\overline{X} \overset{\mathrm{H}_0}{\sim} N\left(\mu_0;\, \sigma/\sqrt{n}\right).$$

So far, we have focused on the location parameter μ. But what about the second central moment that specifies a particular normal distribution, the variance (respectively standard deviation) of the random variable? As you will see, it is critical to the construction of a decision rule to distinguish between situations in which we can regard σ as known and those where we can't.

Known σ: Given a known σ, the distribution of $\overline{X}$ is completely specified. As we cannot analytically integrate the normal density function to get a closed form normal distribution function, we rely on tables of numerical solutions for $N(\mu = 0,\, \sigma = 1)$. We thus standardize $\overline{X}$ and take

$$V = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n}$$

as our test statistic.

Given H_0 is true, V (approximately) follows a standard normal distribution:

$$V \overset{\mathrm{H}_0}{\sim} N(0,\, 1).$$

The critical value corresponding to the relevant significance level α can thus be taken from a standard normal distribution table.

We can now write down the decision regions for the three types of test for significance level α, given the boundary expectation from H_0, i.e., μ_0, is the true population mean.

1. *Two-sided test*

   The probability of V falling into the rejection region for H_0 must equal the given significance level α:

   $$P(V < c_l|\mu_0) + P(V > c_u|\mu_0) = \alpha/2 + \alpha/2 = \alpha.$$

   For $P(V \leq c_u) = 1 - \alpha/2$ we can retrieve the upper critical value from the cumulative standard normal distribution table $N(0,\, 1)$: $c_u = z_{1-\alpha/2}$. Symmetry of the normal (bell) curve implies $c_l = -z_{1-\alpha/2}$.

The *rejection region* for H_0 is thus given by

$$\left\{v|v < -z_{1-\alpha/2} \text{ or } v > z_{1-\alpha/2}\right\}.$$

The *non-rejection region* for H_0 is then

$$\left\{v| - z_{1-\alpha/2} \leq v \leq z_{1-\alpha/2}\right\}.$$

The probability of V assuming a value from the non-rejection region for H_0 is

$$P\left(c_l \leq V \leq c_u|\mu_0\right) = P\left(-z_{1-\alpha/2} \leq V \leq z_{1-\alpha/2}|\mu_0\right) = 1-\alpha$$

2. *Right-sided test*

Deviations of the standardized test statistic V from $E(V) = 0$ to the "right side" (i.e., positive $(V > 0)$) tend to falsify H_0. The rejection region will thus be a range of positive test statistic realizations v (i.e., a positive critical value). The probability of observing realization of V within this region must equal the given significance level α:

$$P\left(V > c|\mu_0\right) = \alpha.$$

For $P(V \leq c) = 1-\alpha$ we find the critical value in the table for the cumulative standard normal distribution $N(0,\ 1)$: $c = z_{1-\alpha}$.

The rejection region for H_0 is given by

$$\left\{v|v > z_{1-\alpha}\right\},$$

and the non-rejection region for H_0 is

$$\left\{v|v \leq z_{1-\alpha}\right\}.$$

The probability of V assuming a value within the non-rejection region for H_0 is

$$P\left(V \leq c \,|\, \mu_0\right) = P\left(V \leq z_{1-\alpha} \,|\, \mu_0\right) = 1-\alpha$$

3. *Left-sided test*

Sample means smaller than μ_0 imply negative realizations of the test statistic V, that is, deviations of V from $E(V) = 0$ to the left side on the real line. In this case, the rejection region for H_0 therefore consists of negative V outcomes. Consequently, the critical value c will be negative.

Once again, we require the probability of observing realization of V within the rejection region to equal α:

$$P\left(V < -c|\mu_0\right) = \alpha.$$

Using the symmetry property of the normal distribution, we can translate $P(V < -c)$ into $1 - P(V < c)$. Thus, the absolute value of the critical value, $|-c| = c$, is the value of the cumulative normal distribution function for probability $(1-\alpha)$, i.e., $c = z_{1-\alpha}$, and $-c = -z_{1-\alpha}$

The rejection region for H_0 is given by

$$\{v | v < -z_{1-\alpha}\},$$

and the non-rejection region for H_0 is

$$\{v | v \geq -z_{1-\alpha}\}.$$

The probability of V taking on a value within the non-rejection region for H_0 is

$$P(V \geq -c \,|\, \mu_0) = P(V \geq -z_{1-\alpha} \,|\, \mu_0) = 1 - \alpha.$$

Unknown σ If we don't have any a priori knowledge about the standard deviation of the random variable under investigation, we need to plug an estimator of it into the test statistic

$$V = \frac{\overline{X} - \mu_0}{\sigma} \sqrt{n}.$$

An unbiased estimator of the population variance is

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}.$$

Replacing σ by the square root of S^2 yields our new test statistic:

$$T = \frac{\overline{X} - \mu_0}{S} \sqrt{n}.$$

If the null hypothesis H_0 is true, T has (at least approximately) a t distribution with $n-1$ degrees of freedom.

For a given significance level α and $n-1$ degrees of freedom, the critical values can be read from the t distribution table.

If we denote the p-quantile of the t distribution with $n-1$ degrees of freedom by $t_{p;n-1}$, and assume μ_0 is the true population mean, we have the following decision regions for the test situations under consideration.

1. *Two-sided test*
   rejection region for H_0:

$$\left\{t | t < -t_{1-\alpha/2;n-1} \text{ or } t > t_{1-\alpha/2;n-1}\right\},$$

   where t is a realization of the random variable T computed from an observed sample of size n.
   Non-rejection region for H_0:

$$\left\{t | -t_{1-\alpha/2;n-1} \leq t \leq t_{1-\alpha/2;n-1}\right\}.$$

2. *Right-sided test*
   rejection region for H_0:

$$\{t | t > t_{1-\alpha;n-1}\}.$$

   Non-rejection region for H_0:

$$\{t | t \leq t_{1-\alpha;n-1}\}.$$

3. *Left-sided test*
   rejection region for H_0:

$$\{t | t < t_{1-\alpha;n-1}\}.$$

   Non-rejection region for H_0:

$$\{t | t \geq t_{1-\alpha;n-1}\}.$$

Note: If the sample size is sufficiently large ($n > 30$), the t distribution can be adequately approximated by the standard normal distribution. That is, T is approximately $N(0; 1)$ distributed. Critical values can then be read from the normal table, and the decision regions equal those derived for known population standard deviation σ. Hence, for large n we can estimate σ by S and abstract from the estimation error (that will occur with probability one, even if the estimator hits the correct parameter on average, i.e., is unbiased).

Calculating the Test Statistic from an Observed Sample

When we have obtained a random sample $x_1, \ldots, x_n$, we can compute the empirical counterparts of the theoretical test statistics we have based our test procedures on. On the theoretical level, we have expressed them in terms of (theoretical) sample

variables, i.e., $X_1, \ldots, X_n$, that is, have denoted them by capital letters: $\overline{X}$, V and S. Actual values calculated from a sample of size n, $x_1, \ldots, x_n$, are denoted by $\bar{x}$, v and s and differ from their theoretical counterparts only in that now the variables stand for real numbers rather than a range of theoretically permissible values. Hence, the respective empirical formulae for sample mean and sample standard deviation are

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and

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

Accordingly, the two realized test statistics for testing normal means for known and unknown variance respectively are

$$v = \frac{\bar{x} - \mu_0}{\sigma} \sqrt{n}$$

and

$$t = \frac{\bar{x} - \mu_0}{s} \sqrt{n}.$$

You may have recognized that we have already applied this notation when specifying the decision regions.

Test Decision and Interpretation

If the test statistic v falls into the rejection region, the null hypothesis H_0 is rejected on the basis of a random sample of size n and a given a significance level α: 'H'_1. Statistically, we have concluded that the true expectation $E(X) = \mu$ does not equal the hypothetical μ_0.

If the *true* parameter *does* belong to the range postulated in the null hypothesis (H_0), we have made a type I error: 'H_1'$|H_0$. In fact, in choosing a particular significance level, we are really deciding about the probability of making exactly this error, since the decision regions are constructed such that the probability of making a type I error equals the significance level: $P(\text{'}H_1\text{'}|H_0) = \alpha$.

If, on the other hand, v falls into the non-rejection region, the particular sample leads to a non-rejection of the null hypothesis for the given significance level: 'H'_0. Thus, we are not able to show statistically, that the true parameter $E(X) = \mu$ deviates from the hypothetical one (μ_0). Chances are, though, nontrivial that we

are making a type II error, i.e., the alternative hypothesis correctly describes reality: 'H_0'|H_1. As already pointed out, the probability of making a β error is, in general, unknown and has to be computed for individual alternative parameter values μ_1.

Power

How can we assess the "goodness" of a test? We have seen that in setting up a test procedure we are controlling the probability of making an α error (by assigning a value to the significance level α). The probability of making a β error is then determined by the true (and unknown) parameter. The smaller β is for a given true parameter μ, the more reliable the test is in that it more frequently rejects the null hypothesis when the alternative hypothesis is really true. Hence, given a specific significance level, we want β to be as small as possible for true parameter ranges outside that specified in the null hypothesis, or, equivalently, we want to maximize the probability of making the correct decision ('H_1'|H_1), that is maximize the quantity $(1-\beta)$ for any given true μ outside the null hypothesis region, i.e., inside that of the alternative hypothesis.

This notion of "goodness" of a test is conceptualized with the so-called *power*, a function assigning probabilities of rejecting H_0 $(1-\beta)$ to true parameter values μ within the H_1 parameter region for given α and hypothetical parameter μ_1. These probabilities represent the theoretical averages of making a right decision in rejecting H_0 over all possible samples (given α and μ). They can thus be computed without utilizing actual samples; in fact, the power is computed because we can obtain only a limited sample and aim to quantify the expected "accuracy" of the individual test procedure.

Technically, the power $P(\mu)$ yields the probability of rejecting H_0 given hypothetical parameters μ:

$$P(\mu) = P(V \in \text{rejection region for } H_0|\mu) = P('H_1'|\mu)$$

1. *Two-sided test*

   In a two-sided test, the null hypothesis is true if and only if $\mu = \mu_0$. Rejecting H_0 given that it is true means we have made a type I error:

   $$P(V \in \text{rejection region for } H_0|\mu = \mu_0) = P('H_1'|H_0) = \alpha.$$

   For all other possible parameter values, rejecting H_0 is a right decision:

   $$P(V \in \text{rejection region for } H_0|\mu \neq \mu_0) = P('H_1'|H_1) = 1-\beta.$$

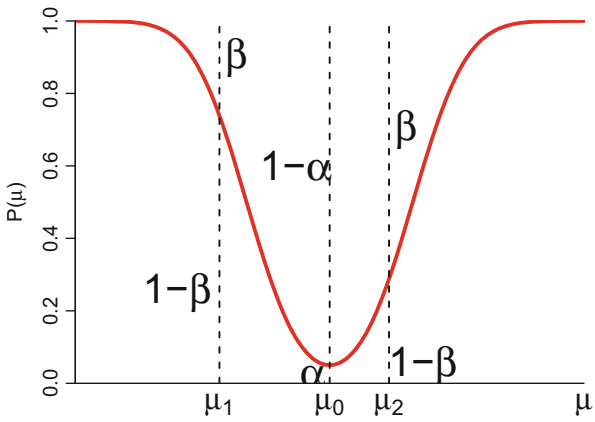


Fig. 9.10 Power function in a two-sided test

We thus have

$$P(\mu) = \begin{cases} P(\text{`H}_1\text{'}|\text{H}_0) = \alpha, & \text{if } \mu = \mu_0; \\ P(\text{`H}_1\text{'}|\text{H}_1) = 1 - \beta, & \text{if } \mu \neq \mu_0. \end{cases}$$

Using our normality assumption about the underlying probability distribution, we can analytically calculate the power for the case of a two-sided test:

$$P(\mu) = 1 - \left[P\left(V \leq z_{1-\alpha/2} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}} \right) - P\left(V \leq -z_{1-\alpha/2} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}} \right) \right].$$

The probability of a type II error can be calculated from the power:

$$P(\text{`H}_0\text{'}|\text{H}_1) = 1 - P(\mu \neq \mu_0) = \beta.$$

Properties of the power for a two-sided test:

- For $\mu = \mu_0$, the power assumes its minimum, α.
- The power is symmetrical around the hypothetical parameter value μ_0
- The power increases with growing distance of the true parameter μ from the hypothetical μ_0 and converges to one as the distance increases to ∞ or $-\infty$ respectively.

The above characteristics are illustrated in the following power curve diagram.

In Fig. 9.10, two alternative true parameter values μ_1 and μ_2 are depicted. If μ_1 is the true parameter, the distance $\mu_1 - \mu_0$ is comparatively high. Consequently, the probability $1 - \beta$ of making a right decision in not-rejecting the alternative hypothesis H_1 (conversely, correctly rejecting the null) is relatively high and the probability of making a type II error, β, small.

The distance of the "hypothetically true" parameter value μ_2 from the hypothetical parameter value μ, $\mu_2 - \mu_0$, is relatively small. Hence, the probability of making a right decision in rejecting the null hypothesis, $1 - \beta$, is smaller than in the first example, and the probability of making a type II error, β, greater. This is intuitively plausible, i.e., that relatively small deviations are less easily discovered by the test.

2. *Right-sided test*

   In a right-sided test, the null hypothesis is true if the true parameter is less than or equal to the hypothetical boundary value μ_0, i.e., if $\mu \leq \mu_0$. If this is the case, the maximum probability of rejecting the null hypothesis and hence making a type I error, equals the significance level α:

$$P\,(V \in \text{rejection region for } \mathrm{H}_0 | \mu \leq \mu_0) = P\,(\text{`}\mathrm{H}_1\text{'} | \mathrm{H}_0) \leq \alpha.$$

If the alternative hypothesis, i.e., $\mu > \mu_0$, is true, rejecting the null hypothesis and hence making a right decision occurs with probability:

$$P\,(V \in \text{rejection region for } \mathrm{H}_0 | \mu \geq \mu_0) = P\,(\text{`}\mathrm{H}_1\text{'} | \mathrm{H}_1) = 1 - \beta.$$

Combining these formulae for the two disjoint subsets of the parameter space gives the power:

$$P\,(\mu) = \begin{cases} P\,(\text{`}\mathrm{H}_1\text{'} | \mathrm{H}_0) \leq \alpha, & \text{if } \mu \leq \mu_0; \\ P\,(\text{`}\mathrm{H}_1\text{'} | \mathrm{H}_1) = 1 - \beta, & \text{if } \mu > \mu_0. \end{cases}$$

We can explicitly calculate the power for our right-sided test problem for all possible true parameter values μ:

$$P\,(\mu) = 1 - P\left(V \leq z_{1-\alpha} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}} \right).$$

Figure 9.11 displays the typical shape of the power for a right-sided test problem. For all values within the parameter set of the alternative hypothesis, the power increases monotonically to one. The greater the distance $\mu - \mu_0$, the higher the probability $1 - \beta$ of making a right decision in not-rejecting the alternative hypothesis, and hence the smaller the probability β of making a type II error. At the point $\mu = \mu_0$ the power is α, the given significance level. For all other values associated with the null hypothesis, i.e., $\mu < \mu_0$, the power is less than α. That's what we assumed when we constructed the test: We want α to be the *maximum* probability of rejecting the null hypothesis for a true null hypothesis. As you can see from the graph, this probability decreases with rising absolute distance $\mu - \mu_0$.

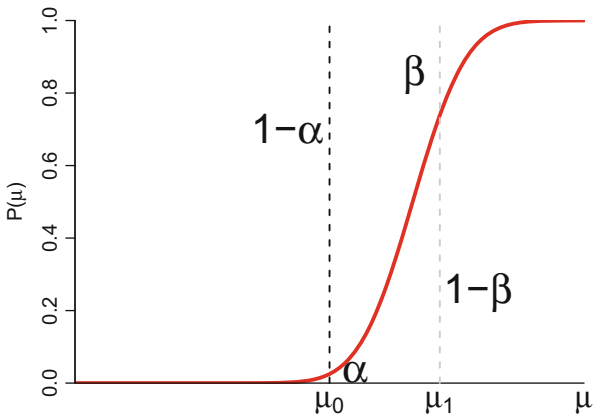


Fig. 9.11 Power function in a right-sided test

3. Left-sided test

In a left-sided test, the null hypothesis is true if the true parameter is greater than or equal to the hypothetical boundary value, that is, if $\mu \geq \mu_0$. In this case, rejecting the null hypothesis and hence making a type I error, will happen with probability of no more than α:

$$P\,(V \in \text{rejection region for } \mathrm{H}_0 | \mu \geq \mu_0) = P\,(\text{`}\mathrm{H}_1\text{'} | \mathrm{H}_0) \leq \alpha.$$

If the alternative hypothesis is true, i.e., $\mu < \mu_0$, the researcher makes a right decision in rejecting the null hypothesis, the odds being:

$$P\,(V \in \text{rejection region for } \mathrm{H}_0 | \mu \leq \mu_0) = P\,(\text{`}\mathrm{H}_1\text{'} | \mathrm{H}_1) = 1 - \beta.$$

For the entire parameter space we thus have:

$$P\,(\mu) = \begin{cases} P\,(\text{`}\mathrm{H}_1\text{'} | \mathrm{H}_0) \leq \alpha, & \text{if } \mu \geq \mu_0; \\ P\,(\text{`}\mathrm{H}_1\text{'} | \mathrm{H}_1) = 1 - \beta, & \text{if } \mu < \mu_0. \end{cases}$$

For our normally distributed population we can calculate the probability of rejecting H_0 as a function of the true parameter value μ (the power) explicitly:

$$P\,(\mu) = P\left(V \leq -z_{1-\alpha} - \frac{\mu - \mu_0}{\sigma / \sqrt{n}} \right).$$

A typical graph of a power for a left-sided test is depicted in Fig. 9.12. The graph is interpreted similar to the right-sided test case.

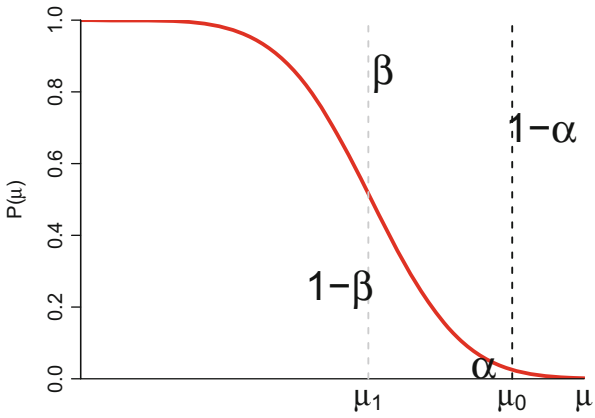


Fig. 9.12 Power function in a left-sided test

More Information: Conducting a Statistical Test

Formulating the Hypotheses

Let us illustrate the problem of choosing an appropriate null (and hence alternative) hypothesis with a real-world example.

Consider a company manufacturing car tires. Alterations in the production process are undertaken in order to increase the tires' lives. Yet competitors will not hesitate to claim that the average life of the tires hasn't increased from the initial, pre-restructuring value of 38,000 kilometers (km). The producers' management wants to justify the investment into the new production process and subsequent advertising campaign (i.e., save their necks) and commissions a scientific, i.e., statistical, investigation.

That's our part. The variable of interest is the life of an individual tire measured in km, denoted by, say, X. It is a random variable, because its fluctuations in magnitude depend on many unknown and known factors, that cannot practically be taken into account (such as speed, weight of the individual car, driving patterns, weather conditions, and even slight variations in the production process, etc.). Before the "improvements" in the production process, the average life of the particular type of car tire was 38,000 km; in theoretical terms, the expectation was $E(X) = \mu_0 = 38,000$ km. The mean value under the new production process is unknown and, in fact, the quantity we want to compare in statistical terms with μ_0: The producer pays the statistician(s) to objectively show, if $\mu > \mu_0 = 38,000$ km. Note that we denote the true expectation under the new regime by μ, as this is the parameter we are interested in and thus want to test. The "old" mean μ_0 "merely" serves as benchmark, and the actual output it represents (the old tires) doesn't receive further attention (and in particular neither does its fluctuations around the mean).

The statement that management hopes that the statistician will "prove" scientifically, $\mu > \mu_0$, looks very much like a readily testable null hypothesis. But as we have emphasized earlier, there is a crucial difference between formalized statements of scientific interest and the means of testing it by stating a null hypothesis suitable to make a reliable decision, that is, a decision that is backed by acceptable type I and II errors.

So which hypothesis shall we test? It should be clear, that the problem at hand is a single-sided one; only deviations of the new expected life from the historical expected life in one direction are of interest. In deciding whether to test the hypothesis as it is already formalized using a left-sided test procedure or testing the negation, $\mu \leq \mu_0$, on a right-sided basis, we have to focus on the actual aim of the investigation: The tire producer intends to verify the claim of μ being greater than μ_0, whilst at the same time controlling the risk of making a wrong decision (type I error) to a level that allows him to regard the (hopefully positive, i.e., a rejection of the null) test decision as statistically proven. This would be the case if the reverse claim of the new tires being less durable can be rejected with an acceptable (i.e., small) significance level, for this would imply that there is only a small probability that the null hypothesis, $\mu \leq \mu_0$, is true and hence the alternative hypothesis, $\mu > \mu_0$, not true. But that's exactly the result the managers want to see. Let's therefore state the negation of the statement to be tested as null hypothesis (and hope it will be rejected on the given significance level):

$$\mathrm{H}_0 : \mu \leq \mu_0 \quad \text{versus} \quad \mathrm{H}_1 : \mu > \mu_0,$$

with $\mu_0 = 38,000$ km.

If the sample of n new tires' usable life leads to a rejection of the null hypothesis H_0 ('H_1'), a type I error will be made if the null hypothesis is true. If the null hypothesis is *not* rejected on the basis of a particular sample of size n, the conjecture stated in the alternative hypothesis may still be true, in which case, the researcher has (unknowingly) made a type II error.

Comparing the implications of type I and type II error for this example shows that the former's impact on the manufacturers fortune is the crucial one, for

- the competitors can carry out (more or less) similar investigations using a left-sided test, leading to the PR nightmare associated with a possible contradiction of the producers' test result,
- future investigation into tires subsequently produced would reveal the actual properties of the tires as the sample size inevitably increases with the amount sold, triggering even more embarrassing questions concerning the integrity and reliability of the manufacturer.

For these reasons, the tire manufacturer is best advised to keep the probability of a type I error, $P(\text{'}\mathrm{H}_1\text{'}|\mathrm{H}_0)$, small, by controlling the significance level, e.g., setting it to $\alpha = 0.05$.

Decision Regions

When testing μ with either single- or two-sided tests the size of the non-rejection and rejection regions on the V or T (standardized test statistic) axis depends only on:

- the given (chosen) level of significance α: ceteris paribus, increasing α will increase the size of the rejection region for H_0, and will reduce the size of the non-rejection region (and vice versa).

Alternatively, when testing μ with either single- or two-sided tests the size of the non-rejection and rejection regions on the X (our original random variable) axis depends on:

- the given (chosen) level of significance α: ceteris paribus, increasing α will increase the size of the rejection region for H_0, and will reduce the size of the non-rejection region (and vice versa);
- the sample size n: ceteris paribus, the larger the sample size, the greater the size of the rejection region for H_0, and the smaller the size of the non-rejection region (and vice versa); and
- the dispersion σ of the variable in the population and therefore S in the sample: ceteris paribus, an increased variability σ or S leads to a decrease in the size of the rejection region for H_0, and increases the size of the non-rejection region (and vice versa).

That is, the critical values on the standardized test statistic axis are independent of the size of n or σ (alternatively, S). The same cannot be said for the "equivalent" critical values for the original X axis where sample size and dispersion affect the magnitude of "acceptable" expected deviations from the null.

If the population variance σ is known, the critical values and therefore the non-rejection/rejection regions for H_0 can easily be calculated for the sample mean $\overline{X}$. We will do this for a two-sided test.

We have derived the test statistic V as a standardization of the estimator $\overline{X}$:

$$V = \frac{\overline{X} - \mu_0}{\sigma} \sqrt{n},$$

and, in terms of realizations x_i's of sample variables X_i's:

$$v = \frac{\overline{x} - \mu_0}{\sigma} \sqrt{n}.$$

In a two-sided test the non-rejection region for H_0 consists of all realization v of V greater than or equal to $-z_{1-\alpha/2}$ *and* less than or equal to $z_{1-\alpha/2}$:

$$\left\{ v | -z_{1-\alpha/2} \leq v \leq z_{1-\alpha/2} \right\}.$$

Thus, the critical values $-z_{1-\alpha/2}$ and $z_{1-\alpha/2}$ are possible realization of the test statistic V. They are subject to the same standardization carried out to convert $\overline{X}$ into V to express it in units comparable with standard normal quantiles:

$$-z_{1-\alpha/2} = \frac{\overline{X}_l - \mu_0}{\sigma}\sqrt{n}\,, \quad z_{1-\alpha/2} = \frac{\overline{X}_u - \mu_0}{\sigma}\sqrt{n}.$$

As $-z_{1-\alpha/2}$ is the lower critical value with respect to V, we similarly have denoted the lower critical value for $\overline{X}$ by $\overline{X}_l$ (the same applies to the upper bound of the non-rejection region, denoted by the subindex u).

We can isolate the upper and lower bound of the rejection region for H_0 in terms of the units of the sample mean:

$$\overline{X}_l = \mu_0 - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\,, \quad \overline{X}_u = \mu_0 + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

The resulting non-rejection region for H_0 in terms of $\overline{X}$ is:

$$\left\{\overline{X} \,|\, \overline{X}_l \le \overline{X} \le \overline{X}_u\right\} = \left\{\overline{X} \,|\, \mu_0 - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \le \overline{X} \le \mu_0 + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right\},$$

and the associated rejection region is given by the complement

$$\left\{\overline{X} \,|\, \overline{X} < \overline{X}_l \text{ or } \overline{X} > \overline{X}_u\right\} =$$

$$\left\{\overline{X} \,|\, \overline{X} > \mu_0 - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \text{ or } \overline{X} > \mu_0 + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right\}.$$

Similar transformations can be imposed on the estimators for one-sided tests.

Power Curve

We will derive the power curve for a two-sided population mean test. The power is calculated as

$$\begin{aligned} P(\mu) &= P(V \in \text{rejection region for } \mathrm{H}_0 \,|\, \mu) \\ &= 1 - P(V \in \text{non-rejection region for } \mathrm{H}_0 \,|\, \mu)\,. \end{aligned}$$

Assuming μ to be the true population mean, we have

$$\begin{aligned} P(\mu) &= 1 - P\left(-z_{1-\alpha/2} \le V \le z_{1-\alpha/2} \,|\, \mu\right) \\ &= 1 - P\left(-z_{1-\alpha/2} \le \frac{\overline{X} - \mu_0}{\sigma\sqrt{n}} \le z_{1-\alpha/2} \,|\, \mu\right). \end{aligned}$$

Adding $\mu - \mu$ to the numerator of the middle term yields

$$
\begin{aligned}
&P(\mu) = \\
&= 1 - P\left(-z_{1-\alpha/2} \leq \frac{\overline{X} - \mu_0 + \mu - \mu}{\sigma\sqrt{n}} \leq z_{1-\alpha/2} \mid \mu\right) \\
&= 1 - P\left(-z_{1-\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma\sqrt{n}} + \frac{\mu - \mu_0}{\sigma\sqrt{n}} \leq z_{1-\alpha/2} \mid \mu\right) \\
&= 1 - P\left(-z_{1-\alpha/2} - \frac{\mu - \mu_0}{\sigma\sqrt{n}} \leq \frac{\overline{X} - \mu}{\sigma\sqrt{n}} \leq z_{1-\alpha/2} - \frac{\mu - \mu_0}{\sigma\sqrt{n}} \mid \mu\right) \\
&= 1 - P\left(-z_{1-\alpha/2} - \frac{\mu - \mu_0}{\sigma\sqrt{n}} \leq V \leq z_{1-\alpha/2} - \frac{\mu - \mu_0}{\sigma\sqrt{n}} \mid \mu\right) \\
&= 1 - \left[P\left(V \leq z_{1-\alpha/2} - \frac{\mu - \mu_0}{\sigma\sqrt{n}} \mid \mu\right) - P\left(V \leq -z_{1-\alpha/2} - \frac{\mu - \mu_0}{\sigma\sqrt{n}} \mid \mu\right)\right].
\end{aligned}
$$

The power for the one-sided tests can be derived in a similar fashion.

From a decision-theoretical point of view it is desirable that the probability of correctly rejecting the null hypothesis increases quickly with a growing distance between the true parameter μ and the hypothetical value μ_0, that is, we want the graph of the power curve to be as steep as possible in that range of the true parameter value. For a given estimator and test statistic, there are two possible ways of improving the "shape" of the power curve.

1. *Increasing the sample size n*

   The above formula for the power of a two-sided test for the mean is clearly positively related to the size of the sample n. In general, ceteris paribus, the graph of the power curve becomes steeper with growing n: For any true parameter value within the H_1 region (i.e., $\mu \neq \mu_0$ for the two-sided, $\mu > \mu_0$ for the right-sided and $\mu < \mu_0$ for the left-sided test), the probability $1 - \beta$ of rejecting the null hypothesis, and hence making a right decision, increases with growing n. That's mirrored by a decreasing probability β of making a type II error. Thus, the probability of correctly discriminating between the true and the hypothetical parameter value grows with increasing sample size. Given a fixed significance level α, the probability of a type II error can be improved (reduced) by "simply" enlarging the sample.

   Figure 9.13 displays the graphs of 4 power curves based on four distinct sample sizes, with $n_1 < n_2 < n_3 < n_4$.

2. *Varying the significance level α*

   Ceteris paribus, allowing for a higher probability of making a type I error, i.e., increasing the significance level α, will shift the graph of the power curve

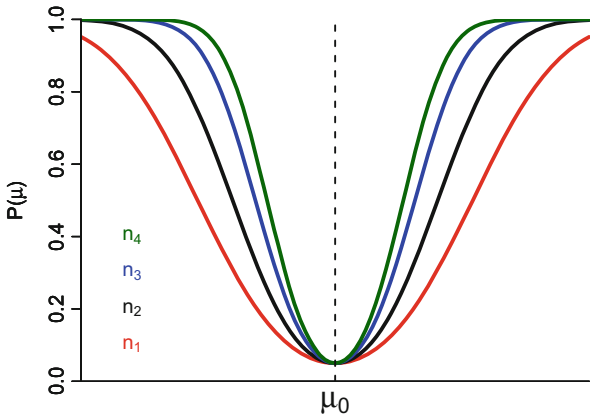


Fig. 9.13 Power of two-sided test for the population mean for alternative sample sizes

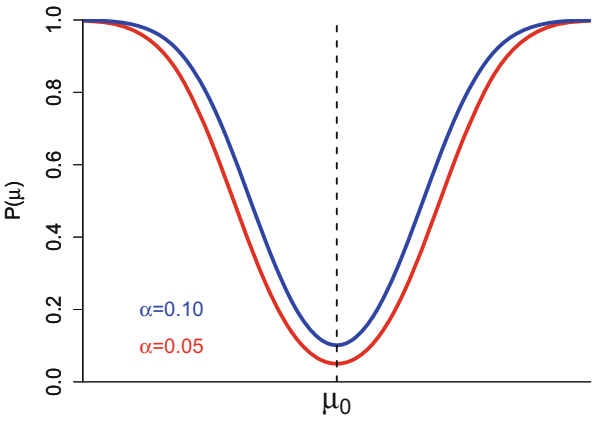


Fig. 9.14 Power of two-sided test for the population mean for alternative significance levels

upwards. This means, that a higher α leads to an increase in the probability of rejecting the null hypothesis for *all* possible true parameter values μ. If the true parameter value within the H_1 region ($\mu \neq \mu_0$ for the two-sided, $\mu > \mu_0$ for the right-sided and $\mu < \mu_0$ for the left-sided test), rejecting the null is a right decision—the probability $1 - \beta$ of correctly rejecting the null hypothesis has increased, the probability β of making a type II error has decreased. **But** the probability of rejecting the null hypothesis has also increased for true parameter values within the H_0 region, increasing the probability of making a type I error. Hence, we encounter a trade-off between the probabilities of making a type I and type II error, a problem that cannot be overcome mechanically, but has to be tackled within some sort of preference-based decision-theoretical approach.

In Fig. 9.14 the power curve of a two-sided test with fixed sample size for two alternative significance levels is depicted. The red graph represents $P(\mu)$ for $\alpha = 0.05$, the blue one $P(\mu)$ for $\alpha = 0.10$.

Explained: Testing the Population Mean

A company is packing wheat flour. The machine has been set up to fill 1,000 grams (g) into each bag. Of course, the probability of any bag containing exactly 1 kg, is zero (as weight is a continuous variable), and even if we take into account the limited precision of measurement, we will still expect some fluctuation around the desired (theoretical) content of 1 kg in actual output. But without prior knowledge we can't even be sure, if the *average* weight of output is actually 1 kg. Fortunately, we have means of testing this statistically. Denote by X the actual net weight per bag. We are interested in the expectation of this random variable, i.e., the average net bag weight, $E(X) = \mu$. Is it sufficiently close to $\mu_0 = 1$ kg, the ideal quantity we want the machine to fill into each bag? As the machine has to be readjusted from time to time to produce output statistically close enough to the required weight, the producer regularly takes samples to assess the then current precision of the packing process. If the mean of any of these samples statistically differs significantly from the hypothetical value μ_0, the machine has to be readjusted.

Hypothesis

The management is interested in deviations of the actual from the desired weight of $\mu_0 = 1$ kg in both directions. Filling in too much isn't cost-effective and putting in too little may trigger investigations from consumer organizations, with all the negative publicity that comes with it. Thus, a two-sided test is indicated:

$$\mathrm{H}_0 : \mu = \mu_0 \quad \text{versus} \quad \mathrm{H}_1 : \mu \neq \mu_0,$$

where $\mu_0 = 1,000$ g.

Sample Size and Significance Level

The statistician decides to test at a 0.05 level and asks a technician to extract a sample of $n = 25$ bags. As the population, that is, the overall production, is large compared to the sample size, the statistician can regard the sample as a simple random sample.

Test Statistic and Its Distribution: Decision Regions

The estimator of the unknown population mean $E(X) = \mu$ is the sample mean $\overline{X}$.

Experience has shown that the actual weight can be approximated sufficiently closely by a normal distributions with standard deviation $\sigma = 10$ g. The estimator $\overline{X}$ is then normally distributed with standard deviation $\sigma = 10/(25)^{1/2} = 2$ g. Under

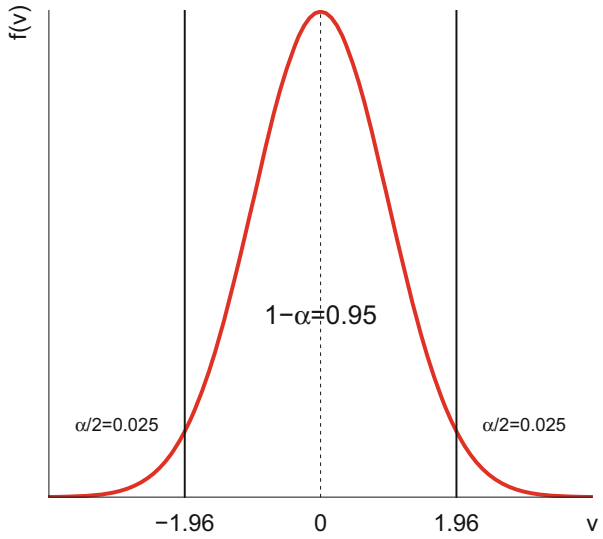


Fig. 9.15 Distribution of V under H_0 and decision regions

H_0, i.e., given, the true population parameter μ equals the hypothetical (desired) one, μ_0, $\overline{X}$ is thus normally distributed with parameters $\mu = 1,000$ g and $\sigma = 2$ g:

$$\overline{X} \overset{H_0}{\sim} N\left(1,000;\, 2\right).$$

The test statistic V is the standardization of the sample mean,

$$V = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n},$$

and follows the standard normal distribution:

$$V \overset{H_0}{\sim} N\left(0;\, 1\right).$$

We can look up the upper critical value in the cumulative standard normal distribution table as $c_u = z_{0.975} = 1.96$ to satisfy $P\left(V \leq c_u\right) = 1 - \alpha/2 = 0.975$. Using symmetry of the normal curve, $c_l = -z_{1-\alpha/2} = -1.96$.

We thus have (Fig. 9.15):

$$\text{The non-rejection region for } H_0 : \{v \mid -1.96 \leq v \leq 1.96\}, \text{ and}$$

$$\text{The rejection region for } H_0 : \{v \mid v < -1.96 \text{ or } v > 1.96\}.$$

Drawing the Sample and Calculating the Test Statistic

25 bags are selected randomly and their net content is weighed. The arithmetic mean of these measurements is $\bar{x} = 996.4$ g. The realized test statistic value is thus

$$v = \frac{996.4 - 1,000}{2} = -1.8.$$

Test Decision and Interpretation

As $v = -1.8$ lies within the non-rejection region for H_0, the hypothesis is not-rejected.

Based on a sample of size $n = 25$, the hypothetical mean value $\mu_0 = 1,000$ g couldn't be shown to differ statistically significantly from the true parameter value μ, i.e., we couldn't verify that the packing process is not precise.

Power

Not having rejected the null hypothesis, we are inevitably taking the risk of making a type II error: 'H_0' | H_1, i.e., the alternative hypothesis is true and we have rejected it. We should therefore assess the reliability of our decision in terms of type II error probabilities for parameter values different from that stated in the null hypothesis, i.e., $\mu \neq \mu_0$. They are given by $1 - P(\mu)$.

Suppose, 1,002 g is the true average weight and the alternative hypothesis therefore a true statement. As the power assigns probabilities for right decisions to alternative true parameter values, $P(1,002)$ is the probability of making a right decision (correctly rejecting the null hypothesis):

$$P(\text{'}H_1\text{'}|H_1) = 1 - \beta.$$

Plugging $\mu_0 = 1,000$, $\alpha = 0.05$, $\sigma = 10$ and $n = 25$ into the formula for the power gives

$$\begin{aligned}
&P(1,002) = \\
&= 1 - \left[P\left(V \leq 1.96 - \frac{1,002 - 1,000}{2} \right) - P\left(V \leq -1.96 - \frac{1,002 - 1,000}{2} \right) \right] \\
&= 1 - [P(V \leq 0.96) - P(V \leq -2.96)] \\
&= 1 - [P(V \leq 0.96) - (1 - P(V \leq 2.96))] \\
&= 1 - [0.831472 - (1 - 0.998462)] \\
&= 1 - 0.829934 \\
&= 0.17 = 1 - \beta.
\end{aligned}$$

The probability of making a type II error if the true population mean is $1,002$, is therefore

$$P\,(\text{`H}_0\text{'}|\text{H}_1) = \beta\,(1,002) = 1 - P\,(1,002) = 0.83.$$

There, if the true average weight is $1,002$, 83 % of all samples of size $n = 25$ would not convert that fact into a correct test decision (rejection of the null) for the given significance level of $\alpha = 0.05$. Since $1,002 - 1,000$ is only a relatively small difference, in statistical terms, the probability of a type II error is large.

If, on the other hand, 989 grams is the true average weight, $P\,(989)$ returns the probability of making a right decision in rejecting the null hypothesis: $P\left(\text{`H}_1'|\text{H}_1\right) = 1 - \beta$, and we can calculate

$$P\,(989) = 1 - \beta = 0.9998 \text{ and } \beta\,(989) = 0.0002.$$

In this case, only 0.02 % of all samples will result in a non-rejection of the null hypothesis and hence a wrong decision. The probability of a type II error is small, because the difference $989 - 1,000$ is large in statistical terms.

Table 9.3 lists values of $P\,(\mu)$ and $1 - P\,(\mu)$ for selected true population averages μ, given the above μ_0, α and σ. Figure 9.16 shows the graph of the power curve.

We can alter the shape of the power curve for a (given) fixed significance level α in our favor by increasing the sample size n. We will illustrate the effect of a change in the sample size for the two "hypothetically" true parameter values $1,002$ and 989. The other test parameters remain constant: $\mu_0 = 1,000$, $\alpha = 0.05$ and $\sigma = 10$.

Figure 9.17 displays the power of the two-sided test for these 4 alternative sample sizes.

When there is reason to believe that the machine produces output with small deviations from the desired weight, an increase of the significance level is advisable to statistically "discover" these deviations reliably and minimize the type II error risk—given the incurred extra sampling costs are outweighed by the information gain.

Table 9.3 Values of power function

μ	True hypothesis	$P\,(\mu)$	$1 - P\,(\mu)$
988.00	H_1	$0.999973 = 1 - \beta$	$0.000027 = \beta$
990.40	H_1	$0.997744 = 1 - \beta$	$0.002256 = \beta$
992.80	H_1	$0.949497 = 1 - \beta$	$0.050503 = \beta$
995.20	H_1	$0.670038 = 1 - \beta$	$0.329962 = \beta$
997.60	H_1	$0.224416 = 1 - \beta$	$0.775584 = \beta$
1,000.00	H_0	$0.05 = \alpha$	$0.95 = 1 - \alpha$
1,002.40	H_1	$0.224416 = 1 - \beta$	$0.775584 = \beta$
1,004.80	H_1	$0.670038 = 1 - \beta$	$0.329962 = \beta$
1,007.20	H_1	$0.949497 = 1 - \beta$	$0.050503 = \beta$
1,009.60	H_1	$0.997744 = 1 - \beta$	$0.002256 = \beta$
1,012.00	H_1	$0.999973 = 1 - \beta$	$0.000027 = \beta$

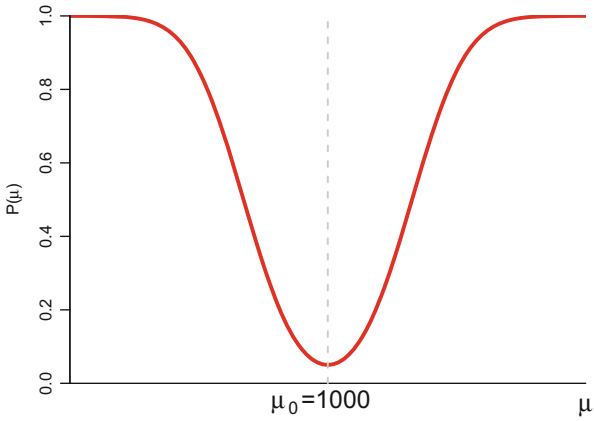


Fig. 9.16 Power of two-sided test for the population mean with $\mu_0 = 1000$, $\alpha = 0.05$, $\sigma = 10$ and $n = 25$

	$n = 9$	$n = 16$	$n = 25$	$n = 36$
$P(1,002) = 1 - \beta$	0.0921	0.126	0.17	0.224
$\beta(1,002)$	0.9079	0.8740	0.8300	0.776000
$P(989) = 1 - \beta$	0.9100	0.9930	0.9998	0.999998
$\beta(989)$	0.0900	0.0070	0.0002	0.000002

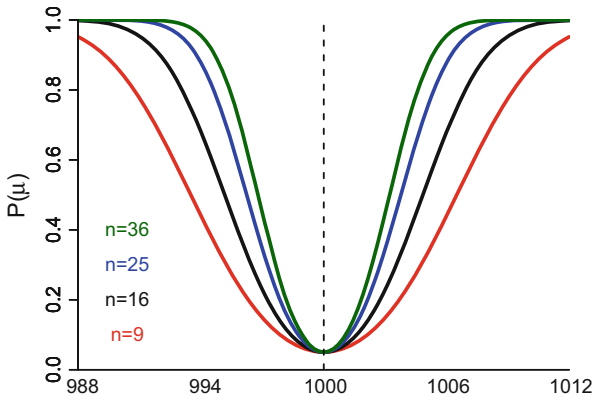


Fig. 9.17 Power of two-sided test for the population mean for alternative sample sizes, with $\mu_0 = 1000$, $\alpha = 0.05$ and $\sigma = 10$

Enhanced: Average Life Time of Car Tires

We will now illustrate how information about the population can influence the choice of the test statistic, the decision regions and—depending on the sample at hand—the test decision.

A car tire producer alters the mix of raw material entering the production process in an attempt to increase the average life of the output. After the first new tires have been sold, competitors criticize that the average life of the new tires doesn't exceed that of the old ones, which is known to be 38,000 km.

The random variable under investigation is the actual life of the population of new tires, measured in km, denoted by X, and the producer's claim is that its expectation $E(X) = \mu$ is higher than the historical one of the old types, $\mu_0 = 38,000$ km. The management wishes to scientifically test this claim and commissions a statistical investigation hoping to verify that the average life has in fact increased, i.e., that $\mu > \mu_0$. But they also want to minimize the risk of making a wrong decision so as not to be exposed to the competitors' (justified) counter arguments.

Hypothesis

Since deviations in one direction are the subject of the dispute, a one-sided test will be conducted. We put the competitors' claim in the null hypothesis with the hope that the sample rejects it, yielding a right-sided test:

$$\mathrm{H}_0 : \mu \leq \mu_0 \quad \text{versus} \quad \mathrm{H}_1 : \mu > \mu_0,$$

where $\mu_0 = 38,000$ km.

Does this operationalization support the producers' intention? We can answer this question by analyzing the possible errors.

Rejecting H_0 gives rise to the possibility of a type I error. Not rejecting the null hypothesis exposes the decision-maker to a type II error.

The producers' emphasis is on keeping the type I error small, as its implications are more severe than those of the type II error: With the production process going ahead and thus the available sample of tires gradually increasing, an actual average life below the acclaimed one would sooner or later be revealed. The maximum probability of the type I error, $P(\text{`}\mathrm{H}_1\text{'}|\mathrm{H}_0)$ is given by the significance level α, a parameter the producer can control. Thus, the test is in line with the producers' requirements.

The probability of making a type II error, $P(\text{`}\mathrm{H}_0\text{'}|\mathrm{H}_1) = \beta$, is unknown, as the true average life of the new processes' output is unknown. The probability of not verifying an increase in the average life of the tires that has actually taken place, can be substantial. That's the price the producer has to pay for choosing the conservative approach of stating the claim as alternative hypothesis and actively controlling the significance level and thus keeping the crucial type I error small. This trade-off makes sense, as the perceived long term reliability of the producer is more important than short term sales gains.

1*st Alternative*

Significance Level and Sample Size

The test will be conducted at a 0.05 significance level. A sample of size $n = 10$ is taken from the output. As the population is reasonably large (a couple of thousand tires have already been produced), the sample can be regarded as a simple random sample.

Test Statistic and Its Distribution: Decision Regions

Sample-based investigations into the tires' properties carried out prior to the implementation of changes in the production process indicate that the fluctuations in the life of the tires can be described "reasonably" well by a normal distribution with standard deviation $\sigma = 1,500$ km. Assuming, this variability is still valid in the new production regime, we have for the distribution of the sample mean under the null hypothesis:

$$\overline{X} \overset{\mathrm{H}_0}{\sim} N\left(38,000;\ \frac{1,500^2}{10}\right).$$

Under H_0, the test statistic

$$V = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n},$$

follows the standard normal distribution:

$$V \overset{\mathrm{H}_0}{\sim} N\left(0;\ 1\right).$$

The critical value c that satisfies $P\left(V \leq c\right) = 1 - \alpha = 0.95$ can be found from the cumulative standard normal distribution table as the 95 % quantile: $c = z_{0.95} = 1.645$. The resulting decision regions are:

$$\text{Non-rejection region for } \mathrm{H}_0 : \left\{v \mid v \leq 1.645\right\},$$
$$\text{Rejection region for } \mathrm{H}_0 : \left\{v \mid v > 1.645\right\}.$$

Sampling and Computing the Test Statistic

Suppose the average life of 10 randomly selected tires is $\bar{x} = 39,100$ km. Then the realized test statistic value is

$$v = \frac{39,100 - 38,000}{1,500}\sqrt{10} = 2.32.$$

Test Decision and Interpretation

As 2.32 is element of the rejection region for H_0, the null hypothesis is rejected. Based on a sample of size $n = 10$ and a significance level of $\alpha = 0.05$, we have shown statistically, that the new tires can be used significantly longer than the old ones, that is, that the true expectation $E(X) = \mu$ of the tires' life is greater than the hypothetical value $\mu_0 = 38,000$ km.

The test has resulted in a non-rejection of the alternative hypothesis H_1: "average life has increased." The producer makes a type I error ('H_1'$|H_0$) if the null hypothesis correctly describes reality (H_0: "average life has not increased"). But the probability of an occurrence of this error has intentionally been kept small with the significance level $\alpha = 0.05$.

If the alternative hypothesis is true, a right decision has been made: 'H_1'$|H_1$. The probability P('H_1'$|H_1$) of this situation can only be computed for specific true population parameters. Assuming this value is $\mu = 39,000$ km, the power is

$$\begin{aligned} P(39,000) &= 1 - P\left(V \leq 1.645 - \frac{39,000 - 38,000}{1,500}\sqrt{10}\right) \\ &= 1 - P(V \leq -0.463) = 1 - [1 - P(V \leq 0.463)] \\ &= 0.6783 = 1 - \beta. \end{aligned}$$

The greater the increase in average life, the higher the power of the test i.e., the probability $1 - \beta$. For example, if an increase to 40,000 had been achieved, the power would be 0.9949: $P(40,000) = 1 - \beta = 0.9949$.

2*nd Alternative*

The significance level $\alpha = 0.05$ and sample size $n = 10$ remain constant, and we continue to assume a normal distribution of the new tires' lives. But we drop the restrictive assumption of a constant standard deviation. We now allow for it to have changed with the introduction of the new production process.

Test Statistic and Its Distribution: Decision Regions

Since we now have to estimate the unknown standard deviation with its empirical counterpart, the square root of the sample variance, S, we must employ the T-statistic

$$T = \frac{\overline{X} - \mu_0}{S}\sqrt{n},$$

which, under H_0, has a t-distribution with $n - 1 = 9$ degrees of freedom.

We can look up the critical value c satisfying $P(T \leq c) = 1 - \alpha = 0.95$ as the upper 0.05 quantile of the t-distribution with 9 degrees of freedom in a t-distribution table and find it to be $t_{0.95;9} = 1.833$. Thus, our decision regions are:

$$\text{Non-rejection region for } \mathrm{H}_0 : \{t \,|\, t \leq 1.833\},$$
$$\text{Rejection region for } \mathrm{H}_0 : \{t \,|\, t > 1.833\}.$$

You will notice that the size of the non-rejection region has increased. This is due to the added uncertainty about the unknown dispersion parameter σ. Consequently, there must be a larger allowance for variability in the test statistic for the same significance level and sample size than in the corresponding test for known standard deviation.

Sampling and Computing the Test Statistic

Along with the sample mean $\bar{x}$ the sample standard deviation s has to be computed. Suppose their realized values are $\bar{x} = 38,900$ km and $s = 1,390$ km. Thus, the realized test statistic value is

$$t = \frac{\bar{x} - \mu_0}{s}\sqrt{n} = \frac{38,900 - 38,000}{1,390}\sqrt{10} = 2.047.$$

Test Decision and Interpretation

As $t = 2.047$ falls into the rejection region, the null hypothesis is rejected. Based on a sample of size $n = 10$ and a significance level of $\alpha = 0.05$, we were again able to statistically show that the true (and unknown) expectation $E(X) = \mu$ of the new tires' lives has increased from its former (i.e., hypothetical) level of $\mu_0 = 38,000$ km.

Of course, we still don't know the true parameter μ, and if it happens to be less than (or equal to) 38,000 km, we have made a type I error, for we have rejected a *true* null hypothesis: 'H_1'$|\mathrm{H}_0$. In choosing a significance level of 5 % we have restricted the probability of this error to a maximum of 5 % (the actual value depending on the true parameter μ).

If the true parameter μ *does* lie within the region specified by the alternative hypothesis, we have made a right decision in rejecting the null hypothesis: 'H_1'$|\mathrm{H}_1$. The probability of this event, P('H_1'$|\mathrm{H}_0) = 1-\beta$, can be (approximately) computed for alternative true population means μ if we assume the sample standard deviation s to be the true one in the population, i.e., $s = \sigma$.

3rd Alternative

Suppose we now drop the assumption of normality, which is a situation more relevant to practical applications. In order to conduct an approximate test about μ, we require the sample size to be greater than 30. If the sample size is smaller than 30, we cannot justify the application of the central limit theorem, as the approximation wouldn't be good enough. The managers decide to pick a sample of $n = 35$ tires, incurring further sampling costs as the price to employ a more suitable and therefore reliable statistical procedure. Further, suppose that the significance level is chosen to be $\alpha = 0.025$.

Test Statistic and Its Distribution: Decision Regions

As in the 2^{nd} alternative, the T-statistic

$$T = \frac{\overline{X} - \mu_0}{S} \sqrt{n},$$

has to be used. Having chosen $n > 30$ independent observations, we can justify to employ the central limit theorem and approximate the distribution of this standardized statistic by a standard normal distribution:

$$V \overset{\text{as}}{\sim} N(0;\, 1)\,.$$

In the above statement, "as" stands for "asymptotically": T is asymptotically standard normal, that is, the standard normal distribution is the limit it converges to as n tends to infinity. For finite samples, the standard normal distribution serves as an approximation. The critical value c satisfying $P(T \leq c) = 1 - \alpha = 0.975$ is then (approximately) the upper 2.5 % quantile of the standard normal distribution, $z_{0.975} = 1.96$, and we have the following decision regions:

$$\text{Non-rejection region for } \mathrm{H}_0 : \{t \,|\, t \leq 1.96\}\,,$$
$$\text{Rejection region for } \mathrm{H}_0 : \{t \,|\, t > 1.96\}\,.$$

Sampling and Computing the Test Statistic

As in the 2^{nd} alternative, we have to compute both the sample mean $\overline{x}$ and the sample standard deviation s as estimators for their population counterparts μ and σ. Suppose, their values are $\overline{X} = 38,500$ km and $s = 1,400$ km for our new sample of size 35. Then the realized test statistic value is:

$$t = \frac{\overline{x} - \mu_0}{s} \sqrt{n} = \frac{38,500 - 38,000}{1,400} \sqrt{35} = 2.11.$$

Test Decision and Interpretation

As $v = 2.11$ lies within the rejection region, the null hypothesis is rejected. On the basis of a particular sample of size $n = 35$ and a significance level of $\alpha = 0.05$ we were able to statistically verify that the true population mean $E(X) = \mu$ of the new tires' lives is greater than the tires' expected life before the implementation of the new process, $\mu_0 = 38,000$ km.

If the null hypothesis is in fact true, we have made a type I error. Fortunately, the probability of this happening (given we *have* rejected H_0 as is the case here) has been chosen not to exceed $\alpha = 0.025$ for any true population mean μ within the parameter space specified in H_0.

Given the small (maximum) type I error probability of 0.025, it is much more likely that we are right in rejecting the null hypothesis: 'H_1'$|H_1$. But the associated probability, $P('H_1'|H_0) = 1 - \beta$, can only be computed for specific true parameter values. As in the 2^{nd} alternative, we have to assume a known σ in order to calculate this quantity by setting $\sigma = s = 1,400$ km.

Interactive: Testing the Population Mean

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please select

- the hypothetical mean μ_0
- the significance level α
- the sample size n

Use "Draw sample" to manually draw a sample and carry out a test.

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

Output

In this interactive example you can study the impact of the significance level α and the sample size n on the test decision of a two-sided test:

$$H_0 : \mu = 0 \quad \text{versus} \quad H_0 : \mu \neq 0.$$

You can carry out this test as often as you like—for every new run a new sample is drawn from the population. You can vary the parameters as you like and isolate

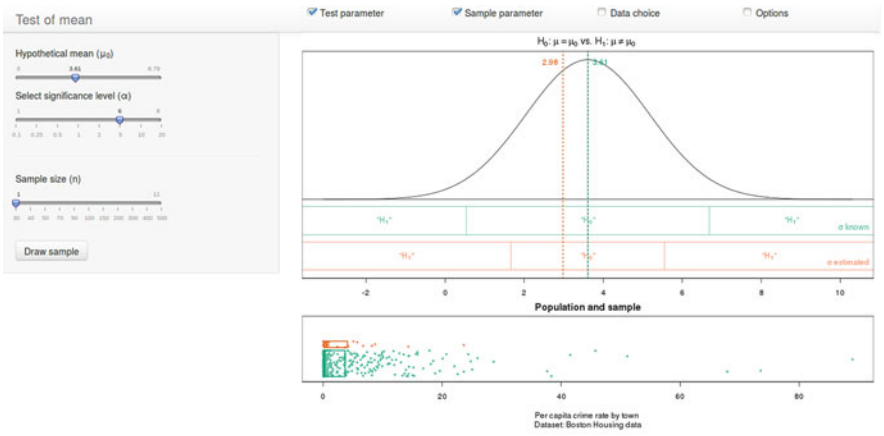


Fig. 9.18 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_tmu1

their effects by holding either of these constant. In particular, you can

- Hold both the significance level α *and* sample size n constant to observe different test decisions based on different samples;
- Vary the significance level α for a fixed sample size n;
- Change the sample size n and leave the significance level α fixed to your chosen level; or
- Vary both the significance level α and the sample size n.

The lower graphic in Fig. 9.18 is a scatterplot including the population (green) and sample (orange).

Interactive: Testing the Population Mean with Type I and II Error

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please select

- the test type
- a hypothetical μ_0
- the significance level α
- the sample size n

Use "Draw sample" to manually draw a sample and carry out a test (Fig. 9.19).

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

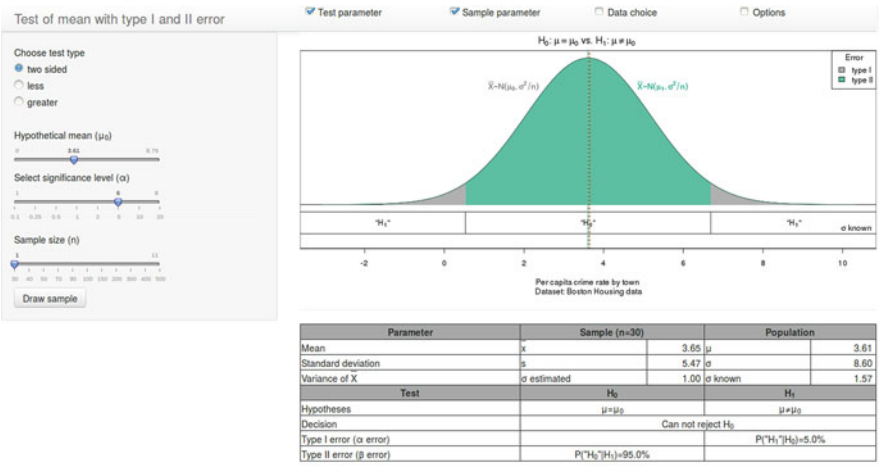


Fig. 9.19 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_terr

Output

In this interactive example you can choose three type of tests for the mean and study the impact of the significance level α, the sample size n, and the choice of μ_0 on the size of the type I and II error.

After you have made your choices you are presented a graphic containing

- the distribution of the sample mean under H_0 (gray bell curve),
- the distribution of the sample mean under H_1 (green bell curve),
- the probability of making a type I error (gray area under the gray bell curve),
- and the probability of making a type II error (green area under the green bell curve).

By varying n, σ and μ_0, you can explore the impact of these test parameters on the type I and II error probability. To isolate the impacts we recommend change the value of only one parameter in successive trials. To facilitate easy diagnostics you are shown a table containing all test values and decisions.

9.3 Testing the Proportion in a Binary Population

Consider a random variable X which has only two possible outcomes. We call the statistical population of X binary, as introduced earlier. If X is an indicator variable storing the information about the existence (or nonexistence) of a feature, we can carry out statistical inference about the proportion of elements within the population

possessing the property of interest (π) or not ($1 - \pi$). As in other parametric tests, the inference relates to a hypothetical value, here π_0, that represents a hypothetical proportion of population elements having the property of interest.

We will introduce statistical test procedures based on a *simple random sample* of size n. This ensures that the sample variables $X_1, \ldots, X_n$, which are indicator variables with outcomes measured as either 0 or 1, are independent and identically distributed Bernoulli variables. As usual the significance level is denoted by α.

Hypotheses

Depending on the application at hand, one- or two-sided tests are formulated:

1. *Two-sided test*

$$\mathrm{H}_0 : \pi = \pi_0, \quad \mathrm{H}_1 : \pi \neq \pi_0.$$

2. *Right-sided test*

$$\mathrm{H}_0 : \pi \leq \pi_0, \quad \mathrm{H}_1 : \pi > \pi_0.$$

3. *Left-sided test*

$$\mathrm{H}_0 : \pi \geq \pi_0, \quad \mathrm{H}_1 : \pi < \pi_0.$$

Our earlier remarks on the choice of null and alternative hypothesis in the section on testing population means also apply in this environment.

Test Statistic and Its Distribution: Decision Regions

The sample proportion

$$\hat{\pi} = \frac{X}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is a suitable estimator of the population parameter π. The estimator

$$X = \sum_{i=1}^{n} X_i,$$

is a simple transformation of $\hat{\pi}$ ($X = n \cdot \hat{\pi}$), which contains all the important information. It counts the number of elements in the sample possessing the property of interest. As has already been shown, X follows a Binomial distribution with parameters n and π: $X \sim B(n;\, \pi)$. As n is chosen by the decision-maker, π is the only remaining parameter needed to completely specify the Binomial distribution. Following the logic applied in all parametric hypothesis testing problems, we assume π to be π_0, that is, we determine the distribution of the test statistic given the hypothetical proportion π_0 is the one prevailing in the population: $\pi = \pi_0$. Hence, the estimator X becomes our test statistic, since it has a Binomial distribution with parameter n and π_0 under H_0:

$$V = X \overset{H_0}{\sim} B(n;\, \pi_0)\,.$$

The rejection region of the null hypothesis contains all realizations of V for which the cumulated probabilities don't exceed the significance level α. The critical values can be read from the numerical table of the cumulative distribution function $F_B(x)$ of $B(n;\, \pi_0)$, by following these rules:

1. *Two-sided test*

   The lower critical value c_l is the realization x of X, for which the cumulative distribution function just exceeds the value $\alpha/2$: $F_B(c_l - 1) \leq \alpha/2$ and $F_B(c_l) > \alpha/2$.

   The upper critical value c_u is the argument x of the cumulative distribution function that returns a probability equal to or greater than $1-\alpha/2$: $F_B(c_u - 1) < 1 - \alpha/2$ and $F_B(c_u) \geq 1 - \alpha/2$. The rejection region for H_0 is given by $\{v \mid v < c_l \text{ or } v > c_u\}$, such that

   $$P(V < c_l|\pi_0) + P(V > c_u|\pi_0) \leq \alpha.$$

   For the non-rejection region for H_0 we have $\{v \mid c_l \leq v \leq c_u\}$, such that

   $$P(c_l \leq V \leq c_u|\pi_0) \geq 1 - \alpha.$$

2. *Right-sided test*

   The critical value c is the smallest realization of the test statistic that occurs with cumulated probability of at least $1-\alpha$: $F_B(c - 1) < 1-\alpha$ and $F_B(c) \geq 1-\alpha$. The rejection region for H_0 is then $\{v \mid v > c\}$, such that

   $$P(V > c|\pi_0) \leq \alpha.$$

   The non-rejection region for H_0 is $\{v \mid v \leq c\}$, such that

   $$P(V \leq c|\pi_0) \geq 1 - \alpha.$$

3. Left-sided test

The critical value c is determined as the smallest realization of the test statistic that occurs with cumulated probability of at least α: $F_B(c-1) \leq \alpha$ and $F_B(c) > \alpha$. The rejection region for H_0 is $\{v \mid v < c\}$, such that

$$P(V < c|\pi_0) \leq \alpha.$$

The non-rejection region for H_0 is given by $\{v \mid v \geq c\}$, such that

$$P(V \geq c|\pi_0) \geq 1 - \alpha.$$

As $V = X$ is a discrete random variable, the given significance level α will generally not be fully utilized (exhausted). The actual significance level α_a will only by chance reach that level and will usually be smaller. The above tests are thus conservative with respect to the utilization of the allowance for the maximum probability of the type I error.

Given the sample size n is sufficiently large, the estimator $\hat{\pi}$ can be standardized to give the test statistic

$$V = \frac{\hat{\pi} - \pi_0}{\sigma_0(\hat{\pi})} = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}.$$

Here, $\sigma_0(\hat{\pi})$ is the standard deviation of the estimation function $\hat{\pi}$ under H_0. Under H_0, V has approximately standard normal distribution (i.e., normal with mean 0 and variance 1). Critical values for the given significance level can be taken from the cumulative standard normal distribution table. Decision regions for the one- and two-sided tests are determined in the same way as those for the approximate population mean test for unknown σ: In fact, a hypothesis about a proportion is a hypothesis about an expectation (of a binary indicator variable): $E(\hat{\pi}) = \pi$.

Sampling and Computing the Test Statistic

Once a sample of size n has been drawn, we have realization $x_1, \ldots, x_n$ of the sampling variables, $X_1, \ldots, X_n$, and can compute the realized value v of the test statistic V.

Test Decision and Interpretation

See the remarks for the μ test.

Power Curve $P(\pi)$

The *power* curve of the large-sample test based on

$$V = \frac{\hat{\pi} - \pi_0}{\sigma_0(\hat{\pi})} = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

can be calculated explicitly for all test situations in the same manner as the power curve for the population mean tests.

The power curve of the exact test based on $V = X$ is computed using the *Binomial distribution* (as this is the distribution underlying the test statistic) for all $0 \leq \pi \leq 1$ and fixed n.

From the definition

$$P(\pi) = P(V = X \in \text{ rejection region for } \mathrm{H}_0 \mid \pi)$$

it follows:

1. *For the two-sided test*

$$\begin{aligned} P(\pi) &= P(V < c_l \mid \pi) + P(V > c_u \mid \pi) \\ &= P(V \leq c_l - 1 \mid \pi) + [1 - P(V \leq c_u \mid \pi)], \end{aligned}$$

2. *For the right-sided test*

$$P(\pi) = P(V > c \mid \pi) = 1 - P(V \leq c \mid \pi),$$

3. *For the left-sided test*

$$P(\pi) = P(V < c \mid \pi) = P(V \leq c - 1 \mid \pi).$$

Given the respective critical values, the probabilities can be looked up in the numerical table of the cumulative Binomial distribution function. For $\pi = \pi_0$, the power curve equals the actual significance level α_a.

Explained: Testing a Population Proportion

A statistics professor has the impression that in the last year the university library has bought proportionally less new statistics books than in the past. Over the last couple of years the relative amount of statistics books amongst new purchases

has consistently been more than 10 %. He asks one of his assistants to investigate whether this has changed in favor of other departments. Acting on behalf of his students whom he wants to secure as many new books as possible, he asks his assistant to minimize the risk of not complaining to the head of the library when the proportion of statistics books *has* decreased.

The assistant decides to have a sample of 25 books taken from the file containing the new purchases over the last 12 months. He wants to know how many of these are statistics books. He is thus dichotomizing the random variable "subject matter" into the outcomes "statistics" and "not statistics." Of course, if you regard the purchases as an outcome of a decision-making process conducted by the librarians, this is anything but a random variable. But for the statisticians who rely on a sample because they don't have access to all relevant information, it appears to be one. From the proportion of statistics books the assistant wants to infer to the population of all newly purchased books, using a statistical test to allow for deviations of the proportion in the sample from those in the population. In particular, he wants to verify whether the proportion has indeed dropped below the past average of 10 %. He will thus test the population proportion π and chooses a "standard" significance level of 0.05.

Hypothesis

As the assistant wants to verify whether the proportion has dropped below 0.1, he has to employ a one-sided test. He recalls that the professor wants him to minimize the probability of not disclosing that the proportion has decreased below $\pi_0 = 0.1$ when in reality it has. He thus opts for a right-sided test, i.e., puts the professors' claim as null hypothesis in the hope of not rejecting it:

$$\mathrm{H}_0 : \pi \leq \pi_0 = 0.1 \quad \text{versus} \quad \mathrm{H}_1 : \pi > \pi_0 = 0.1.$$

The assistant undertakes an investigation into the properties of this test with respect to the professors' intention of minimizing the probability of not detecting a relative decrease in the statistics book supply. A real-world decrease can only not have been detected if the null hypothesis has been rejected even though it is really true. This situation is called type I error:

$$\begin{aligned} \text{`H}_1\text{'}|\mathrm{H}_0 = &\text{`conclude proportion of} \\ &\text{statistics books has } \textit{not} \text{ decreased'}| \\ &\text{in reality, the proportion } \textit{has} \text{ decreased.} \end{aligned}$$

The maximum probability of this situation, $P(\text{`H}_1\text{'}|\mathrm{H}_0)$, is given by the significance level α, which has been set to 0.05. Thus, the risk the professor wanted to "minimize" is under control.

If the null hypothesis is not-rejected, then a type II error can arise:

$$\text{`H}_0\text{'}|\text{H}_1 = \text{`conclude proportion of statistics books has decreased'}| \text{ in reality, the proportion has } \textit{not} \text{ decreased.}$$

The probability of this happening (conditional on the null hypothesis not having been rejected), $P(\text{`H}_1\text{'}|\text{H}_0) = \beta$, is unknown, because the true proportion π (which is element of the parameter set specified by the alternative hypothesis), is unknown. As we have already seen in other examples, it can be substantial, but the professors' priorities lie on trading off type II error for type I error which is under control.

Test Statistic and Its Distribution: Decision Regions

The estimator X : "number of statistics books in a sample of 25 books" can serve as test statistic V. Under H_0, $V = X$ has Binomial distribution with parameter $n = 25$ and $\pi = 0.1$: $V \sim B(25; 0.1)$. A relatively high number of statistics books in the sample supports the alternative hypothesis that the proportion of statistics books has not decreased. The critical value c is the realization of X, for which $F_B(c)$ equals or exceeds $1 - \alpha = 0.95$, that is, we require $F_B(1 - c) < 1 - \alpha = 0.95$ and $F_B(c) \geq 1 - \alpha = 0.95$.

In the table of the cumulative distribution function of $B(25; 0.1)$ you will find $c = 5$. The rejection region for H_0 is thus $\{v \mid v > 5\} = \{6, 7, \ldots, 25\}$, such that

$$P(V > 5|0.1) = 0.0334 = \alpha_a < \alpha.$$

As $V = X$ is a discrete random variable, the given significance level isn't fully utilized:$\alpha_a = 0.0334 < \alpha = 0.05$.

The non-rejection region for H_0 is given by $\{v \mid v \leq 5\} = \{0, 1, 2, 3, 4, 5\}$, such that

$$P(V \leq 5|0.01) = 0.9666.$$

Sampling and Computing the Test Statistic

A subset of 25 books is selected at random from the list of last years' new purchases and categorized in statistics and non-statistics books. As the total amount of new books is sufficiently large from a sample-theoretical point of view, a simple random sample is drawn, i.e., the sampling is carried out without replacement. The amount

of statistics books in the sample is counted to be $x = 3$, which will serve as the realized test statistic value v.

Test Decision and Interpretation

As $v = 3$ falls into the non-rejection region for H_0, the null hypothesis cannot be rejected. On the basis of a random sample of size $n = 25$ and a significance level of $\alpha = 0.05$, the assistant couldn't verify statistically that the proportion of statistics books is still above 10 %. This test result means that a complaint to the library seems to be merited.

Power

Given our test parameters ($\pi_0 = 0.1$, $n = 25$, $\alpha = 0.05$ and $c = 5$), what is the probability of not rejecting the null hypothesis if the true proportion of statistics books is $\pi = 0.2$? That is, we want to calculate the probability of the type II error given a specific element of the parameter set associated with the alternative hypothesis, $\pi = 0.2$:

$$\begin{aligned}\beta\,(0.2) &= P\,(\text{`}H_0\text{'}|H_1) \\ &= P\,(V = X \in \text{ non-rejection region for } H_0 \,|\, \pi = 0.2) \\ &= P\,(V \leq 5 \,|\, \pi = 0.2)\,.\end{aligned}$$

In the table of the cumulative Binomial distribution $B\,(25;\ 0.2)$ we find this probability to be 0.6167. Alas, if the true proportion has increased to 20 %, there is still a 61.67 % chance of not discovering a significant deviation from the hypothetical boundary proportion of 10 %. This is the probability of an unjustified complaint issued by the professor given the proportion has risen to 0.2—a substantial relative increase.

The probability of making a type II error contingent on alternative true proportions π can be computed via the power curve. Levels of $P\,(\pi)$ and $1 - P\,(\pi)$ for several values of π are listed in Table 9.4.

For example, if the true proportion (and therefore absolute amount) of statistics books is $\pi = 0$, the sample cannot contain any statistics books and we will expect $x = 0$ and won't reject the null hypothesis. The rejection of the null hypothesis ('H_1') is an impossible event with associated probability of zero. The power is the conditional probability of rejecting the null hypothesis given the relative amount is zero:

$$P\,(0) = P\,(V = X \in \text{ rejection region for } H_0 \,|\, \pi = 0) = P\,(\text{`}H_1\text{'} \,|\, 0) = 0.$$

Table 9.4 Some values of the power function

π	True hypothesis	$P(\pi)$	$1 - P(\pi)$
0	H_0	$0 = \alpha$	$1 = 1 - \alpha$
0.05	H_0	$0.0012 = \alpha$	$0.9988 = 1 - \alpha$
0.1	H_0	$0.0334 = \alpha_a$	$0.9666 = 1 - \alpha_a$
0.15	H_1	$0.1615 = 1 - \beta$	$0.8385 = \beta$
0.20	H_1	$0.3833 = 1 - \beta$	$0.6167 = \beta$
0.25	H_1	$0.6217 = 1 - \beta$	$0.3783 = \beta$
0.30	H_1	$0.8065 = 1 - \beta$	$0.1935 = \beta$
0.35	H_1	$0.9174 = 1 - \beta$	$0.0826 = \beta$
0.40	H_1	$0.9706 = 1 - \beta$	$0.0294 = \beta$
0.45	H_1	$0.9914 = 1 - \beta$	$0.0086 = \beta$
0.50	H_1	$0.9980 = 1 - \beta$	$0.0020 = \beta$
0.60	H_1	$0.9999 = 1 - \beta$	$0.0001 = \beta$
0.70	H_1	$1 = 1 - \beta$	$0 = \beta$

If, on the other hand, the true proportion of statistics books is $\pi = 0.35$, the power is calculated as

$$\begin{aligned} P(0.35) = P(V > 5 \,|\, \pi = 0.35) &= 1 - P(V \leq 5 \,|\, \pi = 0.35) \\ &= 1 - 0.0826 = 0.9174, \end{aligned}$$

where $P(V \leq 5 \,|\, \pi = 0.35)$ can be looked up in the table of the cumulative distribution function as the value of $B(25;\, 0.2)$ for $c = 5$.

$P(0.35)$ is the probability of correctly rejecting the null hypothesis, $P(\text{`}H_1\text{'}|H_1)$. The probabilities of rejecting the null hypothesis and not-rejecting it must always sum up to one for any given true parameter value within the range specified by the alternative hypothesis:

$$P(\text{`}H_0\text{'}|H_1) + P(\text{`}H_1\text{'}|H_1) = 1.$$

For a true proportion of $\pi = 0.35$, the former sampling result amounts to making a type II error, the probability of which is denoted by $\beta(0.35)$. Thus, we can write

$$\beta(0.35) + P(\text{`}H_1\text{'}|H_1) = 1$$

or

$$P(\text{`}H_1\text{'}|H_1) = 1 - \beta(0.35).$$

As $P(\text{`}H_1\text{'}|H_1)$ is the value of the power at point $\pi = 0.35$, we can calculate the probability of making a type II error as

$$\beta(0.35) = 1 - P(0.35) = 0.0826.$$

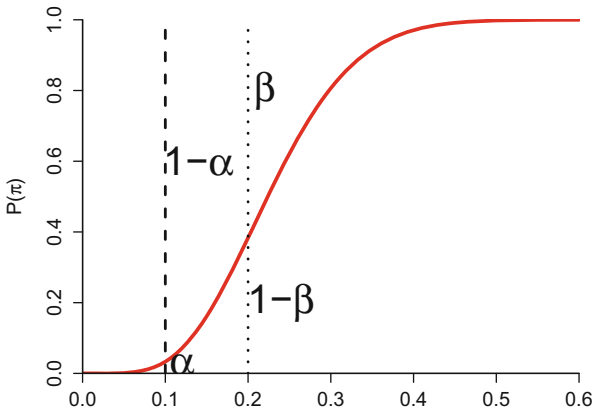


Fig. 9.20 Power curve for the right-sided test

If the true proportion of statistics books is 35 %, 8.26 % of all samples of size $n = 25$ will lead to a non-rejection of the null hypothesis, i.e., won't detect the significant difference between $\pi = 0.35$ and $\pi_0 = 0.10$.

Figure 9.20 depicts the graph of the power curve for the right-sided test we have just discussed: $\pi_0 = 0.10, n = 25, \alpha = 0.05$ and $c = 5$.

Enhanced: Proportion of Credits with Repayment Problems

One of the raison d'etres of financial intermediaries is their ability to efficiently assess the credit-standing ('creditworthiness') of potential borrowers.

The management of ABC bank decides to introduce an extended credit checking scheme if the proportion of customers with repayment irregularities isn't below 20 %. The in-house statistician conducting the statistical test is asked to keep the probability of not deciding to improve the credit rating procedure even though the proportion is "really" above 20 % low (i.e., to keep α low).

The random variable X : "credit event" or "repayment problems" is defined as an indicator variable taking on zero ("no") or one ("yes"). The actual proportion π of clients having trouble with servicing the debt is unknown. The hypothetical boundary value for testing this population proportion is $\pi_0 = 0.2$.

Hypothesis

Deviations from the hypothetical parameter into one direction are of interest; thus, a one-sided test will be employed. As the bank hopes to prove that the evaluation processes in place are sufficient, i.e., the proportion of debtors displaying

irregularities in repaying their loans is less than 20 %, this claim is formulated as the alternative hypothesis:

$$H_0 : \pi \geq \pi_0 = 0.2 \quad \text{versus} \quad H_1 : \pi < \pi_0 = 0.2$$

The properties of this test with respect to the bank managers' requirements have to be evaluated to ensure the test really meets their needs. The type I error, which can be made if the null hypothesis is rejected, is here:

$$\begin{aligned} \text{`}H_1\text{'}|H_0 = &\text{`do not-reject that the proportion of} \\ &\text{problematic debtors} < 0.2\text{; no new guidelines'} \mid \\ &\text{in reality, unreliable debtors make up} \\ &\text{at least 20 \%; credit process has to be reviewed.} \end{aligned}$$

If the test results in the non-rejection of the null hypothesis, a type II error might occur:

$$\begin{aligned} \text{`}H_0\text{'}|H_1 = &\text{`do not-reject that the proportion of} \\ &\text{problematic debtors} \geq 0.2\text{;} \\ &\text{new evaluation process to be developed'} \mid \\ &\text{in reality, unreliable debtors make up no more than 20} \\ &\text{per cent; no need for action.} \end{aligned}$$

The type I error represents the risk the managers of the ABC bank want to cap. Its maximum level is given by the significance level, which has been set to a sufficiently low level of 0.05.

The type II error represents the risk of a costly introduction of new credit evaluation processes without management-approved need. The impact of this scenario on the banks' profitability is difficult to assess, as the new process will lead to a repricing of credits and thus may also generate cost *savings*. The following two alternatives are both based on the above test.

A random sample is drawn from the population of 10,000 debtors without replacement. This is reasonable, if $n/N \leq 0.05$, as the random sample can then be regarded as "simple" anyway.

1*st* Alternative

To curb costs, a sample size of $n = 30$ is chosen. The sampling-theoretical requirement $n/N \leq 0.05$ is fulfilled.

Test Statistic and Its Distribution: Decision Regions

The estimator X : "Number of clients with irregularities in debt servicing in sample of size 30" can directly serve as our test statistic V. Under H_0, $V = X$ has Binomial distribution $B(30; 0.2)$. A small V supports the alternative hypothesis. The critical value c is the smallest realization of X, for which $F_B(x)$ equals to or is greater than α , i.e., it has to satisfy: $F_B(c-1) \leq \alpha = 0.05$ and $F_B(c) > \alpha = 0.05$. In the numerical table of the cumulative distribution function of $B(30; 0.2)$ we find $c = 3$, and thus we have the following decision regions:

Rejection region for H_0:

$$\{v \mid v < 3\} = \{0, 1, 2\}, \text{ with } P(V < 3|0.2) = 0.0442.$$

Non-rejection region for H_0:

$$\{v \mid v \geq 3\} = \{3, 4, \ldots, 30\}, \text{ with } P(V \geq 3|0.02) = 0.9558.$$

Because $V = X$ is a discrete random variable, the given significance level isn't exhausted: i.e., $\alpha_a = 0.0442 < \alpha = 0.05$.

Sampling and Computing the Test Statistic

30 randomly selected debtors are investigated with respect to reliability in debt servicing. Assume 5 of them haven't always fulfilled their contractual obligations: $v = 5$.

Test Decision and Interpretation

As $v = 5$ belongs to the non-rejection region for H_0, the null hypothesis is not-rejected. Even though the sample proportion $x/n = 5/30 = 0.167$ is smaller than the hypothetical boundary proportion $\pi_0 = 0.2$, which should favor H_1, we cannot conclude H_0 is false: at a significance level of 0.05, the difference cannot be regarded as statistically significant. In other words: It is far too likely that the difference has arisen from sampling variability due to the small sample size to be able to reject the null hypothesis. It is important to observe that it's not merely the value of the point estimator compared to the hypothetical value that leads to a non-rejection or rejection of the null hypothesis, but intervals that take into account the random character of the estimator (i.e., the difference is compared to an appropriate, case specific, statistical yardstick to determine what is statistically significant large, and hence small, deviations/differences). Based on a random sample of size $n = 30$ and a significance level $\alpha = 0.05$, we were unable to show statistically, that the proportion of trouble debtors is significantly smaller than 20 %. Consequently, the ABC bank will review and try to improve the credit approval procedures.

Power

Not having rejected the null hypothesis, we are vulnerable to a type II error, which occurs when the alternative hypothesis is a true statement: 'H_0'$|H_1$.

Let's calculate the type II error probability for a true parameter value $\pi = 0.15$: What is the probability of not rejecting the null hypothesis in a left-sided test with $\pi_0 = 0.2$, $n = 30$, $\alpha = 0.05$ and $c = 3$, given the true population proportion is $\pi = 0.15$ and hence the null hypothesis actually wrong?

$$\begin{aligned}\beta\,(\pi = 0.15) &= P\,(\text{'H}_0\text{'}|\text{H}_1)\\ &= P\,(V = X \in \text{non-rejection region for H}_0|\pi = 0.15)\\ &= P\,(V \geq 3|\pi = 0.15)\,.\end{aligned}$$

We compute

$$\begin{aligned}P\,(V \geq 3\,|\,\pi = 0.15) &= 1 - P\,(V < 3\,|\,\pi = 0.15)\\ &= 1 - P\,(V \leq 2\,|\,\pi = 0.15) = 1 - 0.1514 = 0.8486,\end{aligned}$$

where $P\,(V \leq 2\,|\,\pi = 0.15)$ is taken from the table of the cumulative distribution function $B\,(30;\,0.15)$ for $c = 2$, that is $F_B\,(2)$.

Interpretation Given the true proportion is $\pi = 0.15$, 84.86 % of all samples of size $n = 30$ will not be able to discriminate between the true parameter and the hypothetical $\pi_0 = .20$, inducing the bank to undertake suboptimal improvements of the credit assessment process with probability 0.8486. In deciding to control the maximum error I probability, the bank is accepting type II error probabilities of such magnitude, statisticians can provide management with power function graphs for any desired true parameter value π.

Of course, not rejecting the null hypothesis can also be the right decision: 'H_0'$|H_1$. Suppose, for example, that the true proportion of unreliable debtors is $\pi = 0.25$. The probability of not rejecting the null hypothesis and hence (unknowingly) making the right decision given our current test setting (left sided with $\pi_0 = 0.20$, $n = 30$, $\alpha = 0.05$ and thus $c = 3$) is

$$\begin{aligned}&P\,(V = X \in \text{ non-rejection region for H}_0\,|\,\pi = 0.25)\\ &= P\,(V \geq 3\,|\,\pi = 0.25) = P\,(\text{'H}_0\text{'}|\text{H}_1) = 1 - \alpha.\end{aligned}$$

We have

$$\begin{aligned}&P\,(V \geq 3\,|\,\pi = 0.25) = 1 - P\,(V < 3\,|\,\pi = 0.25)\\ &= 1 - P\,(V \leq 2\,|\,\pi = 0.25) = 1 - 0.0106 = 0.9894,\end{aligned}$$

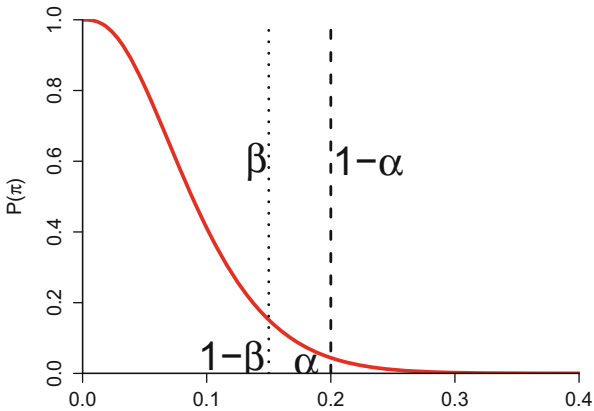


Fig. 9.21 Power curve of the left-sided test with parameters $\pi_0 = 0.20$, $n = 30$, $\alpha = 0.05$ and $c = 3$

Table 9.5 Some values of the power function

π	True hypothesis	$P(\pi)$	$1 - P(\pi)$
0	H_1	$1 = 1 - \beta$	$0 = \beta$
0.05	H_1	$0.8122 = 1 - \beta$	$0.1878 = \beta$
0.10	H_1	$0.4114 = 1 - \beta$	$0.5886 = \beta$
0.15	H_1	$0.1514 = 1 - \beta$	$0.8486 = \beta$
0.20	H_0	$0.0442 = \alpha_a$	$0.9558 = 1 - \alpha_a$
0.25	H_0	$0.0106 = \alpha$	$0.9894 = 1 - \alpha$
0.30	H_0	$0.0021 = \alpha$	$0.9979 = 1 - \alpha$
0.35	H_0	$0.0003 = \alpha$	$0.9997 = 1 - \alpha$
0.40	H_0	$0 = \alpha$	$1 = 1 - \alpha$

where $P(V \leq 2 \,|\, \pi = 0.25)$ can be looked up in a numerical table of $B(30;\ 0.25)$ as the cumulative probability for values less than or equal to $c = 2$, i.e., $F_B(2)$.

These calculations can be carried out for any desired parameter value within the overall parameter space (here: $\pi \in (0, 1)$). Depending on which hypothesis the individual parameter adheres to, the power curve $P(\pi)$ or $1 - P(\pi)$ returns probabilities for making a right decision or a type I or type II error. Figure 9.21 shows the graph of the power curve of the left-sided test with parameters $\pi_0 = 0.20$, $n = 30$, $\alpha = 0.05$ and $c = 3$ (Table 9.5).

2*nd* Alternative

Now the statistician tries to both satisfy the parameter $\alpha = 0.05$ set by the management to contain the probability of the crucial type I error *and* keep the type II error as low as possible. She is aware of the trade-off relationship between α and β error and focuses on possibilities of reducing the associated probabilities simultaneously by increasing the sample size n and thus making the decision an economic one. Cost projections in conjunction with a valuation of the benefit

of higher reliability lead to a choice of $n = 350$, still small enough to satisfy $n/N \leq 0.05$ as basis for simple random sampling without replacement.

Test Statistic and Its Distribution: Decision Regions

The standardized test statistic

$$V = \frac{\hat{\pi} - \pi_0}{\sigma_0(\hat{\pi})} = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

is used. Under H_0, it is approximately normally distributed with parameters $\mu = 0$ and $\sigma = 1$. Large sample theory suggests that the approximation is sufficiently accurate for a sample size of $n = 350$. From the cumulative standard normal distribution table we can take $c = z_{0.95} = 1.645$ to satisfy $P(V \leq c) = 1 - \alpha = 0.95$. From symmetry it follows that $-c = -1.645$, and we have $\{v \mid v < -1.645\}$ as the approximated rejection region for H_0 and $\{v \mid v \geq -1.645\}$ as the approximated non-rejection region for H_0.

Sampling and Computing the Test Statistic

From the universe of 10,000 debtors, 350 are selected and random, of which 63 turn out to have displayed problems in debt servicing at least once in their repayment history. Their proportion within the sample is thus 0.18. Plugging this into the test statistic yields

$$v = \frac{0.18 - 0.2}{\sqrt{\frac{0.2 \cdot (0.8)}{350}}} = -0.935.$$

Test Decision and Interpretation

As $v = -.0935$ falls into the non-rejection region for H_0, the null hypothesis is not rejected. On the basis of this particular sample of size $n = 350$, it cannot be statistically claimed, that the proportion of problematic debtors is less than 20 %. The ABC bank management will thus initiate a review of their credit procedures.

Type II Error Probability

As the bank management has been induced to not-reject the statement in the null hypothesis, it may have made a type II error, which occurs if the true proportion

amongst the 10,000 is actually smaller than 0.2: 'H_0'|H_1. Let's examine the probability of this happening for a "hypothetical" true population proportion of $\pi = 0.15$, i.e., $P(\text{'H}_0\text{'}|\text{H}_1) = \beta(\pi = 0.15)$.

First we must determine the critical proportion p_c corresponding to the critical value calculated using the normal approximation. From

$$-c = (p_c - \pi_0)/\sigma(\hat{\pi})$$

follows

$$p_c = \pi_0 - c \cdot \sigma(\hat{\pi}) = 0.2 - 1.645\ (0.2 \cdot 0.8/350) = 0.1648.$$

$\beta(\pi = 0.15)$ is the probability of the sample function $\hat{\pi}$ assuming a value from the non-rejection region of the null hypothesis, given the true parameter π belongs to the alternative hypothesis:

$$\beta(\pi = 0.15) = P(\hat{\pi} \geq p_c \,|\, \pi = 0.15) = P(\hat{\pi} \geq 0.1648 \,|\, \pi = 0.15).$$

In order to determine this probability on the basis of a numerical table for the *standard* normal distribution, we must standardize using $E(\hat{\pi}) = \pi = 0.15$ and $Var(\hat{\pi}) = \pi(1-\pi)/n = 0.15 \cdot 0.85/350$:

$$\beta(\pi = 0.15) = P(\hat{\pi} \geq p_c \,|\, \pi = 0.15) = P\left(\frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi(1-\pi)}{n}}} \geq \frac{p_c - \pi_0}{\sqrt{\frac{\pi(1-\pi)}{n}}} \,|\, \pi = 0.15\right)$$

$$= P\left(\frac{0.1648 - 0.15}{\sqrt{\frac{0.15 \cdot (0.85)}{350}}} \,|\, \pi = 0.15\right) = P(V \geq 0.775 \,|\, \pi = 0.15).$$

In the standard normal distribution table we find $P(V \leq 0.775) = 0.7808$ and thus have

$$\beta(\pi = 0.15) = 1 - P(V \leq 0.775) = 1 - 0.7808 = 0.2192.$$

Thus, compared to $\beta(\pi = 0.15)$ from the 1^{st} alternative, the increase in the sample size has resulted in a sizeable reduction in the error type II probability for a true population proportion of $\pi = 0.15$.

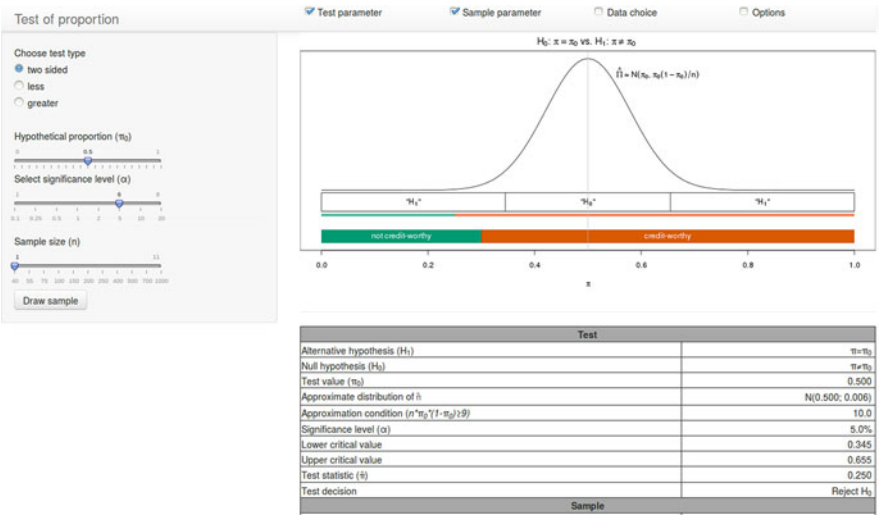


Fig. 9.22 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_tprop

Interactive: Testing a Proportion in a Binary Population

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please select

- the test type
- a hypothetical proportion π_0
- the significance level α
- the sample size n

Use "Draw sample" to manually draw a sample and carry out a test (Fig. 9.22).

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

Output

In this interactive example you can choose three type of tests for the proportion and study the impact of the significance level α, the sample size n and the choice of π_0 on the test decision.

After you have made your choices you are presented a graphic containing

- the distribution of the sample proportion under H_0 (bell curve),
- a vertical line displaying your π_0
- a green and orange area showing the test decisions

By varying n, σ and μ_0, you can explore the impact of these test parameters on the test decision. To isolate the impacts we recommend change the value of only one parameter in successive trials. To facilitate easy diagnostics you are shown a table containing all test values and decisions.

9.4 Testing the Difference of Two Population Means

The unknown parameter to be tested now is the difference of two expectations in two distinguishable populations, $(\mu_1 - \mu_2)$. Our parameter tests will be based on individual samples arising from these two populations; we will thus be dealing with two-sample tests.

There are many different ways of constructing tests for the difference in two population expectations. Our tests will be suited to the following assumptions:

- There are two populations. The random variable observed in the first, X_1 has expectation $E(X_1) = \mu_1$ and variance $Var(X_1) = \sigma_1^2$; the parameters of the random variable observed in the second population, X_2, are $E(X_2) = \mu_2$ and $Var(X_2) = \sigma_2^2$. We test for the difference in their expected values, because we have to regard μ_1 and μ_2 as unknown.
- The sizes of the two populations, N_1 and N_2, are sufficiently large to base the test procedures on simple random samples drawn without replacement. The sample sizes are denoted by n_1 and n_2, respectively.
- The two samples are independent. This means they are drawn independently of each other so as to not convey any cross-sample information.
- Either the random variables X_1 and X_2 are normally distributed ($X_1 \sim N(\mu_1; \sigma_1)$ and $X_2 \sim N(\mu_2; \sigma_2)$), or the sums of observations from both populations can be approximated sufficiently accurately by a normal distribution via the central limit theorem. For this to be feasible, the sample sizes n_1 and n_2 have to be sufficiently large.

There is a hypothesis about the difference, expressed in terms of $\omega_0 = \mu_1 - \mu_2$. A special case of particular practical interest is that of hypothetical equality of the two population means, i.e., $\omega_0 = 0$. The test will be conducted at a significance level of α.

Hypotheses

Depending on the application at hand, a two- or one-sided test will be carried out:

1. Two-sided test

$$\mathrm{H}_0 : \mu_1 - \mu_2 = \omega_0 \quad \text{versus} \quad \mathrm{H}_1 : \mu_1 - \mu_2 \neq \omega_0.$$

2. Right-sided test

$$\mathrm{H}_0 : \mu_1 - \mu_2 \leq \omega_0 \quad \text{versus} \quad \mathrm{H}_1 : \mu_1 - \mu_2 > \omega_0.$$

3. Left-sided test

$$\mathrm{H}_0 : \mu_1 - \mu_2 \geq \omega_0 \quad \text{versus} \quad \mathrm{H}_1 : \mu_1 - \mu_2 < \omega_0.$$

The choice of the appropriate test should be guided by the considerations laid out in the section on one-sample tests of μ.

Test Statistic and Its Distribution: Decision Regions

We have already shown that the estimator of the difference of two expectations,

$$D = \overline{X}_1 - \overline{X}_2,$$

where $\overline{X}_1$ and $\overline{X}_2$ are the sample means, that is,

$$\overline{X}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i} \quad \overline{X}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} X_{2i},$$

has normal distribution with expectation $E(D) = \omega = \mu_1 - \mu_2$. Independence of the sample variables implies the variance of the sample mean differential is the difference of the variances of the sample means:

$$Var(D) = \sigma_D^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Assume that ω_0 is the true distance between the population expectations: $\omega = \omega_0$. Then D follows a normal distribution with expectation $E(D) = \omega_0$ and variance σ_D^2.

In constructing an appropriate test statistic, we have to make the same distinction concerning our knowledge about the standard deviations σ_1 and σ_2 as in the one-sample case. Let's start with the simplifying (and unrealistic) assumption that, for some miraculous reason, we know the standard deviations in both populations, σ_1 and σ_2.

Known Standard Deviations σ_1 and σ_2
If we know σ_1 and σ_2, the distribution of D is fully specified as above, and we can standardize D to ensure the applicability of numerical tables for the standard normal distribution:

$$V = \frac{D - \omega_0}{\sigma_D} = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \omega_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Table 9.6 Rejection and non-rejection region for known standard deviations

Test	Rejection region for H_0	Non-rejection region for H_0
Two- sided	$\{v \mid v < -z_{1-\alpha/2}, \text{ or } v > z_{1-\alpha/2}\}$	$\{v \mid -z_{1-\alpha/2} \leq v \leq z_{1-\alpha/2}\}$
Right-sided	$\{v \mid v > z_{1-\alpha}\}$	$\{v \mid v \leq z_{1-\alpha}\}$
Left-sided	$\{v \mid v < z_{-1-\alpha}\}$	$\{v \mid v \geq -z_{1-\alpha}\}$

Under H_0, V has (at least approximately) *standard normal distribution*, and the table of numerical values of the cumulative standard normal distribution can be used to determine *critical values*. These normal quantiles translate into the decision regions for tests at a significance level α shown in Table 9.6.

Unknown Standard Deviations σ_1 and σ_2
We have to estimate the unknown quantities σ_1 and σ_2 using their sample counterparts:

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(X_{1i} - \overline{X}_1\right)^2, \quad S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} \left(X_{2i} - \overline{X}_2\right)^2.$$

Assuming *homogeneity in variances*, i.e., the random variable under consideration has the same dispersion in both populations, $\sigma_1^2 = \sigma_2^2$, the estimation function S^2 of the joint variance σ^2 is a weighted arithmetic average of the two variance estimators S_1^2 and S_2^2:

$$S^2 = \frac{(n_1 - 1)\ S_1^2 + (n_2 - 1)\ S_2^2}{n_1 + n_2 - 2}.$$

Thus, we can write the estimator S_D^2 of σ_D^2 as

$$S_D^2 = S^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right) = \frac{n_1 + n_2}{n_1\, n_2} \frac{(n_1 - 1)\ S_1^2 + (n_2 - 1)\ S_2^2}{n_1 + n_2 - 2}.$$

The test statistic V is then calculated as

$$V = \frac{D - \omega_0}{S_D} = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \omega_0}{\sqrt{\frac{n_1+n_2}{n_1\, n_2} \frac{(n_1-1)\, S_1^2 + (n_2-1) \cdot S_2^2}{n_1+n_2-2}}},$$

and has $t-$distribution with $n_1 + n_2 - 2$ degrees of freedom.

Under the assumption of *heterogeneous variances*, $\sigma_1^2 \neq \sigma_2^2$, the estimator S_D^2 can only be approximated as

$$S_D^2 = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}.$$

Table 9.7 Rejection and non-rejection region for unknown standard deviations

Test	Rejection region for H_0	Non-rejection region for H_0
Two-sided	$\{v \mid v < -t_{1-\alpha/2;n_1+n_2-2}$, or $v > t_{1-\alpha/2;n_1+n_2-2}\}$	$\{v \mid -t_{1-\alpha/2;n_1+n_2-2} \leq v \leq t_{1-\alpha/2;n_1+n_2-2}\}$
Right-sided	$\{v \mid v > t_{1-\alpha;n_1+n_2-2}\}$	$\{v \mid v \leq t_{1-\alpha;n_1+n_2-2}\}$
Left-sided	$\{v \mid v < t_{-1-\alpha;n_1+n_2-2}\}$	$\{v \mid v \geq -t_{1-\alpha;n_1+n_2-2}\}$

Welsh has suggested to base the test statistic on this approximation and use

$$V = \frac{D - \omega_0}{S_D} = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \omega_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

as test statistic.

Under the null hypothesis, V can be approximated by a $t-$distribution with f degrees of freedom calculated as follows:

$$f = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{S_2^2}{n_2}\right)^2}.$$

In both cases (homogenous and heterogeneous variances) critical values can be taken from the $t-$distribution table. Table 9.7 shows the derived decision regions for the three test situations (for significance level α).

Note that the $t-$distribution quantiles in Table 9.7 can be approximated by standard normal quantiles, if both sample sizes n_1 and n_2 are big enough to justify the application of the central limit theorem ($n_1 > 30$ *and* $n_2 > 30$). The resulting decision regions are then similar to those in the case of known variances.

Sampling and Computing the Test Statistic

On the basis of an observed sample, the two sample means $\bar{x}_1$ and $\bar{x}_2$ and, if needed, the empirical standard deviations s_1 and s_2 can be computed. Plugging these values into the test statistic formula gives the realized test statistic value v.

Test Decision and Interpretation

Test decision and interpretation are carried out analogously to the one-sample mean test.

Explained: Testing the Difference of Two Population Means

Student Sabine visits two farms to buy fresh eggs. The farms are populated by two different breeds of hens—one on each. Sabine randomly picks 10 eggs from the first and 15 eggs from the second farm. Back home, she has the impression that the eggs produced by the hens on the first farm are heavier than those from the second. To verify this suspicion, she conducts a statistical test at a significance level of α. Sabine compares two (weight) averages by testing for the difference $\mu_1 - \mu_2$ of two means.

Hypothesis

As Sabine has reason to believe that the average weight of one egg variety is greater than that of the other, a single-sided test is indicated. She wants to prove statistically, that the first farm produces heavier eggs and consequently puts her conjecture as alternative hypothesis, hoping that her sample will reject the null hypothesis which states the negation of the statement she wants to verify positively. But Sabine has no idea as to how great the average weight difference could be and thus sets the hypothetical difference that has to be exceeded to prove her right to zero: $\mu_1 - \mu_2 = \omega_0 = 0$. She can formalize her test as

$$\mathrm{H}_0 : \mu_1 - \mu_2 \leq 0 \quad \text{versus} \quad \mathrm{H}_1 : \mu_1 - \mu_2 > 0,$$

or, equivalently,

$$\mathrm{H}_0 : \mu_1 \leq \mu_2 \quad \text{versus} \quad \mathrm{H}_1 : \mu_1 > \mu_2.$$

Test Statistic and Its Distribution: Decision Regions

Sabine has picked the eggs at random—in particular, she hasn't tried to get hold of the biggest ones on either farm. Naturally, she sampled without replacement, but we must also assume that the population of daily produced eggs on both farms is sufficiently large to justify the assumption of a simple random sample. Clearly, Sabine has drawn the samples independently, for she sampled on two unrelated farms.

Sabine assumes that the random variables X_1 : "egg weight of first breed" and X_2 : "egg weight of second breed" are normally distributed: $X_1 \sim N(\mu_1; \sigma_1)$ and $X_2 \sim N(\mu_2; \sigma_2)$. Expectations $E(X_1) = \mu_1$ and $E(X_2) = \mu_2$ and variances $Var(X_1) = \sigma_1^2$ and $Var(X_2) = \sigma_2^2$ are unknown. To simplify matters, Sabine assumes that the population variances are homogenous: $\sigma_1^2 = \sigma_2^2$. This assumption implies that a differential in the expectation doesn't induce a differential in the variances—a rather adventurous assumption. Nevertheless, acknowledging the

above assumptions (and the possibility of their violation), Sabine can base her test on the test statistic

$$V = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \omega_0}{\sqrt{\frac{n_1+n_2}{n_1\, n_2}\, \frac{(n_1-1)\, S_1^2+(n_2-1)\cdot S_2^2}{n_1+n_2-2}}}.$$

Here, $n_1 = 10$ and $n_2 = 15$ are the sample sizes, $\overline{X}_1$ and $\overline{X}_2$ are the sample means and S_1^1 and S_1^2 are the estimators of σ_1^2 and σ_2^2. Under H_0, V has $t-$distribution with $n_1 + n_2 - 2 = 10 + 15 - 2 = 23$ degrees of freedom. In the corresponding t-table we find the quantile $t_{0.95;23} = 1.714$ to be the critical value c satisfying $P(V \leq c) = 1 - \alpha = 0.95$ and hence have the following decision regions:

$$\text{Non-rejection region for } H_0 : \{v \mid v \leq 1.714\}.$$
$$\text{Rejection region for } H_0 : \{v \mid v > 1.714\}.$$

Sampling and Computing the Test Statistic

Sabine weighs the eggs and computes the sample-specific arithmetic averages and variances:

1^{st} *breed:*

$$\bar{x}_1 = 65.700 \quad s_1^2 = 50.35.$$

2^{nd} *breed:*

$$\bar{x}_2 = 60.433 \quad s_1^2 = 42.46.$$

Using $\omega_0 = 0$ she calculates a test statistic value of $v = 1.91$.

Test Decision and Interpretation

The test statistic realization $v = 1.91$ falls into the rejection region for H_0. Thus, Sabine couldn't prove statistically on the basis of two independent random samples of sizes $n_1 = 10$ and $n_2 = 15$ and a significance level of $\alpha = 0.05$, that the difference $\mu_1 - \mu_2$ of the population averages of the eggs' weights is significantly negative. As the type I error probability $P(\text{'H}_1\text{'} \mid H_0)$ cannot exceed α, Sabine has scientific backing for her claim that the eggs from breed 1 hens are heavier than those from the second farm—on average!

Enhanced: Average Age Difference of Female and Male Bank Employees

Mr. Schmidt and Mr. Maier, two senior bank managers, enjoy lunch hours that are long enough to start arguing about the average age of their colleagues.

1st Dispute

Mr. Schmidt claims that the average age of female employees differs from that of the male employees—an opinion Mr. Maier cannot and, more importantly, doesn't want to share.

2nd Dispute

Mr. Schmidt even believes to know the direction of the deviation: Female workers are older on average, it appears to him. Being opposed to Schmidt's first claim, Maier cannot but disagree with his second.

3rd Dispute

The above is not enough confrontation to override the boredom that has spread after numerous discussions about the fair value of the Euro and the best national football team coach. Mr. Schmidt cannot help himself and switches to attack: "On average, the women in our bank are 5 years older than the men!" Mr. Maier is more than happy to disagree, even though he suddenly concedes that the average male colleague might be younger than the average female. But he cannot rule out the possibility that these subjective impressions could be subject to a focus bias arising from a more critical examination of their female colleagues (Maier and Schmidt are both married).

To settle their disputes and hence make space for other future discussions, Maier and Schmidt decide to carry out a statistical investigation. They are both surprised that they can agree on the following settings:

The statistical test will be based on the difference of two population means $\mu_1 - \mu_2$; significance level is α.

Random variable X_1 captures the age of a female banker, X_2 the age of a male banker. Expectations $E(X_1) = \mu_1$, $E(X_2) = \mu_2$ and variances $Var(X_1) = \sigma_1$, $Var(X_2) = \sigma_2$ are unknown. *Homogeneity of variances* cannot be assumed, Maier and Schmidt agree. Furthermore, there is no prior knowledge about the shape of the distribution of X_1 and X_2 . Consequently, sample sizes n_1 and n_2 will have to be sufficiently large to justify the application of the central limit theorem. Maier and Schmidt know that there are approximately as many female as male workers in the

bank, and they thus choose equal sample sizes: $n_1 = n_2 = 50$. They ask human resources for support in there ground-breaking investigations. Of course, personnel could simply provide them with the exact data, but they agree to draw two samples of size 50 at random, without replacing the sampled entity after each draw. They assure that the two samples from the male and female population can be regarded as independent. Sample averages and variances are computed for both samples.

Test Statistic and Its Distribution: Decision Regions

As σ_1 and σ_2 are unknown and Maier&Schmidt have to assume *heterogeneity of variances*, they employ the test statistic

$$V = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \omega_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}},$$

where

$$\overline{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i}, \quad \overline{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i}$$

are the sample means and

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(X_{1i} - \overline{X}_1\right)^2, \quad S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}\left(X_{2i} - \overline{X}_2\right)^2$$

are estimators of the population variances σ_1 and σ_2.

As the sample sizes satisfy $n_1 > 30$ respectively $n_2 > 30$, the central limit theorem can be applied, and the distribution of V can, under H_0, be approximated by the standard normal distribution (bell curve). Maier&Schmidt will thus apply an asymptotic or approximate test for $\mu_1 - \mu_2$.

1*st Dispute*

Hypothesis

Mr. Schmidt's first claim is general in that he doesn't specify direction or size of the proposed average age differential. Thus, a two-sided test with $\omega_0 = 0$ has to be specified:

$$\mathrm{H}_0 : \mu_1 - \mu_2 = \omega_0 = 0 \quad \text{versus} \quad \mathrm{H}_1 : \mu_1 - \mu_2 \neq \omega_0 = 0,$$

or, equivalently,

$$\mathrm{H}_0 : \mu_1 = \mu_2 \quad \text{versus} \quad \mathrm{H}_1 : \mu_1 \neq \mu_2.$$

Determining the Decision Regions for H_0

The upper critical value satisfying $P(V \geq c_u) = 1 - \alpha/2 = 0.975$ can be looked up in the normal distribution table as the 97.5 % quantile: $c_u = z_{0.975} = 1.96$. From the symmetry of the normal distribution around zero follows for the lower critical value $c_l = -z_{1-\alpha/2} = -1.96$, such that $P(V \leq c_l) = \alpha/2 = 0.025$. We thus have the following decision regions:

Approximate non-rejection region for H_0:

$$\{v \mid -1.96 \leq v \leq 1.96\}.$$

Approximate rejection region for H_0:

$$\{v \mid v < -1.96 \text{ or } v > 1.96\}.$$

Sampling and Computing the Test Statistic

Personnel submits the following data computed from the two samples:

- Female bank clerks: $\bar{x}_1 = 47.71, \quad s_1^2 = 260.875$.
- Male bank clerks: $\bar{x}_2 = 41.80, \quad s_2^2 = 237.681$.
- Using $\omega_0 = 0$, Maier&Schmidt derive a test statistic value of $v = 1.87$.

Test Decision and Interpretation

The test statistic value of $v = 1.87$ falls into the non-rejection region for H_0, and consequently the null hypothesis is not rejected. Based on two independent random samples of sizes $n_1 = n_2 = 50$, Maier&Schmidt couldn't prove statistically the existence of a significant difference in the population averages of female and male bank clerks' ages, μ_1 and μ_2.

Having not-rejected the null hypothesis, Maier&Schmidt may have made a wrong decision. This is the case, if in reality the two population means *do* differ. The probability of the occurrence of a type II error ('H_0'$|\mathrm{H}_1$) can only be computed for "hypothetical" true parameter values, i.e., the parameter region of the alternative hypothesis is narrowed to a single parameter point.

2nd Dispute

Hypothesis

Mr. Schmidt believes that subsequently he has come up with some substantial new arguments in favor of his proposition and insists in putting it as the alternative hypothesis in a further test to be conducted. If the null hypothesis is rejected and thus his hypothesis verified, he can quantify the maximum type I error probability to be α and has thus scientific backing for maintaining his position. The resulting test is a right-sided one, still without quantification of the suggested positive difference: $\omega_0 = 0$:

$$\mathrm{H}_0 : \mu_1 - \mu_2 \leq \omega_0 = 0 \quad \text{versus} \quad \mathrm{H}_1 : \mu_1 - \mu_2 > \omega_0 = 0,$$

or, equivalently,

$$\mathrm{H}_0 : \mu_1 \leq \mu_2 \quad \text{versus} \quad \mathrm{H}_1 : \mu_1 > \mu_2.$$

Determining the Decision Regions for $\mathrm{H_0}$

The critical value satisfying $P(V \leq c) = 1 - \alpha = 0.95$ can be found in the normal distribution table to be $c = z_{0.95} = 1.645$. The decision regions are then:

Approximative non-rejection region for H_0:

$$\{v \mid v \leq 1.645\}.$$

Approximative rejection region for H_0:

$$\{v \mid v > 1.645\}.$$

Sampling and Computing the Test Statistic

Human resources supplies Mr. Maier and Mr. Schmidt with the following sample characteristics:

- Female bank clerks: $\bar{x}_1 = 51.71, \quad s_1^2 = 385.509.$
- Male bank clerks: $\bar{x}_2 = 45.16, \quad s_2^2 = 283.985$
- Using $\omega_0 = 0$, Maier&Schmidt compute the test statistic value as $v = 1.79$.

Test Decision and Interpretation

As the test statistic value of $v = 1.87$ falls into the rejection region for H_0, the null hypothesis is rejected. Maier&Schmidt could show on the basis of two independent

random samples of sizes $n_1 = n_2 = 50$, that the difference $\mu_1 - \mu_2$ is significant at the $\alpha = 0.05$ level. Thus, Schmidt has reason to maintain his claim that the average female bank clerk is older than the average male.

The probability of having made a wrong conclusion in a repeated test context, i.e., the type I error probability $P(\text{'H}_1\text{'}|\text{H}_0)$, is constrained by the significance level $\alpha = 0.05$.

Compared to the two-sided test, the rejection region for H_0 doesn't consist of two segments, but is located on the right-hand side of $E(V) = 0$. As the area under the normal curve corresponding to this region has to equal the "entire" quantity α, the critical value is smaller than that for the two-sided version. For this reason the null hypothesis is more likely to be rejected for the same significance level α and sample sizes n_1 and n_2 in the one-sided test than in the two-sided test for equal deviations of the test statistic from the hypothetical boundary parameter value in the same direction.

3rd Dispute

Hypothesis

In his third claim, Mr. Schmidt has gone one step further in that he has quantified the average age of his female colleagues to be at least 5 years higher than the average age of his male coworkers. Translated into our test formalization, the hypothetical difference is $\omega_0 = 5$. Maier agrees to adopt the same test structure as in the second dispute, leaving Schmidt's claim as alternative hypothesis. The resulting right-sided test is:

$$\text{H}_0 : \mu_1 - \mu_2 \leq \omega_0 = 5 \quad \text{versus} \quad \text{H}_1 : \mu_1 - \mu_2 > \omega_0 = 5.$$

Determining the Decision Regions for H_0

The critical value for $P(V \leq c) = 1 - \alpha = 0.95$ is looked up in the normal distribution table: $c = z_{0.95} = 1.645$. The resulting approximate decision regions are the same as in the second dispute:

Approximative non-rejection region for H_0:

$$\{v \mid v \leq 1.645\}.$$

Approximative rejection region for H_0:

$$\{v \mid v > 1.645\}.$$

Sampling and Computing the Test Statistic

Human resources submit the following statistics:

- Female bank clerks: $\bar{x}_1 = 52.22, \quad s_1^2 = 321.914$.
- Male bank clerks: $\bar{x}_2 = 43.13, \quad s_2^2 = 306.527$
- This time Maier&Schmidt compute the test statistic value using $\omega_0 = 5$, yielding $v = 1.154$.

Test Decision and Interpretation

The test statistic value $v = 1.154$ belongs to the non-rejection region for H_0, and the null hypothesis is thus not rejected. On the basis of two independent random samples of sizes $n_1 = n_2 = 50$, Maier&Schmidt couldn't verify statistically, that the difference $\mu_1 - \mu_2$ is significantly greater than 5. Schmidt hence couldn't prove statistically at a significance level of $\alpha = 0.05$, that the average female bank clerk is 5 years older than the average male bank worker. The test delivers an objective decision basis for a proposed difference of exactly 5—nothing can be said about any other positive difference smaller than 5 (neither for true differences greater than 5, owing to the possibility of the type II error). Thus, if the average female banker is older than the average male banker in the population, Mr. Schmidt has either overstated the difference or is a victim of the type II error, 'H_0'|H_1, the probability of which can only be computed for specific values of the true population parameter differential.

Interactive: Testing the Difference of Two Population Means

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please select

- the significance level α
- the sample size for group one n_1 and two n_2

Use "Draw sample" to manually draw a sample and carry out a test (Fig. 9.23).

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

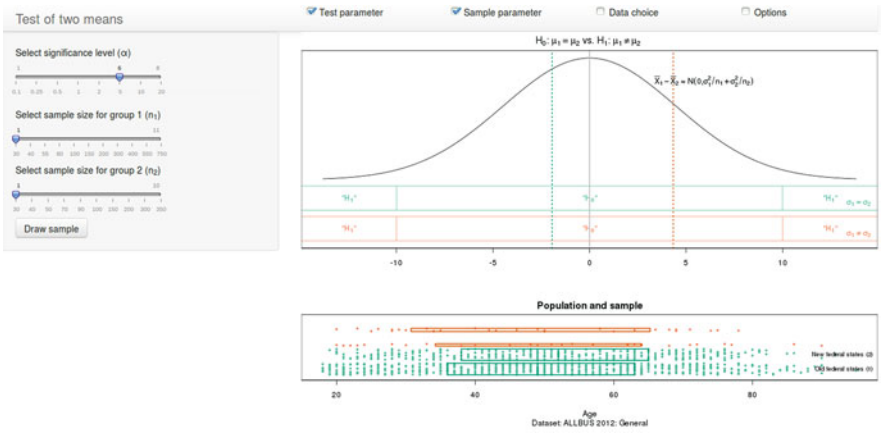


Fig. 9.23 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_tmu2

Output

In this interactive example you can study the impact of the significance level α and the sample sizes n_1, n_2 on the test decision of a two-sided test:

$$\mathrm{H}_0 : \mu_1 = \mu_2 \quad \text{versus} \quad \mathrm{H}_0 : \mu_1 \neq \mu_2.$$

You may conduct this test as often as you like. Each repetition is based on freshly simulated random samples of X_1 and X_2 and carried out using your specified test parameters. You can:

- repeatedly observe test decisions on the basis of unchanged significance level α and sample sizes n_1 and n_2;
- alter α, for constant n_1 and n_2;
- vary the sample sizes n_1 and n_2, holding the significance level α constant; or
- vary α, n_1 and n_2 simultaneously.

To facilitate easy diagnostics you are shown a table containing all test values and decisions.

9.5 Chi-Square Goodness-of-Fit Test

The chi-square goodness-of-fit test allows us to test the unknown population distribution of a random variable X. In the test procedures we have introduced so far we have assumed that the distribution of X can be described (at least approximately) by a function that is specified up to some parameter (e.g., μ and σ, or π and n). Our tests were "merely" designed to verify whether certain hypothetical values for these

unknown parameters can be reconciled with the sample at hand. Our goal now is to verify whether the data can be fit by a *fully specified* probability model. That means that there aren't any parameters left to test for, and we are thus moving on from a parametric to a nonparametric approach. The chi-square goodness-of-fit test is based on a simple random sample. As usual the significance level α has to be fixed before the test is conducted. Note that the chi-square test represents only one approach to testing the fit of a probability model. Consult the literature for other tests.

A random variable X has probability distribution $F(x)$. No restrictions are imposed on the measurement level of X. The probability distribution is unknown, but there is a hypothesis about it, denoted by $F_0(x)$.

If X is a discrete random variable, we denote the set of possible outcomes by $x_1, \ldots, x_k$. We define:

- $h(x_j) = h_j$ is the observed absolute frequency of x_j in the sample, $j = 1, \ldots, k$,
- $P(X = x_j)$ is the probability of X assuming the value x_j, $j = 1, \ldots, k$.

If X is a continuous random variable (which we understand to include quasi-continuous variables, i.e., discrete variables with infinitely many possible realization), we have to partition the set of possible outcomes. If $k \geq 2$ is the number of classes, the classes are given by the following exhaustive sequence of disjoint intervals:

$$\left(x_0^*, x_1^*\right], \left(x_1^*, x_2^*\right], \ldots, \left(x_{k-1}^*, x_k^*\right] \text{ respectively } \left(x_{j-1}^*, x_j^*\right], \quad j = 1, \ldots, k.$$

We define for the continuous case:

- $h\left(x_{j-1}^* < X \leq x_j^*\right) = h_j$ is the observed absolute frequency in the j^{th} class in the sample, $j = 1, \ldots, k$,
- $P\left(x_{j-1}^* < X \leq x_j^*\right)$ is the probability of X assuming values within the j^{th} class, $\left(x_{j-1}^*, x_j^*\right], j = 1, \ldots, k$.

Hypothesis

The null hypothesis in a goodness-of-fit test states that the proposed probability model correctly describes the distribution of the data in the population; the alternative hypothesis contains the negation of this statement. Applied to the chi-square test using above conventions, the test is formalized as follows:

Discrete X

$$\mathrm{H}_0 : P(X = x_j) = p_j \quad \forall j = 1, \ldots, k$$

versus

$$\mathrm{H}_1 : P(X = x_j) \neq p_j \quad \text{for at least one } j.$$

Continuous X

$$\mathrm{H}_0 : P\left(x_{j-1}^* < X \leq x_j^*\right) = p_j \quad \forall j = 1, \ldots, k$$

versus

$$\mathrm{H}_1 : P\left(x_{j-1}^* < X \leq x_j^*\right) \neq p_j \quad \text{for at least one } j.$$

In both cases p_j denotes the probability of X assuming the value x_j (or falling into the jth class, $\left(x_{j-1}^*, x_j^*\right]$), given the null hypothesis is true and hence $F_0(x)$ is the true probability distribution:

$$p_j = P\left(X = x_j \,|\, \mathrm{H}_0\right) \quad \text{respectively} \quad p_j = P\left(x_{j-1}^* < X \leq x_j^* \,|\, \mathrm{H}_0\right).$$

How Is p_j Computed?

Fully Specified Parametric Distribution Function
The quantities p_j can be readily calculated if the hypothetical distribution is a fully specified function. If $F_0(x)$ is a member of some parametric class, all parameters have to be known.

Example X has Poisson distribution $PO(\lambda)$ with given parameter λ.

Partially Specified Parametric Distribution Function
If the hypothetical distribution belongs to a parametric family involving one or more parameters, of which at least one is unknown, they will have to be estimated before the p_j's can be calculated.

Example We want to test whether X has a normal distribution $N(\mu, \sigma)$, where expectation μ and variance σ are unknown. We will have to estimate these parameters using the information conveyed by the sample to obtain a completely specified distribution function and calculate the hypothetical probabilities p_j.

Frequency Distribution
The null hypothesis may state the hypothetical probability model in the form of a numerical frequency distribution.

Example The random variable X can take on 4 possible values with associated probabilities $p_1 = 0.2$, $p_1 = 0.4$, $p_1 = 0.1$ and $p_1 = 0.3$.

Test Statistic and Its Distribution: Decision Regions

The tests' principle is to compare the hypothetical probabilities derived from the hypothetical distribution stated in the null hypothesis with observed relative

frequencies. The underlying test statistic is based on observed absolute frequencies h_j. Once we have drawn a sample, of size n, we can calculate them as frequencies of the events $\{X = x_j\}$ respectively $\left\{x^*_{j-1} < X \leq x^*_j\right\}$. The set of all absolute frequencies $h_j, j = 1, \ldots, k$ constitutes the samples' distribution. They are random, because by sampling randomly we carry out a random experiment. We must therefore regard the absolute frequencies h_j as realizations of random variables $H_j, j = 1, \ldots, k$.

If the *null hypothesis is true*, the expected values of the relative frequencies in the sample are given by the probabilities p_j. The expectations of the absolute frequencies are thus np_j.

The comparison between observed and expected (under H_0) frequencies is constructed around the differences $H_j - np_j$, $j = 1, \ldots, k$. Small differences count in favor of the null hypothesis. A way of consolidating the differences across possible outcomes/classes is expressed by the following test statistic:

$$V = \sum_{j=1}^{k} \frac{\left(H_j - np_j\right)^2}{np_j}.$$

Under H_0, V has approximately chi-square distribution with $k - m - 1$ degrees of freedom—independent of the distribution that is being tested for.

Approximation Conditions

The approximation can be assumed to be sufficiently accurate if

- $np_j \geq 1$ for all j *and*
- $np_j \geq 5$ for at least 80 % of all expected absolute frequencies.

A means of ensuring the applicability of the chi-square goodness-of-fit test when these conditions aren't fulfilled in the original setting is to combine bordering classes or outcomes into larger classes. As the hypothetical probabilities p_j are fixed (through the null hypothesis), an increase in the sample size n will always result in an improved approximation accuracy.

In determining the degrees of freedom we have to take into account:

- k is the number of classes *after* a possibly necessary combination of classes,
- m is the number of parameters that has to be estimated from the sample in order to fully specify the distribution. If the probability distribution proposed in H_0 is completely specified, m is zero.

Observe that $\left(H_j - np_j\right)^2 / np_j$ cannot be negative. The test statistic v as the sum of these ratios can thus assume only positive values. Large (absolute) deviations $H_j - np_j$ are translated into high positive contributions to the test statistic value,

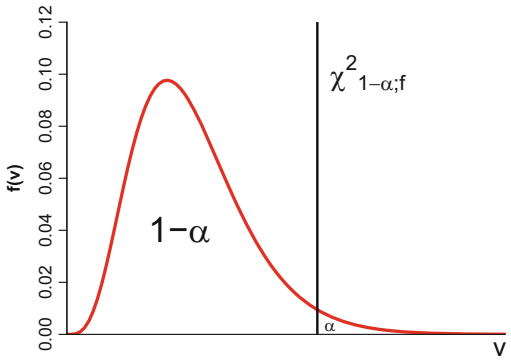


Fig. 9.24 Distribution for the chi-square goodness-of-fit test

increasing the likelihood of a rejection for H_0. Small deviations are more likely to stem from the noise introduced by the sampling process rather than being the result of a distribution function fundamentally (i.e., significantly) different from the hypothetical one. They thus tend to support the null hypothesis. Small deviations are marginalized through the square operator. As only relatively high values of the test statistic lead to a rejection of the null hypothesis, the chi-square test is a right-sided test. The *critical value* c satisfying $P(V \leq c_u) = 1 - \alpha$ for the given degrees of freedom is taken from the chi-square *distribution function* table. The decision regions are thus:

$$\text{Rejection region for } H_0 : \left\{v \mid v > \chi^2_{1-\alpha;k-m-1}\right\}.$$

$$\text{Non-rejection region for } H_0 : \left\{v \mid v \leq \chi^2_{1-\alpha;k-m-1}\right\}.$$

The probability of V assuming a value from the rejection region for H_0, given H_0 is true, equals the significance level: $\alpha = P\left(V > \chi^2_{1-\alpha;k-m-1} \mid H_0\right)$. The probability of V falling into the non-rejection region under H_0 is $P(V \leq \chi^2_{1-\alpha;k-m-1} | H_0) = 1-\alpha$ (Fig. 9.24).

Sampling and Computing the Test Statistic

Once a random sample of size n has been observed, the absolute frequencies h_j can be computed. If needed, unknown parameters in the hypothetical distribution can be estimated and the expected absolute frequencies, np_j, can be calculated. Plugging this condensed data into the test statistic formula gives the test statistic value.

Test Decision and Interpretation

If v belongs to the rejection region for H_0, the null hypothesis will be rejected on the basis of a random sample of size n and a significance level α: 'H_1'. In this case the researcher could show statistically, that the population distribution of the random variable X is not given by $F_0(x)$.

Rejecting the null hypothesis makes the researchers' conclusions subject to the risk of a type I error: 'H_1'|H_0. Fortunately, that's a quantity the empirical scientist can control: the probability of the null hypothesis being true, given it has been rejected, cannot—by construction—exceed the significance level α.

If v is observed within the non-rejection region, the null hypothesis is not rejected on the basis of a particular sample of size n for a given significance level α: 'H_0'. One could not verify statistically that the true population distribution generating the data differs significantly from the hypothetical distribution $F_0(x)$.

Of course, this decision doesn't imply, that the true distribution *does* coincide with the proposed one. The actual sample merely couldn't falsify this possibility, and in a certain number of samples it won't do so even though the null hypothesis is not true. Such a case is an example of a type II error: 'H_0'|H_1.

More Information

In principle the general approach used in goodness-of-fit tests resembles that of parametric tests. A test statistic is constructed summarizing or condensing the information about the hypothetical distribution and that conveyed by the sample to form the basis for a probabilistic statement about the null hypothesis. The test statistic distribution has to be derived (at least approximately) under the null hypothesis. Thus, the decision about the non-rejection or rejection of a probability model to describe the real-world data generating process (a nonparametric test) is subject to the same possible errors as in parametric tests: In repeated samples (tests), a type I error will be made with (conditional) probability $P(\text{'}H_0\text{'}|H_1) = \alpha$, if H_0 has been not-rejected and a type II error will be made with probability $P(\text{'}H_1\text{'}|H_0) = \beta$ if it has been rejected. The α error probability is controlled by the researcher through the significance level α, but the type II error probability cannot be computed, for it is not clear what the alternative probability model is—we only know that it is *not* the one specified in the null hypothesis, but one can make up infinitely many models that are arbitrarily close to the hypothetical one. It should thus always be the goal of the researcher to reject the null hypothesis, as this caps the probability of making a wrong decision. On the other hand it's not possible to "accept" the null hypothesis (hypothetical model)—we either reject it or do not reject it—non-rejection does not imply that the null is necessarily true (recall we may have experienced a type II error).

Hypothesis

If the hypothetical distribution is the true distribution actually generating the data throughout the population, we expect to encounter this distributional pattern in the sample. As the sample is a randomly chosen subset of the population, it will more or less accurately reflect the true population pattern, and only on average will the samples (in a large-sample context) reveal the true (and correctly "guessed") distribution. Deviations of the empirical distribution encountered in the sample will then be a result of noise introduced by the sampling process (due the fact, that only a limited number of statistical elements making up the entire population is represented in the sample). Statistical goodness-of-fit tests are designed to reliably discriminate between this sampling noise and deviations of the hypothetical distributions from the actual. The "reliability" concept is based on a repeated sampling context—on average, we want the test to discriminate properly, as there is always a (albeit small) probability of drawing a sample that is atypical (or nonrepresentative) for the underlying true distribution. The question that the chi-square goodness-of-fit test tries to answer is thus whether the encountered deviation of the empirical from the hypothetical (theoretical) distribution is significant in that it exceeds the average sample noise expected for the given sample size n. The pair of hypotheses is always:

- H_0 : The distribution of the random variable in the population is the hypothetical one.
- H_1 : The distribution of the random variable in the population differs from the hypothetical one.

As already mentioned, large deviations of the sample distribution from the hypothetical distribution tend to falsify the null hypothesis, indicating that a different distribution is governing the population data.

The pair of hypotheses underlying the chi-square goodness-of-fit test contains the probabilities $p_j\,(j = 1, \ldots, k)$, which are calculated from the hypothetical distribution. If X is a discrete random variable, the probabilities $p_j = P\left(X = x_j \mid \mathrm{H}_0\right)$ are explicitly given with the probability function. In the case of continuous random variables the probability of one specific value having been realized is always zero. For this reason intervals have to be formed, within which realizations can be observed. The probability $p_j = P\left(x_{j-1}^* < X \leq x_j^* \mid \mathrm{H}_0\right)$, that the continuous random variable X assumes values in the interval (class) $\left(x_{j-1}^*, x_j^*\right]$ can be calculated from the given probability density function. Note that it may be necessary to group (quasi-continuous) discrete variables into classes—if only for the sake of improving the approximation accuracy of the chi-square distribution.

Test Statistic

We will now illustrate the random nature of the observed absolute frequencies H_j. Our reasoning is valid both for continuous and discrete variables, but we will refer to the discrete case for simplicity.

Suppose we randomly pick a statistical element from the population of all elements (objects/subjects) displaying the random variable X under consideration. If we want to compute the absolute frequency H_j for one specific outcome x_j, the only information we are interested in is whether X assumes this value on that particular element or not. Thus, there are only two possible outcomes: Under H_0, the probability of X being observed in x_j is p_j, and the probability of this element not counting towards the absolute frequency H_j is $1 - p_j$. Drawing a sample of size n means to independently repeat this random experiment n times. As the hypothetical distribution and therefore the derived quantity p_j remains unchanged, we are carrying out a Bernoulli experiment, if we focus on one single absolute frequency H_j.

Having repeated the Bernoulli experiment n times, we are interested in the overall number of realization of $\{X = x_j\}$, i.e., the absolute frequency of x_j in the sample. This frequency can (and most certainly will) vary across samples. Hence, H_j : Number of observations $X = x_j$ in a simple random sample of size n is a discrete random variable with possible outcomes $0, \ldots, n$. More specifically, under H_0 the random variable H_j has Binomial distribution with parameters n and p_j: $H_j \sim B(n; p_j)$. Its expectation is given by $E(H_j) = np_j$, the expected absolute frequency $\{X = x_j\}$ under the null hypothesis. The variance $Var(H_j) = np_j(1 - p_j)$ captures the variation in the observed absolute frequency of $\{X = x_j\}$.

The test statistic is based on deviations of the random variables from their expectation: $H_j - np_j$. Summing over these quantities would result in negative and positive deviations offsetting each other. Squaring the terms before summation prevents from that: $(H_j - np_j)^2$. Dividing by the sample size n and the probabilities p_j weights the squared deviations by their "importance" in terms of contribution to the overall probability distribution. A difference $h_j - np_j = 5$ receives a higher weighting for $np_j = 10$ than for $np_j = 100$—for a fixed sample size, the difference accounts for a higher proportion in the test statistic, if the value x_j is expected with low probability and thus accounts for a smaller part of the distribution in terms of probability (the tails of the distribution for example). These considerations apply for all $j = 1, \ldots, k$.

Summation over all normalized deviations consolidates the overall deviation of the empirical distribution function from the hypothetical, yielding an adequate test statistic with a known asymptotic distribution:

$$V = \sum_{j=1}^{k} \frac{(H_j - np_j)^2}{np_j}.$$

Because H_j, $j = 1, \ldots, k$ are random variables, V is also a random variable. When n is sufficiently large and the approximation conditions on the np_j hold, V is approximately chi-square distributed with $k - m - 1$ degrees of freedom under the null hypothesis, regardless of the shape of the hypothetical distribution. If the approximation conditions aren't fulfilled, combining classes may provide a fix. This may require the construction of classes for discrete random variables (or a broadening of one or more classes, if the data has already been grouped). Since we have summarized, or condensed, our data into k classes we have k pieces of information to provide information about the null. The '-1' term in the formula for the degrees of freedom reflects the fact, that any absolute frequency h_j is determined by the other $k - 1$ frequencies, as the overall number of absolute frequencies must satisfy $\sum_j h_j = n$. The absolute frequencies are thus not (linearly) independent of each other. The need for estimation of parameters in the hypothetical distribution $F_0(x)$ results in a further loss of degrees of freedom. If m is the number of parameters to be estimated, we have $k - m - 1$ degrees of freedom.

Explained: Conducting a Chi-Square Goodness-of-Fit Test

A die is claimed to be fair. We want to verify this statement using a chi-square goodness-of-fit test at a significance level of $\alpha = 0.1$. The size of the sample is $n = 240$.

Hypothesis

The random variable we are dealing with, X : "Number on top of the die," is a discrete variable that can assume the values $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4, x_5 = 5$ and $x_6 = 6$. Its distribution $F(x)$ is unknown, but the null hypothesis is that the die is fair and hence all outcomes (sides) are equally probable. Hence, the null hypothesis states that X has discrete uniform distribution:

$$\mathrm{H}_0 : P\left(X = x_j\right) = p_j = 1/6\,, \quad \forall j = 1, \ldots, 6$$

vs.

$$\mathrm{H}_1 : P\left(X = x_j\right) = p_j \neq 1/6\,, \quad \text{for at least one (and hence at least one further) } j.$$

Test Statistic and Its Distribution: Decision Regions

We use the test statistic for the chi-square goodness-of-fit test:

$$V = \sum_{j=1}^{k} \frac{\left(H_j - np_j\right)^2}{np_j}.$$

Under H_0, V is approximately chi-square distributed. The approximation conditions are fulfilled, as $np_j = 40 > 5$ for all $j = 1, \ldots, 6$. The discrete uniform distribution is fully specified and there are hence no parameters to be estimated ($m = 0$). We thus have $k - m - 1 = 6 - 0 - 1 = 5$ degrees of freedom.

Looking up the critical value c for which $P(V \leq c) = 1 - \alpha = 0.9$ in the table of the chi-square distribution with 5 degrees of freedom gives $c = \chi^2_{1-\alpha;k-m-1} = \chi^2_{0.90;5} = 9.24$. The resulting decision regions are:

Rejection region for H_0:

$$\{v \mid v > 9.24\}.$$

Non-rejection region for H_0:

$$\{v \mid v \leq 9.24\}.$$

Sampling and Computing the Test Statistic

The die is rolled 240 times. The resulting sequence of observations constitutes a simple random sample, because the individual trials are independent of each other. Table 9.8 summarizes the data.

Take a look at the deviations of the observed frequencies from the frequencies expected under the null hypothesis. Can they be regarded as random variations around the expected value arising from the finite size of the sample? The test statistic value is given by the sum of the last column: $v = 9.8$.

Table 9.8 Data on 240 throws of a die

x_j	Observed frequencies h_j	Expected frequencies np_j	$h_j - np_j$	$\left(h_j - np_j\right)^2$	$\frac{\left(h_j - np_j\right)^2}{np_j}$
1	52	40	12	144	3.6
2	50	40	10	100	2.5
3	32	40	−8	64	1.6
4	36	40	−4	16	0.4
5	32	40	−8	64	1.6
6	38	40	−2	4	0.2

Test Decision and Interpretation

As the test statistic value falls into the rejection region for H_0, the null hypothesis is rejected ('H_1'). On the basis of a random sample of size $n = 240$ and a significance level $\alpha = 0.1$, we couldn't prove statistically that the die is fair, i.e., that the true probability distribution of X : "Number on top of the die" is the discrete uniform distribution. In repeated samples (tests) the probability of making a type I error, $P('H_1'|H_0)$, doesn't exceed a chosen significance level $\alpha = 0.1$ by construction. Therefore, any faith that we associate with this test conclusion (result) stems from the fact that we have conducted a test that is "accurate" on average.

Enhanced: Goodness-of-Fit Test for Product Demand

The management of a wholesaler analyzes the business. The focus is on demand for a certain specialized product. Which distribution can describe the variation in demand?

The demand for a product unfolds continuously in time. Customers place their orders independently of each other, and the distributor cannot trace back the individual orders to common underlying factors. As a consequence, the overall demand is a random phenomenon. Let's partition continuous time into intervals of one days' length. Then the random variable X denotes the discretely measured demand for the product under investigation. These settings suggest that the Poisson distribution may be a suitable model for the random variations in the demand: $X \sim PO(\lambda)$.

The test has to be conducted at a significance level of $\alpha = 0.05$. The data encountered in a simple random sample of size of $n = 50$ days is summarized in Table 9.9.

Table 9.9 Probabilities and expected absolute frequencies under H_0

j	Demand x_j	Observed frequencies hj	$p_j = P(X = x_j \mid H_0)$	$np_j \mid H_0$
1	0	3	0.1653	8.265
2	1	9	0.2975	14.875
3	2	14	0.2678	13.390
4	3	13	0.1607	8.035
5	4	6	0.0723	3.615
6	5	5	0.0260	1.300
7	≥ 6	0	0.0104	0.520

1*st Version*

Hypothesis

An experienced member of the staff believes that the average quantity sold in any 5 days period is 9. As the mean of the Poisson distribution is given by $E(X) = \lambda$ and we are observing in intervals of 1 day, we must scale the expectation to $\lambda = 9/5 = 1.8$. Our test is then

$$\mathrm{H}_0 : X \text{ has Poisson distribution with parameter } \lambda = 1.8, \text{ i.e., } X \sim PO(1.8)$$

vs.

$$\mathrm{H}_1 : X \text{ doesn't have Poisson distribution with parameter } \lambda = 1.8.$$

Columns 4 and 5 of Table 9.9 contain the probabilities under H_0, $P\left(X = x_j \mid \mathrm{H}_0\right) = p_j$ (taken from the $PO(1.8)$ table) and the associated expected absolute frequencies np_j.

Test Statistic and Its Distribution: Decision Regions

The test statistic for the chi-square goodness-if-fit test is :

$$V = \sum_{j=1}^{k} \frac{\left(H_j - np_j\right)^2}{np_j}.$$

Under H_0, V is asymptotically chi-square distributed with $k - m - 1$ degrees of freedom.

Are the approximation conditions satisfied? As you can see in the fifth column of Table 9.9, the realizations $x_5 = 4$ and $x_6 = 5$ do not satisfy $np_j \geq 5$. Realization $x_7 \geq 6$ doesn't even satisfy $np_j \geq 1$. We thus combine these three realization into one class.

Determining degrees of freedom:

There are $k = 5$ classes left after the re-grouping. The hypothetical Poisson distribution had been fully specified; the given parameter $\lambda = 1.8$ didn't have to be estimated: $m = 0$. Thus we have V has approximately chi-square distribution with $k - m - 1 = 5 - 0 - 1 = 4$ degrees of freedom.

We find the critical value c satisfying $P(V \leq c) = 1 - \alpha = 0.95$ in the table of the chi-square distribution with 4 degrees of freedom: $c = \chi^2_{1-\alpha;k-m-1} = \chi^2_{0.95;4} = 9.49$. The decision regions are:

Rejection region for H_0:

$$\{v \mid v > 9.49\}.$$

Table 9.10 Test statistic components for new grouping

x_j	hj	np_j	$h_j - np_j$	$(h_j - np_j)^2$	$(h_j - np_j)^2 / np_j$
0	3	8.265	−5.265	27.7202	3.3539
1	9	14.875	−5.875	34.5156	2.3204
2	14	13.390	0.610	0.3721	0.0278
3	13	8.035	4.965	24.6512	3.0680
≥ 4	11	5.435	5.565	30.9692	5.6981

Non-rejection region for H_0:

$$\{v \mid v \leq 9.49\}.$$

Calculating the Test Statistic Value

Table 9.10 summarizes the sample data in terms of test statistic components for the new grouping.

Summing over all five values in the last columns gives the realized (observed) test statistic value: $v = 14.4682$.

Test Decision and Interpretation

The test statistic value belongs to the rejection region for H_0; consequently, the null hypothesis is rejected ('H_1'). On the basis of a random sample of size $n = 50$ and a significance level $\alpha = 0.05$, we could prove statistically that the random variable X : "Daily demand for considered product" does *not* have Poisson distribution with parameter $\lambda = 1.8$. Note, this doesn't imply that we have to leave the class of Poisson distribution when searching for an appropriate probability model, for we have only tested for the assumed parameterization $\lambda = 1.8$.

Having decided in favor of the alternative hypothesis we may have made a type I error: 'H_1'|H_0. This is the case if $PO(1.8)$ *is* the true distribution of X. The probability of this happening in a repeated sampling (i.e., over many tests) is given by the significance level $\alpha = 0.05$.

2nd Version

Hypothesis

We maintain our assumption that the class of Poisson distributions is an adequate model for the demand: $X \sim PO(\lambda)$. This time we don't have any prior knowledge (or belief) about the parameter λ and thus have to estimate its value from the data.

Table 9.11 Probabilities and expected absolute frequencies under H_0

j	Demand x_j	Observed frequencies hj	$p_j = P(X = x_j \mid H_0)$	$np_j \mid H_0$
1	0	3	0.0821	4.105
2	1	9	0.2052	10.260
3	2	14	0.2565	12.825
4	3	13	0.2138	10.690
5	4	6	0.1336	6.680
6	5	5	0.0668	3.340
7	≥ 6	0	0.0420	2.100

We will use the sample of size $n = 50$ as in the first version. Applying the method of moments estimation principle, we can estimate $\lambda = E(X)$ with the first sample moment

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

The arithmetic mean in the observed sample is $\bar{x} = 125/50 = 2.5$, and we have the following pair of hypotheses:

$$H_0 : X \text{ has Poisson distribution with parameter } \lambda = 2.5, \text{ i.e., } X \sim PO(2.5)$$

vs.

$$H_1 : X \text{ doesn't have Poisson distribution with parameter } \lambda = 2.5.$$

In columns 4 and 5 of the Table 9.11 you find the probabilities implied by H_0: $P(X = x_j \mid H_0) = p_j$ (taken from the $PO(2.5)$ table) and the associated expected absolute frequencies np_j.

Test Statistic and Its Distribution: Decision Regions

Once again, we use the test statistic

$$V = \sum_{j=1}^{k} \frac{(H_j - np_j)^2}{np_j},$$

which we know is approximately chi-square distributed with $k - m - 1$ degrees of freedom.

Verifying the Approximation Conditions
As can be seen in the fifth column of Table 9.11, the realization $x_1 = 0$ doesn't satisfy the approximation condition $np_j \geq 5$. We fix this by combining it with the second realization x_2. The sixth and seventh possible outcomes ($h_6 = 5$, $h_7 \geq 6$) still aren't expected to be observed sufficiently frequently under the null hypothesis. We group them into one class.

Calculating Degrees of Freedom
After the re-grouping there are $k = 5$ classes left. The Poisson distribution parameter has been estimated, imposing a reduction of one: $m = 1$. Hence we have $k - m - 1 = 5 - 1 - 1 = 3$ degrees of freedom. V has approximately chi-square distribution with 3 degrees of freedom.

The value v satisfying $P(V \leq c) = 1 - \alpha = 0.95$ can be looked up in the table of the chi-square distribution with 3 degrees of freedom: $c = \chi^2_{1-\alpha;k-m-1} = \chi^2_{0.95;3} = 7.81$. The critical value specifies the decision regions:

Rejection region for H_0:

$$\{v \mid v > 7.81\}.$$

Non-rejection region for H_0:

$$\{v \mid v \leq 7.81\}.$$

Calculating the Test Statistic Value

Table 9.12 contains the sampled data in terms of test statistic components.

The test statistic value is computed by taking the sum of the last column: $v = 1.101$.

Table 9.12 Test statistic components for new grouping

x_j	hj	np_j	$h_j - np_j$	$(h_j - np_j)^2$	$(h_j - np_j)^2 / np_j$
0-1	12	14.365	−2.365	5.5932	0.3894
2	14	12.825	1.175	1.3806	0.1076
3	13	10.690	2.310	5.3361	0.4992
4	6	6.680	−.680	0.4624	0.0692
≥ 5	5	5.440	−0.440	0.1936	0.0356

Test Decision and Interpretation

As the test statistic value belongs to the non-rejection region for H_0, the null hypothesis is not rejected ('H_0'). On the basis of a random sample of size $n = 50$ and a significance level $\alpha = 0.05$, we could *not* prove statistically that the random variable X : "Daily demand for considered product" does *not* follow a Poisson distribution with parameter $\lambda = 2.5$, $PO(2.5)$.

We have made a type II error if the underlying isn't $PO(2.5)$ and thus the null hypothesis not true: 'H_0'$|H_1$. In repeated samples (i.e., over repeated tests), the probability of this error, P ('H_1'$|H_0$), is unknown.

9.6 Chi-Square Test of Independence

The chi-square test of independence allows us to test for statistical (stochastic) independence. It is a nonparametric test applicable to all measurement scales.

We assume that two random variables X and Y are observed simultaneously on $i = 1, \ldots, n$ statistical elements, the observed pairs being mutually independent (simple random sample). If X and Y are discrete random variables, they can be observed in the realization x_k, $k = 1, \ldots, K$ respectively y_j, $j = 1, \ldots, J$. If X and Y are continuous (including quasi-continuous discrete variables), the sample space has to be partitioned into disjoint exhaustive classes (intervals). In this case, x_k, $k = 1, \ldots, K$ and y_j, $j = 1, \ldots, J$ denote representative values within the classes (usually the class midpoints) and J and K denote the overall number of classes. A suitable representation of the observed joint *frequency distribution* is the two-dimensional frequency table, also known as *bivariate contingency table* (see Chap. 10 for additional material on contingency tables).

Here, h_{kj} denotes the absolute frequency of the observed pair $\left(x_k, y_j\right)$, i.e., that X assumes x_k or a value from the *kth* class, *and* Y assumes y_j or a value within the *jth* class:

$$h_{kj} = h\left(\{X = x_k\} \cap \left\{Y = y_j\right\}\right) ; \quad k = 1, \ldots, K\,, j = 1, \ldots, J.$$

The last column contains the observed *marginal distribution* (md) of X, composed of the absolute marginal frequencies $h_{k\bullet} = h\left(X = x_k\right) ; k = 1, \ldots, K$, denoting the frequencies with which X has been observed in x_k (discrete realization or class midpoint) regardless of the value of Y. In the last row you find the observed marginal distribution of Y, given by the absolute marginal frequencies $h_{j\bullet} = h\left(Y = y_j\right) ; j = 1, \ldots, J$, the frequencies of Y being observed in y_j regardless of X. The following definitions are used in the construction of the two-dimensional

Table 9.13 Two-dimensional contingency table

x	y	y_1	$\cdots$	y_j	$\cdots$	y_J	md x
x_1		h_{11}	$\cdots$	h_{1j}	$\cdots$	h_{1J}	$h_{1\bullet}$
$\vdots$		$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_k		h_{k1}	$\cdots$	h_{kj}	$\cdots$	h_{kJ}	$h_{k\bullet}$
$\vdots$		$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_K		h_{K1}	$\cdots$	h_{Kj}	$\cdots$	h_{KJ}	$h_{K\bullet}$
md x		$h_{\bullet 1}$	$\cdots$	$h_{\bullet j}$	$\cdots$	$h_{\bullet J}$	$h_{\bullet\bullet} = n$

contingency table (Table 9.13):

$$h_{k\bullet} = \sum_{j=1}^{J} h_{kj}\,; \quad k = 1, \ldots, K;$$

$$h_{\bullet j} = \sum_{k=1}^{K} h_{kj}\,; \quad j = 1, \ldots, J;$$

$$h_{\bullet\bullet} = \sum_{k=1}^{K} h_{k\bullet} = \sum_{j=1}^{J} h_{\bullet j} = \sum_{k=1}^{K}\sum_{j=1}^{J} h_{kj} = n.$$

Hypothesis

The null hypothesis in a chi-square test of independence states that X and Y are statistically (stochastically) independent; the alternative hypothesis negates this.

$$\mathrm{H}_0 : X \text{ and } Y \text{ are statistically independent}$$

vs.

$$\mathrm{H}_1 : X \text{ and } Y \text{ are } \textit{not} \text{ statistically independent.}$$

If the null hypothesis is true, the multiplication rule for independent events gives

$$P\left(X = x_k\right\} \cap \left\{Y = y_j\right) = P\left(X = x_k\right) \cdot P\left(Y = y_j\right) = p_{k\bullet} \cdot p_{\bullet j}.$$

In above formula,

- p_{kj} denotes the probability of X assuming x_k (or a value belonging to the class represented by x_k) *and* Y assuming y_j (or a value within the *jth* class),
- $p_{k\bullet}$ is the probability of X being observed in x_k respectively the *kth* class (marginal probabilities of X),

- $p_{\bullet j}$ is the probability that Y assumes the value x_k or is observed in the *jth* class (marginal probabilities of Y).

The pair of hypotheses can thus be written

$$\mathrm{H}_0 : p_{kj} = p_{k\bullet} \cdot p_{\bullet j} \quad \forall\,(k,j)$$

vs.

$$\mathrm{H}_1 : p_{kj} \neq p_{k\bullet} \cdot p_{\bullet j} \quad \text{for at least one pair } (k,j)\,.$$

As usual the significance level α and sample size n have to be fixed before the test is conducted.

Test Statistic and Its Distribution: Decision Regions

As the test is based on a comparison between observed absolute frequencies and absolute frequencies expected under the null hypothesis, the test statistic is built around absolute frequencies.

An observed sample is summarized in the bivariate contingency table in terms of joint absolute frequencies h_{kj} $(k = 1,\ldots,K\,,\, j = 1,\ldots J)$. These quantities are outcomes of a random experiment and thus vary across samples. They are realizations of their theoretical counterparts, the random variables denoted by H_{kj}.

If the null hypothesis is true, the expected joint frequencies are $e_{kj} = n \cdot p_{k\bullet} \cdot p_{\bullet j}$. The joint probabilities p_{kj} and marginal probabilities $p_{k\bullet}$ and $p_{\bullet j}$ are unknown and have to be estimated from the sample. Unbiased and consistent estimators for $p_{k\bullet}$ and $p_{\bullet j}$ are the relative marginal frequencies (sample proportions) $f_{k\bullet} = h_{k\bullet}/n$ and $f_{\bullet j} = h_{\bullet j}/n$. This implies, that we are assuming fixed marginal frequencies in the two-dimensional contingency table. Our estimators for the expected joint absolute frequencies under H_0 are given by

$$\hat{e}_{kj} = n \cdot f_{k\bullet} \cdot f_{\bullet j} = n \cdot \frac{h_{k\bullet}}{n} \cdot \frac{h_{\bullet j}}{n} = \frac{h_{k\bullet} \cdot h_{\bullet j}}{n}.$$

The comparison between the joint absolute frequencies encountered in the sample and those expected under the null hypothesis is based on the differences $H_{kj} - \hat{e}_{kj}$ $(k = 1,\ldots,K\,;j = 1,\ldots J)$. A test statistic weighting these differences is the sum

$$V = \sum_{k=1}^{K} \sum_{j=1}^{J} \frac{\left(H_{kj} - \hat{e}_{kj}\right)^2}{\hat{e}_{kj}}.$$

Under H_0, the test statistic V has approximately chi-square distribution with $(K-1)\cdot(J-1)$ degrees of freedom. The approximation is sufficient, if $\hat{e}_{kj} \geq 5$ for all pairs (k,j). When these conditions aren't fulfilled, adjoining realizations (or classes)

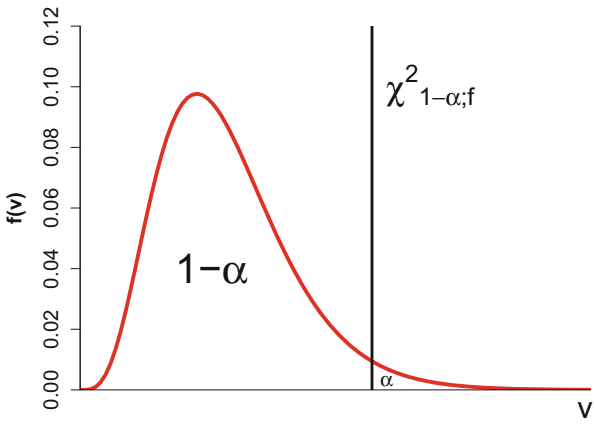


Fig. 9.25 Distribution for the chi-square test of independence

have to be combined into larger sets of (possible) observations. K and J denote the numbers of classes in both variables *after* such necessary (re-)grouping.

The *critical value* c satisfying $P(V \leq c) = 1 - \alpha$ has to be looked up in the table of the cumulative chi-square distribution function with appropriate degrees of freedom ($= (K-1) \cdot (J-1)$). The decision regions are

Rejection region for H_0:

$$\left\{ v \mid v > \chi^2_{1-\alpha;(K-1)\cdot(J-1)} \right\}.$$

Non-rejection region for H_0:

$$\left\{ v \mid v \leq \chi^2_{1-\alpha;(K-1)\cdot(J-1)} \right\}.$$

Under the null, the probability of the test statistic V assuming a value from the rejection region for H_0 equals the significance level $\alpha = P(V > \chi^2_{1-\alpha;(K-1)\cdot(J-1)} | H_0)$. The probability of the test statistic V being observed in the non-rejection region for H_0 is $P(V \leq \chi^2_{1-\alpha;(K-1)\cdot(J-1)} | H_0) = 1 - \alpha$ (Fig. 9.25).

Sampling and Computing the Test Statistic

After a sample of size n has been drawn, the absolute frequencies h_{kj} of all observed realization pairs (x_k, y_j) can be calculated. We can consolidate these into the empirical marginal frequencies for X and Y and derive the expected absolute frequencies $\hat{e}_{kj}$ from these according to the above formulas. If violated the approximation conditions necessitate further grouping, and the frequencies h_{kj}, $h_{k\bullet}$, $h_{\bullet j}$ and $\hat{e}_{kj}$ have to be recalculated. Plugging h_{kj} and $\hat{e}_{kj}$ into the test statistic formula yields the realized test statistic value v.

Test Decision and Interpretation

If v falls into the rejection region for H_0, the null hypothesis is rejected on the basis of a random sample of size n at a significance level of α ('H_1'). In this case it had been shown that the random variables X and Y are statistically dependent. If they *are* actually independent in the population, a type I error has been made ('H_1'| H_0), the probability of which in repeated samples (tests) equals the significance level: P ('H_1'|H_0) $= \alpha$.

If v belongs to the non-rejection region for H_0, the null hypothesis is not rejected on the basis of a random sample of size n ('H_0'). The sample doesn't statistically contradict the assumption of independence. A type II error has been made, in this case, if the alternative hypothesis is actually true ('H_0'|H_1).

More Information

The principle underlying the independence tests resembles that of the parametric tests. A test statistic is constructed to summarize (consolidate) the distance of the relevant information about the theoretical distribution under the null hypothesis from the corresponding structure in the sample (i.e., measure the distance between the two distributions). The distribution of the test statistic has to be determined—either exactly or approximately. The null hypothesis is being tested, and the decision can result in a type I error with probability P ('H_1'|H_0) $= \alpha$, if the null hypothesis has been rejected, or in a type II error, if it has not been rejected with probability P ('H_0'|H_1) $= \beta$. The error I probability is controlled by setting the significance level, but the type II error probability cannot be calculated, as there are infinitely many probability models different from that claimed to be the true one in the null hypothesis. For this reason one will try to reject the null hypothesis and thus back a possible rejection by a known maximum probability of making a wrong decision (in repeated samples).

Hypothesis

If the random variables *are* independent in the population, we expect this to be reflected in the sample. But a sample cannot convey all the information embedded in the population, and we have to account for random variation introduced by the sampling process. If the hypothesis is true, we expect it to be reflected accurately on statistical average only and have to determine what the expected deviation of the sample characteristics from the hypothetical ones arising from sampling noise is. Deviations of the observed joint absolute frequencies from those implied by independence, $\hat{e}_{kj}$, will occur with probability one. The task is to quantify them relative to the expected variation—an excess disagreement, i.e., a significant deviation, leading to a rejection of the null hypothesis. As it is always the null

hypothesis that is being tested, the independence of X and Y has to be proposed as null hypothesis. Only this way the expected absolute frequencies can be calculated; after all we need some probability model that allows us to derive the distribution of the test statistic and thus assess its intrinsic variation. Large deviations of the observed joint absolute frequencies h_{kj} from those expected if X and Y *are* independent, e_{kj}, contradict the independence assumption and thus increase the likelihood of rejecting the null hypothesis (everything else equal).

The test statistic underlying the chi-square test of independence is calculated using observed frequencies and the theoretical probabilities p_{kj}, $p_{k\bullet}$, and $p_{\bullet j}$ ($k = 1, \ldots, K; j = 1, \ldots J$). If X and Y are discrete random variables, the joint probabilities are related to exactly one pair of realizations:

$$p_{kj} = P\left(\{X = x_k\} \cap \left\{Y = y_j\right\}\right), \; p_{k\bullet} = P\left(\{X = x_k\}\right), \; p_{\bullet j} = P\left(\left\{Y = y_j\right\}\right).$$

Continuous random variables assume specific values with probability zero. Thus, the sample space has to be partitioned into exhaustive disjoint intervals. In the continuous case the probabilities are defined as follows:

- p_{kj} is the probability of X assuming a value belonging to the class $\left(x^*_{k-1}, x^*_k\right)$ *and* Y assuming a value from the class $\left(y^*_{j-1}, y^*_j\right)$,
- $p_{k\bullet}$ is the probability of X being observed in *kth* class $\left(x^*_{k-1}, x^*_k\right)$ (marginal probabilities of X),
- $p_{\bullet j}$ is the probability that Y takes values from the *jth* class $\left(y^*_{j-1}, y^*_j\right)$ (marginal probabilities of Y).

Formally:

$$\begin{aligned} p_{kj} &= P\left(\left\{x^*_{k-1} < X \leq x^*_k\right\} \cap \left\{y^*_{j-1} < Y \leq y^*_j\right\}\right), \\ p_{k\bullet} &= P\left(x^*_{k-1} < X \leq x^*_k\right), \\ p_{\bullet j} &= P\left(y^*_{j-1} < Y \leq y^*_j\right). \end{aligned}$$

To simplify and unify exposition for discrete and continuous random variables, x_k, $(k = 1, \ldots K)$ and y_j, $(j = 1, \ldots J)$ are taken to be values representative for the classes in the continuous case (e. g. midpoints). K and J denote the number of classes constructed for X and Y.

Note that it may prove necessary to group observations from discrete variables into classes—if only to improve approximation accuracy (for the price of a less detailed probability model).

Test Statistic

We want to illustrate why the joint absolute frequencies H_{kj} are random variables. Our argumentation is valid both for discrete and continuous variables.

Suppose we sample one statistical element from the population with respect to the random variables X and Y and check whether the observation pair equals (x_k, y_j), i.e., whether the event $\{X = x_k\} \cap \{Y = y_j\}$ has been realized. There are only two possible outcomes to this random experiment. The probability of the event $\{X = x_k\} \cap \{Y = y_j\}$ happening is p_{kj}, and the probability of one single element not being observed in this particular pair of X and Y realizations is $1 - p_{kj}$. If we draw a sample of n independent pairs of observations, we repeat this random experiment n times under the same conditions and thus with constant p_{kj}. In other words, we are carrying out a Bernoulli experiment with n replications.

In doing so, we are interested in the total number of occurrences of the event $\{X = x_k\} \cap \{Y = y_j\}$, i.e., the absolute frequency of the value pair (x_k, y_j) in the sample. This frequency is the outcome of a Bernoulli experiment and thus varies across samples. Thus, H_{kj} : "Number of occurrences of $\{X = x_k\} \cap \{Y = y_j\}$ in a simple random sample of size n" is a discrete random variable with possible outcomes $0, 1, \ldots, n$. The random variable H_{kj} has Binomial distribution with parameters n and p_{kj}: $H_{kj} \sim B(n; p_{kj})$. Expectation for H_{kj} is given by $E(H_{kj}) = np_{kj}$. If the null hypothesis is true and thus X and Y are statistically independent, the joint probability p_{kj} is calculated according to the multiplication rule for independent events as the product of the marginal probabilities $p_{k\bullet}$ and $p_{\bullet j}$: $p_{kj} = p_{k\bullet} \cdot p_{\bullet j}$. The expected joint absolute frequencies are then given by $e_{kj} = n \cdot P_{kj} = n \cdot p_{k\bullet} \cdot p_{\bullet j}$. This result applies to all $k = 1, \ldots, K$ and $j = 1, \ldots J$.

The test statistic is based on a comparison of the joint absolute frequencies encountered in the sample with those to be expected given the null hypothesis is true. The probabilities underlying the expected frequencies are unknown and have to be estimated from the sample. The comparison is based on the differences $H_{kj} - \hat{e}_{kj}$ as distance measures. To prevent negative differences from offsetting positive ones (or vice versa), the difference is squared: $(H_{kj} - \hat{e}_{kj})^2$. To account for varying importance of these squared deviations, they are weighted by dividing by $\hat{e}_{kj}$: A difference of $h_{kj} - \hat{e}_{kj} = 5$ receives a higher weighting if $\hat{e}_{kj} = 10$ than if $\hat{e}_{kj} = 100$. Summing over all pairs (k, j) summarizes (condenses) all weighted squared deviations into one test statistic:

$$V = \sum_{k=1}^{K} \sum_{j=1}^{J} \frac{(H_{kj} - \hat{e}_{kj})^2}{\hat{e}_{kj}}.$$

As the H_{kj} are random variables, so is V. Under the null hypothesis, for a sufficiently large sample size n and validity of the approximation conditions, V is approximately chi-square distributed with $(K - 1) \cdot (J - 1)$ degrees of freedom. If the approximation requirements aren't fulfilled, bordering classes or values have to be combined in a suitable way. The outcomes of discretely measured random experiments are then being grouped into classes. K and J are the numbers of classes remaining after such a necessary re-grouping.

Determining the Degrees of Freedom
There is a total of $K \cdot J$ probabilities p_{kj} constituting the bivariate distribution of the random variables X and Y as categorized in the two-dimensional contingency table. We lose one degree of freedom because the probabilities aren't independent of each other: From $\sum_k \sum_j p_{kj} = 1$ follows that any probability p_{kj} is determined by the other $K \cdot J - 1$ joint probabilities. If we could derive all probabilities joint probabilities from both variables' marginal distributions (probabilities) applying $p_{kj} = p_{k\bullet} \cdot p_{\bullet j}$, we had thus $K \cdot J - 1$ degrees of freedom. Unfortunately the marginal probabilities $p_{k\bullet}$ and $p_{\bullet j}$ are unknown and have to be estimated from the data, further reducing the degrees of freedom. The marginal distribution of X encompasses K probabilities $p_{k\bullet}$, of which only $K-1$ have to be estimated because $\sum_k p_{k\bullet} = 1$. The same applies to the marginal distribution of Y: As $\sum_j p_{\bullet j} = 1$, only $J - 1$ marginal probabilities $p_{\bullet j}$ have to be estimated. Thus, a total of $(K - 1) + (J - 1)$ marginal probabilities has to be estimated, and the overall degrees of freedom are:

$$K \cdot J - 1 - [(K - 1) + (J - 1)] = K \cdot J - K - J + 1 = (K - 1) \cdot (J - 1) .$$

As $\left(H_{kj} - \hat{e}_{kj}\right)^2 / \hat{e}_{kj}$ is positive for all pairs (k, j), the test statistic V will always be positive. Large deviations $H_{kj} - \hat{e}_{kj}$ translate into a high test statistic value. The null hypothesis is thus rejected for high values of V. Hence, the chi-square test of independence is a right-sided test.

Explained: The Chi-Square Test of Independence in Action

Someone suggests that the number of defects on a car is statistically independent from its age. We want to test this hypothesis at a significance level of $\alpha = 0.05$ using the chi-square test of independence.

The random variable X : "number of defects" is measured in the realization x_1 : "no defect," x_2 : "one defect" and x_3 : "two or more defects"; random variable Y : "cars' age" is categorized as x_1 : "$\leq$ 1 year," x_2 : "$>$ 1 year and $\leq$ 2 years" and x_2 : "$>$ 2 years."

Hypothesis

As the test statistic underlying the chi-square test of independence uses as inputs the expected joint frequencies, which are in turn calculated using the assumption of independence, the independence hypothesis must to be stated as null hypothesis:

$$\mathrm{H}_0 : X \text{ and } Y \text{ are statistically independent}$$

vs.

$$\mathrm{H}_1 : p_{kj} = X \text{ and } Y \text{ are } \textit{not} \text{ statistically independent}$$

or

$$\mathrm{H}_0 : p_{kj} = p_{k\bullet} \cdot p_{\bullet j} \forall\ (k,j)$$

vs.

$$\mathrm{H}_1 : p_{kj} \neq p_{k\bullet} \cdot p_{\bullet j} \text{ for at least one pair } (k,j)\,.$$

Test Statistic and Its Distribution: Decision Regions

We use the test statistic of the chi-square test of independence:

$$V = \sum_{k=1}^{K} \sum_{j=1}^{J} \frac{\left(H_{kj} - \hat{e}_{kj}\right)^2}{\hat{e}_{kj}}.$$

Under H_0, V is approximately chi-square distributed with $(K-1)\cdot(J-1)$ degrees of freedom. The decision regions of the null hypothesis can only be determined after the sample has been drawn and analyzed:

- First, the expected joint absolute frequencies have to be estimated.
- Then the approximation conditions can (must) be checked and necessary combinations of classes (or values) can be established.
- Once the two above steps have been concluded, and not before, the degrees of freedom can be determined and the critical values looked up.

Sampling and Computing the Test Statistic

Police officers positioned at various locations randomly stop 110 cars and record age and number of defects. In Table 9.14 , the absolute joint and marginal frequencies in this sample are listed together with the expected frequencies under the null hypothesis, calculated as

$$\hat{e}_{kj} = \frac{h_{k\bullet} \cdot h_{\bullet j}}{n}.$$

The approximation conditions are fulfilled, as all expected absolute joint frequencies are equal to or greater than five: $\hat{e}_{kj} \geq 5$. We are observing X and Y in $K = 3$ respectively $J = 3$ classes and thus have $(K-1)\cdot(J-1) = 4$ degrees of freedom. The critical value satisfying $P\,(V \leq c) = 1 - \alpha = 0.95$ is looked up in the table of the chi-square distribution as $c = \chi^2_{1-\alpha;(K-1)\cdot(J-1)} = \chi^2_{0.95;4} = 9.49$, implying the following decision regions

Rejection region for H_0:

$$\{v \mid v > 9.49\}\,.$$

Table 9.14 Absolute joint and marginal frequencies

# Defects (x_k)		Age (y_j)			
		< 1	$1-2$	> 2	MD X
0	Observed	30.0	14.0	5.0	49.0
	Expected	26.7	13.4	8.9	
1	Observed	18.0	10.0	4.0	32.0
	Expected	17.5	8.7	5.8	
≥ 2	Observed	12.0	6.0	11.0	29.0
	Expected	15.8	7.9	5.3	
MD Y		60.0	30.0	20.0	110.0

Non-rejection region for H_0:

$$\{v \mid v \leq 9.49\}.$$

The realized test statistic value is

$$v = \frac{(30-26.7)^2}{26.7} + \frac{(14-13.4)^2}{13.4} + \ldots + \frac{(11-5.3)^2}{5.3} = 10.5.$$

Test Decision and Interpretation

Since the test statistic value $v = 10.5$ falls into the rejection region the null hypothesis is rejected. Given our test parameters (sample size $n = 110$ and significance level $\alpha = 0.05$), we could verify the random variables X : "number of defects" and Y : "cars' age" to be statistically dependent. If this is not true in the population, we have made a type I error ('H_1'$|H_0$). In repeated samples (tests) the probability of this happening is given by the significance level $\alpha = 0.05$.

Enhanced: Chi-Square Test of Independence for Economic Situation and Outlook

In 1991 and 1996, randomly selected German citizens over 18 have been presented the following two questions:

- Q1) Assess the current economic situation
- Q2) What is the economic outlook for the upcoming year The participants we asked to express their opinion on the following scale:
- Possible answers for Q1): 1 = "Very Good", 2 = "Good", 3 = "Satisfactory", 4 = "Fair", 5 = "Poor"
- Possible answers for Q2): 1 = "Significantly improved", 2 = "Improved", 3 = "Unchanged", 4 = "Deteriorated", 5 = "Significantly deteriorated"

The questions are translated into the random variables X_1 : "Current economic situation" and X_2 : "Economic outlook," with the above realizations. In addition, a third variable Y : "Survey region" with the categories "West Germany" and "East Germany" has been recorded.

We want to test at a significance level of $\alpha = 0.05$, whether the random variables X_1 and Y respectively X_2 and Y as surveyed in 1991 and 1996 are statistically independent.

Hypothesis: Test Statistic and Its Distribution

The independence of the random variables has to be stated in H_0 to facilitate the computation of the expected absolute joint frequencies and thus the test statistic:

$$H_0 : X_1 \text{ and } Y \text{ are statistically independent}$$

vs.

$$H_1 : X_1 \text{ and } Y \text{ are } \textit{not} \text{ statistically independent}$$

and

$$H_0 : X_2 \text{ and } Y \text{ are statistically independent}$$

vs.

$$H_1 : X_2 \text{ and } Y \text{ are } \textit{not} \text{ statistically independent.}$$

We use the test statistic for the chi-square test of independence,

$$V = \sum_{k=1}^{K} \sum_{j=1}^{J} \frac{\left(H_{kj} - \hat{e}_{kj}\right)^2}{\hat{e}_{kj}},$$

which, under H_0, has approximately a chi-square distribution with $(K-1) \cdot (J-1)$ degrees of freedom. The decision regions of the null hypothesis cannot be determined before the sample has been drawn and analyzed, because we have to follow a sequential approach:

- First, we estimate the expected joint absolute frequencies.
- On this basis we can check the approximation conditions and, if necessary, combine values or classes.
- Now we can determine the degrees of freedom and retrieve the critical values.

Sampling and Computing the Test Statistic: Test Decision

Tables 9.15, 9.16, 9.17, and 9.18 contain the joint absolute frequencies in the samples of the years 1991 and 1996 as well as the expected absolute joint frequencies for true null hypothesis, calculated as

$$\hat{e}_{kj} = \frac{h_{k\bullet} \cdot h_{\bullet j}}{n},$$

and the differences $h_{kj} - \hat{e}_{kj}$.

The approximation conditions are fulfilled for all 4 tests to be conducted, i.e., $\hat{e}_{kj} \geq 5$ for all pairs (k, j). The critical value satisfying $P(V \leq c) = 0.95$ is $\chi^2_{1-\alpha;(K-1)\cdot(J-1)} = \chi^2_{0.95;4} = 9.49$ as we have $(K - 1 \cdot (J - 1) = 4$ degrees of freedom. The decision regions are thus

Rejection region for H_0:

$$\{v \mid v > 9.49\}.$$

Non-rejection region for H_0:

$$\{v \mid v \leq 9.49\}.$$

Chi-square values and resulting decisions for the 4 tests are given in Table 9.19.

Table 9.15 Current economic situation (X_1) versus region (Y), 1991

Current economic situation (x_{1k})		Region (y_j) West	Region (y_j) East	MD X_1
Very good	Observed	209.0	165.0	374.0
	Expected	184.8	189.2	
	Difference	24.2	−24.2	
Good	Observed	744.0	592.0	1,336.0
	Expected	660.1	675.9	
	Difference	83.9	−83.9	
Satisfactory	Observed	431.0	647.0	1,078.0
	Expected	532.6	545.5	
	Difference	−101.6	101.6	
Fair	Observed	36.0	39.0	75.0
	Expected	37.1	37.9	
	Difference	−1.1	1.1	
Poor	Observed	4.0	15.0	19.0
	Expected	9.4	9.6	
	Difference	−5.4	5.4	
MD Y		1,424.0	1,458.0	2,882.0

Table 9.16 Current economic situation (X_1) versus region (Y), 1996

Current economic situation (x_{1k})		Region (y_j) West	Region (y_j) East	MD X_1
Very good	Observed	20.0	6.0	26.0
	Expected	17.2	8.8	
	Difference	2.8	−2.8	
Good	Observed	264.0	116.0	380.0
	Expected	251.3	128.7	
	Difference	12.7	−12.7	
Satisfactory	Observed	1,006.0	557.0	1,563.0
	Expected	1,033.7	529.3	
	Difference	−27.7	27.7	
Fair	Observed	692.0	335.0	1,027.0
	Expected	679.2	347.8	
	Difference	12.8	−12.8	
Poor	Observed	141.0	73.0	214.0
	Expected	141.5	72.5	
	Difference	−0.5	0.5	
MD Y		2,123.0	1,087.0	3,210.0

Table 9.17 Economic outlook (X_1) versus region (Y), 1991

Economic outlook (x_{2k})		Region (y_j) West	Region (y_j) East	MD X_2
Significantly improved	Observed	75.0	203.0	278.0
	Expected	137.4	140.6	
	Difference	−62.4	62.4	
Improved	Observed	449.0	763.0	1,212.0
	Expected	598.9	613.1	
	Difference	−149.9	149.9	
Unchanged	Observed	684.0	414.0	1,108.0
	Expected	547.5	560.5	
	Difference	136.5	−136.5	
Deteriorated	Observed	200.0	62.0	262.0
	Expected	129.5	132.5	
	Difference	70.5	−70.5	
Significantly deteriorated	Observed	16.0	6.0	22.0
	Expected	10.9	11.1	
	Difference	5.1	−5.1	
MD Y		1,424.0	1,458.0	2,882.0

Table 9.18 Economic outlook (X_1) versus region (Y), 1996

Economic outlook (x_{2k})		Region (y_j) West	Region (y_j) East	MD X_2
Significantly improved	Observed	9.0	6.0	15.0
	Expected	9.9	5.1	
	Difference	−0.9	0.9	
Improved	Observed	190.0	131.0	321.0
	Expected	212.3	108.7	
	Difference	−22.3	22.3	
Unchanged	Observed	809.0	444.0	1, 253.0
	Expected	828.7	42.3	
	Difference	−19.7	19.7	
Deteriorated	Observed	960.0	426.0	1, 386.0
	Expected	916.7	469.3	
	Difference	43.3	−43.3	
Significantly deteriorated	Observed	155.0	80.0	235.0
	Expected	155.4	79.6	
	Difference	−0.4	0.4	
MD Y		2, 123.0	1, 087.0	3, 210.0

Table 9.19 Chi-square values and resulting decisions

Year	Random variables	Test statistic value v	Test decision
1991	X_1, Y	71.85	Reject H_0
1996	X_1, Y	6.15	Do not-reject H_0
1991	X_2, Y	278.17	Reject H_0
1996	X_2, Y	14.61	Reject H_0

Interpretation

Perception of Current Economic Situation

While the 1991 data rejects the null hypothesis of statistical independence, at a significance level of 0.05, the proposition that the random variables X_1 : "Current economic situation" and Y : "Survey region" are statistically independent is not-rejected for the 1996 data. But we can extract more qualitative information if we look at the contingency tables. As can be seen from the comparatively high positive differences $h_{kj} - \hat{e}_{kj}$ for the positive statements in Table 9.15, in 1991 West Germans tended to classify the economic situation more positively compared to the East Germans. In 1996, there are still positive differences $h_{kj} - \hat{e}_{kj}$, but their sum isn't significant anymore. Some kind of convergence in the assessment of the (then) current economic situation has taken place.

Economic Outlook
Both surveys' data reject the null hypothesis that the random variables X_2 : "Economic outlook" and Y : "Survey region" are statistically independent at a significance level of $\alpha = 0.05$. Observe that in both years the East Germans have been more positive about the future of the economy than the West Germans. If you compare the differences $h_{kj} - \hat{e}_{kj}$ for both years, you will notice the same qualitative tendency to homogeneity in opinions across (East and West) Germany as in the assessment of the current economic environment. Yet quantitatively they are still large enough (in total) to be significant in 1996, and we cannot but conclude (at least within the assumed test parameter setting) that the East and West Germans have structurally different opinions. The type of dependency between X_2 and Y can be explored using suitable statistical tools for dependence analysis (e.g., categorical regression).

Chapter 10
Two-Dimensional Frequency Distribution

10.1 Introduction

In the natural sciences, we can often clearly represent the relationship between two variables by means of a function because it has its origin in physical laws.

That is quite different in socio-economic studies. What kind of relationship for instance exists between income and consumption expenditures? In this case, we may not be able to clearly describe the relationship by means of a simple deterministic function. However, there may be a statistical relationship which can be uncovered. Tools such as Scatterplots, Scatterplot-Matrices, and 3D-scatterplots can be used for exploratory analysis. We can also use contingency tables and measures of dependence to evaluate the strength of a relationship.

In the following chapter, we will analyze n statistical observations on a pair of variables X and Y. Questions that we will want to answer include:

- Is there a relationship or a dependency between the variables X and Y?
- How pronounced is this relationship?
- Can we describe the relationship by means of a function?

10.2 Two-Dimensional Frequency Tables

We are given:

- Variable X which takes on possible values x_i $(i = 1, \ldots, m)$
- Variable Y which takes on possible values y_j $(j = 1, \ldots, r)$ (Table 10.1)

W.K. Härdle et al., *Introduction to Statistics*, DOI 10.1007/978-3-319-17704-5_10

Table 10.1 Structure of a contingency table

	Variable Y					
Variable X	y_i	$\cdots$	y_j	$\cdots$	y_r	MD X
x_1	h_{11}	$\cdots$	h_{1j}	$\cdots$	h_{1r}	$h_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_i	h_{i1}	$\cdots$	h_{ij}	$\cdots$	h_{ir}	$h_{i\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
MD Y	$h_{\cdot 1}$	$\cdots$	$h_{\cdot j}$	$\cdots$	$h_{\cdot r}$	$h_{\cdot\cdot} = n$

Realizations $m \cdot r$

By enumerating all possible ordered pairs, we can calculate their number to be $m \cdot r$.

$$(x_i, y_j) = (X = x_i) \bigcap (Y = y_j)$$

Absolute Frequency

The absolute frequency for a particular ordered pair (x_i, y_j) is the number of observations which take on that specific combination of values:

$$h(x_i, y_j) = h_{ij}.$$

Relative Frequency

The relative frequency is the proportion of observations with a specific combination of values (x_i, y_j)

$$f(x_i, y_j) = f_{ij} = h(x_i, y_j)/n.$$

Properties

$$\sum_{i=1}^{m} \sum_{j=1}^{r} h(x_i, y_j) = n, \qquad \sum_{i=1}^{m} \sum_{j=1}^{r} f(x_i, y_j) = 1$$

A two-dimensional absolute frequency distribution or contingency table tabulates the number of occurrences of each ordered pair. In the right and bottom margins it tabulates the marginal frequencies or marginal distribution (MD) for each variable individually. These are the row sums $h_{i.}$ and column sums $h_{.j}$ of the joint frequencies (Table 10.2).

The two-dimensional relative frequency distribution is defined similarly using the relative frequencies (f_{ij}). (Note, this may be accomplished simply by dividing all of the absolute frequencies in the absolute frequency distribution table by n.)

- 5×3 contingency table
- X—occupation; Y—athletic activity
- $n = 1000$ working people

Explained: Two-Dimensional Frequency Distribution

For $n = 100$ randomly selected persons it has been determined whether they smoke and whether they have had lung cancer. The variables are:

- **X - Smoker** with realizations x_1 = "yes" and x_2 = "no"
- **Y - Lung cancer** with realizations y_1 = "yes" and y_2 = "no"

The two-dimensional frequency distribution is provided in a 2×2 contingency table shown in Table 10.3.

The numbers in Table 10.3 have the following meaning: Among smokers there were 10 cases of lung cancer, among nonsmokers only 5 cases. Among all surveyed persons there were 25 smokers; 85 of the surveyed persons did not have lung cancer.

Table 10.2 Example of a contingency table

	Athletic activity Y			
Occupation X	Rarely	Sometimes	Regularly	MD X
Worker	240	120	70	430
Salaried	160	90	90	340
Civil servant	30	30	30	90
Farmer	37	7	6	50
Self-employed	40	32	18	90
MD Y	507	279	214	1000

Table 10.3 Two-dimensional frequency distribution for X and Y

	Lung cancer		
Smoker	Yes(y_1)	No(y_2)	MD X
Smoker(x_1)	10	15	25 ($h_{1.}$)
Nonsmoker (x_2)	5	70	75 ($h_{2.}$)
MD Y	15 ($h_{.1}$)	85 ($h_{.2}$)	100 (n)

Enhanced: Department Store

The "department store" data set contains the following variables recorded for $n = 165$ randomly selected customers:

Variable	Possible realizations
X gender	1—male
	2—female
Y method of payment	1—cash
	2—ATM card
	3—credit card
Z residence	1—Berlin
	2—not in Berlin

Below, the three possible two-dimensional frequency distributions are given that can be constructed from this data. Absolute frequencies h_{ij} and relative frequencies f_{ij} (in brackets and rounded to three decimals) are given.

The two-dimensional frequency distribution for the variables **gender** and **method of payment** is a 2×3 contingency table (Table 10.4).

The two-dimensional frequency distribution for the variables **gender** and **residence** is a 2×2 contingency table (Table 10.5).

The two-dimensional frequency distribution for the variables **residence** and **method of payment** is a 2×3 contingency table (Table 10.6).

Table 10.4 Two-dimensional frequency distribution for gender and method of payment

	Method of payment (Y)			
Gender (X)	(y_1)	(y_2)	(y_3)	MD X
Male (x_1)	31 (0.188)	32 (0.194)	23 (0.139)	86 (0.521)
Female (x_2)	30 (0.182)	29 (0.176)	20 (0.121)	79 (0.479)
MD Y	61 (0.370)	61 (0.370)	43 (0.260)	165 (1.00)

Table 10.5 Two-dimensional frequency distribution for gender and residence

	Residence (Z)		
Gender (X)	Berlin (z_1)	Not in Berlin (z_2)	MD X
Male (x_1)	50 (0.303)	36 (0.218)	86 (0.521)
Female (x_2)	37 (0.224)	42 (0.255)	79 (0.429)
MD Y	87 (0.527)	78 (0.473)	165 (1.00)

Table 10.6 Two-dimensional frequency distribution for residence and method of payment

	Method of payment (Y)			
Residence (Z)	(y_1)	(y_2)	(y_3)	MR X
Berlin (z_1)	44 (0.267)	22 (0.133)	21 (0.127)	87 (0.527)
Not in Berlin (z_2)	17 (0.103)	39 (0.237)	22 (0.133)	78 (0.473)
MD Y	61 (0.370)	61 (0.370)	43 (0.260)	165 (1.00)

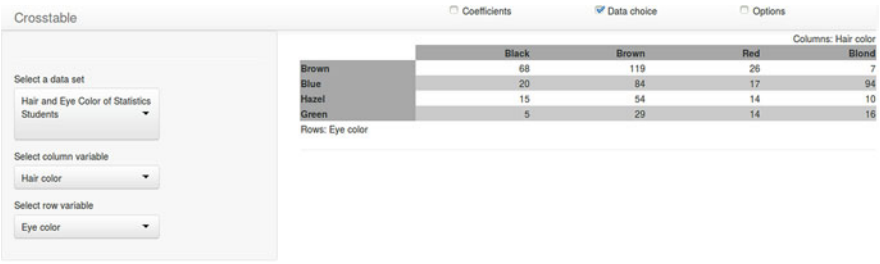


Fig. 10.1 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_tab2

Interactive: Example for Two-Dimensional Frequency Distribution

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right (Fig. 10.1).

Please select

- a dataset
- a column and row variable

Output

The interactive example allows us to display a two-dimensional frequency distribution in the form of a crosstable for a variety of variables.

10.3 Graphical Representation of Multidimensional Data

Frequency Distributions

When there are two variables, a three-dimensional graph is required to depict the frequency distribution where the vertical dimension corresponds to frequencies. Alternatively, one can use a grouped bar chart.

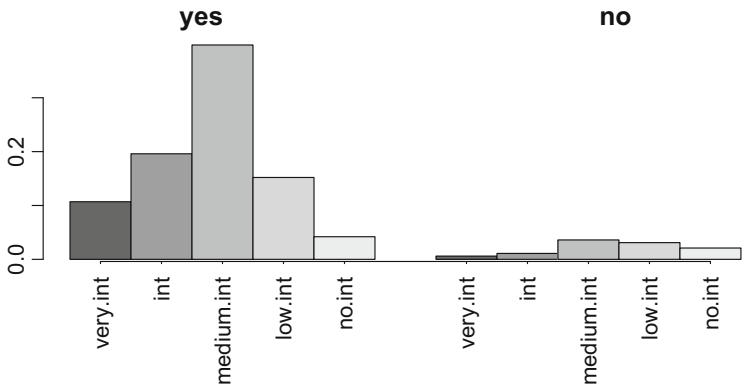


Fig. 10.2 Example of a grouped bar chart

Grouped Bar Chart

For each value of one variable a group of bars corresponding to the value of the second "grouping variable" are drawn (Fig. 10.2).

3D-Bar Chart

For each combination of values of the two variables, a vertical bar is drawn with height proportional to the frequency (Fig. 10.3).

Scatterplots

Scatterplot

We can represent observations on two continuous variables as points in a plane (a scatterplot). Scatterplots are very useful to show possible relations between two variables measured on a metric scale (example: increase of variable X leads to a visible increase of variable Y; Fig. 10.4).

3D-Scatterplot

We can represent simultaneously three continuous variables in a 3D-scatterplot. Different statistical software can be used to rotate the 3D-scatterplot which helps us to see possible relations (Fig. 10.5).

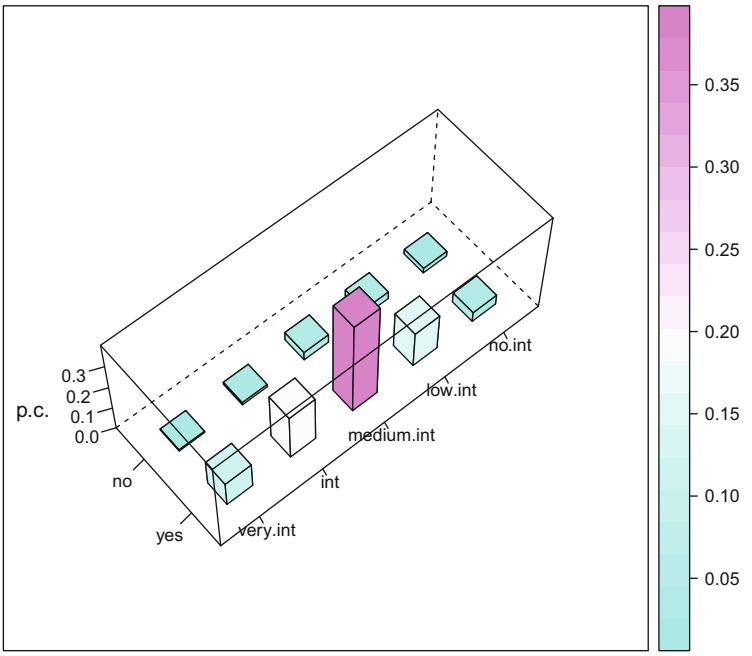


Fig. 10.3 Example of a 3D-bar chart

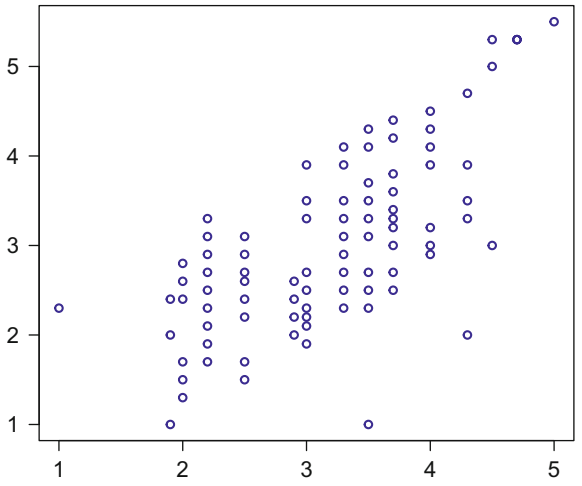


Fig. 10.4 Example of a scatterplot

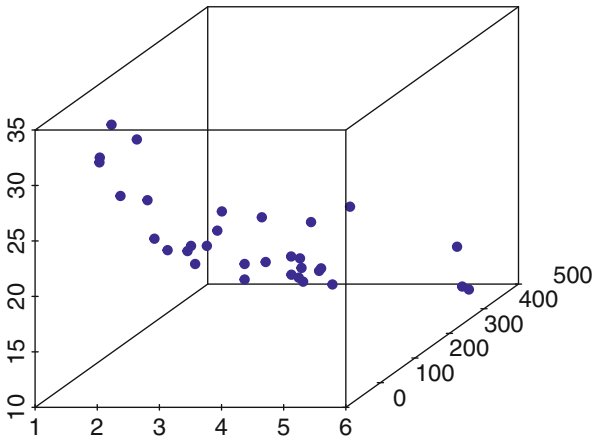


Fig. 10.5 Example of a 3D-scatterplot

Scatterplot-Matrix

If we need to analyze more than two continuous variables, we can use the scatterplot-matrix to represent them. Here, we produce the scatterplots of all possible pairs of two variables and put them together as a matrix. However, interpretation and clarity becomes increasingly difficult the greater the number of variables being studied (Fig. 10.6).

Explained: Graphical Representation of a Two- or Higher Dimensional Frequency Distribution

In 1985, the following variables describing criminal activity were recorded for each of the 50 states of the USA:

*X*1—land area
*X*2—population
*X*3—murder
*X*4—rape
*X*5—robbery
*X*6—assault
*X*7—burglary
*X*8—larceny
*X*9—auto theft
*X*10—US states region number
*X*11—US states division number

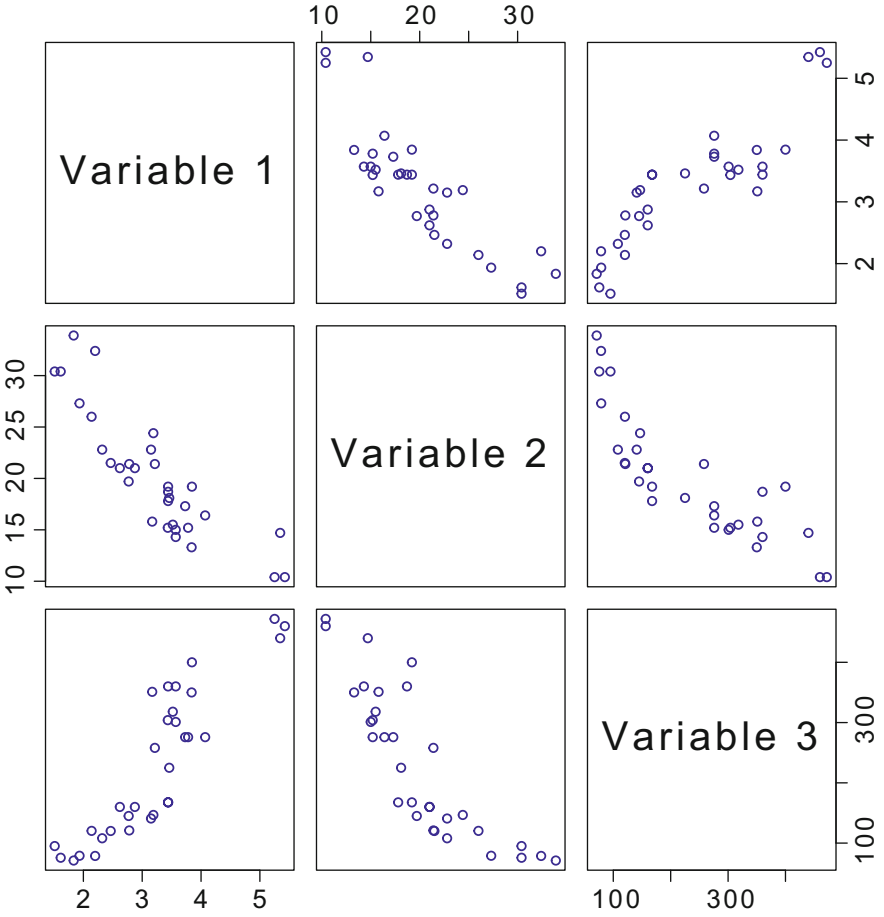


Fig. 10.6 Example of a scatterplot-matrix

The relationship between rate of "murder" ($X3$) and "population size" ($X2$) can be visualized in a scatterplot. Each state is represented in the scatterplot by a point with coordinates $(x2, x3)$ (Fig. 10.7).

The scatterplot shows a tendency of the rate of murder to increase with population size.

The three variables "population" ($X2$), "murder" ($X3$), and "robbery" ($X5$) can be visualized simultaneously in a 3D-scatterplot (Fig. 10.8).

Note: You can use this section's interactive example to visualize the relationships between the other variables of this data set (Fig. 10.9).

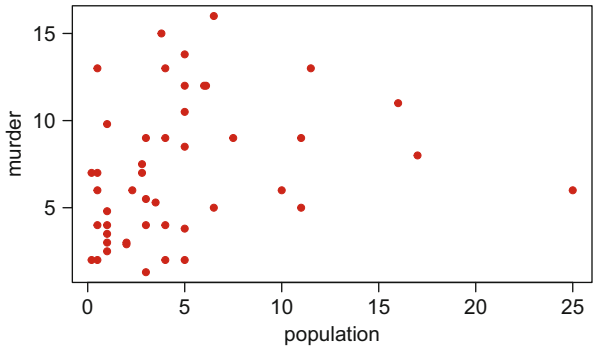


Fig. 10.7 Scatterplot of murder ($X3$) and population size ($X2$)

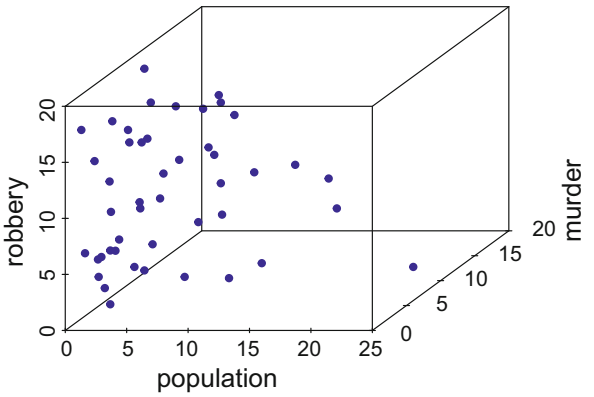


Fig. 10.8 3D-scatterplot of population ($X2$), murder ($X3$) and robbery ($X5$)

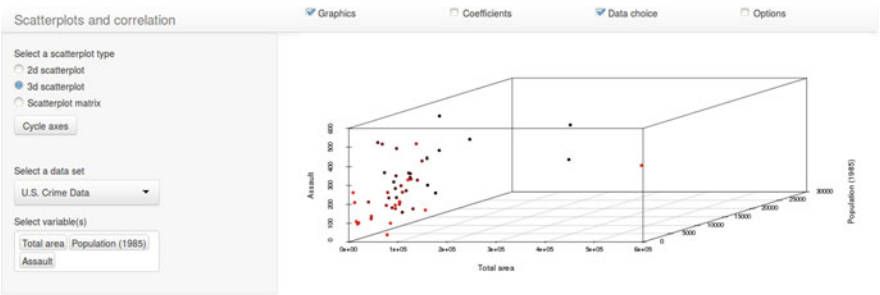


Fig. 10.9 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_plot

Interactive: Example for the Graphical Representation of a Two- or Higher Dimensional Frequency Distribution

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please select a scatterplot type, e.g., scatterplot matrix.

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

Output

The interactive example allows us to display a two- or three-dimensional frequency distribution in the form of a 2D/3D-scatterplot or a scatterplot matrix. After choosing a set of variables (attention: three variables are required for the 3D-plot), the output window shows the corresponding scatterplot.

10.4 Marginal and Conditional Distributions

Marginal Distribution

Suppose one is given a two-dimensional frequency distribution of the variables X and Y. The marginal distribution of X (respectively Y) is the one-dimensional distribution of variable X (respectively Y), in which we do not consider what happens to variable Y (respectively X).

The marginal distribution is the result of "adding up" the frequencies of the realizations. For example for the marginal (absolute) distribution of X as shown in Table 10.7.

Table 10.7 Marginal distributions

Variable X	Variable Y			Marginal distribution of X
	y_1	y_2	y_3	
...	...	...	...	...
x_i	$h(x_i, y_1)$	$h(x_i, y_2)$	$h(x_i, y_3)$	$= h(x_i, y_1)$ $+h(x_i, y_2)$ $+h(x_i, y_3)$
...	...	...	...	...
Marginal distribution of Y	...	...	...	

Marginal absolute distribution of variable X with the values x_j:

$$h_{i\cdot} = \sum_{j=1}^{r} h_{ij}; \quad i = 1, \ldots, , m$$

Marginal absolute distribution of variable Y with the values y_j:

$$h_{\cdot j} = \sum_{i=1}^{m} h_{ij}; \quad j = 1, \ldots, r$$

Total number of observations equals n:

$$h_{\cdot\cdot} = \sum_{i=1}^{m} \sum_{j=1}^{r} h_{ij} = \sum_{i=1}^{m} h_{i\cdot} = \sum_{j=1}^{r} h_{\cdot j} = n$$

The marginal relative distribution is defined similarly using the relative frequencies (f_{ij}). (Note, this may be accomplished simply by dividing all of the absolute frequencies in the marginal absolute distribution table by n.)

Marginal relative distribution of variable X with the values x_j:

$$f_{i\cdot} = \sum_{j=1}^{r} f_{ij}; \quad i = 1, \ldots, , m$$

Marginal relative distribution of variable Y with the values y_j:

$$f_{\cdot j} = \sum_{i=1}^{m} f_{ij}; \quad j = 1, \ldots, r$$

Total of all relative frequencies equals 1:

$$f_{\cdot\cdot} = \sum_{i=1}^{m} \sum_{j=1}^{r} f_{ij} = \sum_{i=1}^{m} f_{i\cdot} = \sum_{j=1}^{r} f_{\cdot j} = 1$$

Conditional Distribution

Suppose one is given a two-dimensional frequency distribution of two variables X and Y. The frequency distribution of X given a particular value of Y is called the conditional distribution or conditional distribution of X given y_j. (The conditional distribution of Y given x_i is defined analogously.)

Conditional relative frequency distribution of X for a given $Y = y_j$:

$$f(x_i|Y = y_j) = f(x_i|y_j) = \frac{f_{ij}}{f_{\cdot j}} = \frac{h_{ij}}{h_{\cdot j}}$$

Conditional relative frequency distribution of Y for a given $X = x_i$:

$$f(y_j|X = x_i) = f(y_j|x_i) = \frac{f_{ij}}{f_{i\cdot}} = \frac{h_{ij}}{h_{i\cdot}}$$

Like marginal distributions, conditional distributions are one-dimensional distributions.

Example The starting point is the 5×3 contingency table of the two variables:

- X—occupation
- Y—athletic activity

which have been observed for $n = 1000$ employed persons (Table 10.8).

The conditional distribution of the variable Y (athletic activity) for a given x_i (occupational group) are summarized in Table 10.9.

Table 10.8 Contingency table of X and Y

	Athletic activity Y			
Occupation X	Rarely	Sometimes	Regularly	MDX
Worker	240	120	70	430
Salaried	160	90	90	340
Civil servant	30	30	30	90
Farmer	37	7	6	50
Self-employed	40	32	18	90
MD Y	507	279	214	1000

Table 10.9 Conditional distribution of Y given X

	Athletic activity Y			
Occupation X	Rarely	Sometimes	Regularly	
Worker	0.56	0.28	0.16	1.00
Salaried	0.47	0.26	0.26	1.00
Civil servant	0.33	0.33	0.33	1.00
Farmer	0.74	0.14	0.12	1.00
Self-employed	0.44	0.36	0.20	1.00

Explained: Conditional Distributions

In a survey of 107 students their major and gender were recorded. The responses were used to produce the 9×2 contingency table given in Table 10.10.

What are the shares of females and males in each major? The answer is given by the conditional distributions of gender, given the major. The frequencies of the conditional distribution are computed as the ratio of the corresponding cells of the joint distribution table and the marginal distribution (i.e., row sum in this case) of the respective major (Table 10.11).

The results show that business is dominated by males who account for 73.7 % of all students majoring in business. In theology, on the other hand, women are the majority comprising 77.8 % of theology majors.

Table 10.10 Contingency table of gender and university major for 107 students

	Gender		
Major	Female	Male	MD (Major)
Social sc.	12	13	25
Engineering	1	1	2
Law	8	13	21
Medicine	6	4	10
Natural sc.	1	8	9
Psychology	3	8	11
Other	1	0	1
Theology	7	2	9
Business	5	14	19
MD (Gender)	44	63	107

Table 10.11 Conditional distribution of gender given the university major

	Female	Male	
Social sc.	0.480	0.520	1.000
Engineering	0.500	0.500	1.000
Law	0.381	0.619	1.000
Medicine	0.600	0.400	1.000
Natural sc.	0.111	0.889	1.000
Psychology	0.273	0.727	1.000
Other	1.000	0.000	1.000
Theology	0.778	0.222	1.000
Business	0.263	0.737	1.000
Total	0.411	0.589	1.000

Enhanced: Smokers and Lung Cancer

For $n = 100$ randomly selected persons it has been determined whether they smoke and whether they have had lung cancer. The variables are:

- **X—Smoker** with realizations x_1 = "yes" and x_2 = "no",
- **Y—Lung cancer** with realizations y_1 = "yes" and y_2 = "no."

The two-dimensional frequency distribution is a 2×2 contingency table shown in Table 10.12.

The conditional distributions of the variable X (smoker) for a given y_j (lung cancer) are shown in Table 10.13.

Each element of the conditional distribution has been calculated as the ratio of the respective cell of the joint distribution and the corresponding element of the Y marginal distribution.

From Table 10.13 we learn that 66.7 % of all persons diagnosed with lung cancer are smokers. 82.4 % of the persons not diagnosed with lung cancer are nonsmokers.

The conditional distribution of the variable Y (lung cancer) for a given value x_i (smoker/nonsmoker) is constructed analogously and shown in Table 10.14.

Hence, 40 % of all smokers but only 6.7 % of all nonsmokers have been diagnosed with lung cancer.

Table 10.12 Two-dimensional frequency distribution for X and Y

	Lung cancer		
Smoker	Yes (y_1)	No (y_2)	MD X
Smoking yes (x_1)	10	15	25
Smoking no (x_2)	5	70	75
MD Y	15	85	100

Table 10.13 Conditional distribution of X given Y

	Lung cancer	
Smoker	Yes (y_1)	No (y_2)
Smoker yes	0.667	0.176
Smoker no	0.333	0.824
	1.000	1.000

Table 10.14 Conditional distribution of X given Y

	Lung cancer		
Smoker	Yes (y_1)	No (y_2)	
Smoker yes (x_1)	0.400	0.600	1.000
Smoker no (x_2)	0.067	0.933	1.000

Table 10.15 Contingency table of age and education for 941 persons

Age	Education				MD (Age)
	University	High school	Middle school	Lower school	
18–29	38	93	134	42	307
30–39	23	94	168	70	355
40–49	12	39	129	99	279
MD (Education)	73	226	431	211	941

Table 10.16 Conditional distribution of education given age

	University	High school	Middle school	Lower school	
18–29	0.124	0.303	0.436	0.137	1.000
30–39	0.065	0.265	0.473	0.197	1.000
40–49	0.043	0.140	0.462	0.355	1.000

Table 10.17 Conditional distribution of age given education

	University	High school	Middle school	Lower school
18–29	0.521	0.411	0.311	0.199
30–39	0.315	0.416	0.390	0.332
40–49	0.164	0.173	0.299	0.469
	1.000	1.000	1.000	1.000

Enhanced: Educational Level and Age

In a survey of 941 persons, respondents' age (grouped as 18–29, 30–39, and 40–49) and the highest level of education attained (university, high school, middle school, and lower school) were recorded. The observed frequencies are shown in the 3×4 contingency table in Table 10.15.

The conditional distributions of educational attainment, given age, are summarized in Table 10.16.

Each element of the distribution has been calculated as the ratio of the respective cell of the joint distribution and the corresponding element of the marginal distribution of age.

Table 10.16 shows that among the 18–29-year-olds 12.4 % have completed a university education, 30.3 % graduated from high school, and 43.6 % finished middle school. In the group of 40–49-year-olds the fraction of persons with a university degree is only 4.3 %.

The conditional distribution of age, for a given level of educational attainment, is constructed analogously and shown in Table 10.17.

It can be seen that among those with at most a high school education, 41.1 % belong to the age group 18–29, 41.6 % to the age group 30–39, and 17.3 % to the age group 40–49.

10.5 Characteristics of Two-Dimensional Distributions

For the marginal distribution and the conditional distribution we can use the location- and dispersion measures the same way as for one-dimensional distributions (see preceding chapter), because they are also one-dimensional distributions.

Covariance

The covariance is a special characteristic for two-dimensional distributions that measures the common variation of two variables X and Y on a continuous scale.

The covariance for a pair of discrete random variables with true probabilities $p_{ij} = p(x_i; y_j)$, $(i = 1, \ldots, m;\ j = 1, \ldots, r)$ is given by:

$$Cov(X, Y) = \sum_{i=1}^{m} \sum_{j=1}^{r} (x_i - E(X))(y_j - E(Y)) \cdot p_{ij}$$

If one has n observations on these variables with absolute frequencies $h(x_i; y_j)$ and relative frequencies $f(x_i; y_j)$ $(i = 1, \ldots, m;\ j = 1, \ldots, r)$ one can calculate the sample covariance:

$$\begin{aligned} s_{xy} &= \frac{1}{n(n-1)} \sum_{i=1}^{m} \sum_{j=1}^{r} (x_i - \bar{x})(y_j - \bar{y}) \cdot h_{ij} \\ &= \frac{1}{n-1} \sum_{i=1}^{m} \sum_{j=1}^{r} (x_i - \bar{x})(y_j - \bar{y}) \cdot f_{ij} \\ &= \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y}) \end{aligned}$$

In contrast to the variance, the covariance can also take on negative values.

Properties of Covariance

- If the variables X and Y are independent, then the covariance is zero.
  Note: This does not automatically work the other way round. If the covariance between the variables X and Y is zero we cannot conclude that they are statistically independent.
- The contribution of a realization $(x_i; y_j)$ to the covariance is positive if the differences $(x_i - \bar{x})$ and $(y_j - \bar{y})$ have the same sign. It is negative if the differences $(x_i - \bar{x})$ and $(y_j - \bar{y})$ have opposite signs.

- The covariance of a variable with itself is equal to the variance of this variable: $s_x^2 = s_{xx}$
- Linear transformation: $X^* = a + bX$, and $Y^* = c + dY$, then $s_{X^*Y^*} = b \cdot d \cdot s_{XY}$

Independent Variables

Independence means that the distribution of a variable does not depend on the values of another variable. If two variables X and Y are independent:

1. All the conditional distributions of X are equal to each other and to the corresponding marginal distribution, that is for the conditional distribution of X: $f(x_i|y_j) = f(x_i|y_k) = f(x_i)$ for all $j, k = 1, \ldots, r$ and for all $i = 1, \ldots, m$ and the same holds for the conditional distribution of Y: $f(y_j|x_i) = f(y_j|x_h) = f(y_j)$ for all $i, h = 1, \ldots, m$ and for all $j = 1, \ldots, r$.
2. The joint distribution is equal to the product of the marginal distributions:

$$\begin{aligned} f(x_i|y_j) &= f(x_i) = \frac{f(x_i, y_j)}{f(y_j)} \\ &\Leftrightarrow f(x_i, y_j) = f(x_i) \cdot f(y_j) \\ f(y_j|x_i) &= f(y_j) = \frac{f(x_i, y_j)}{f(x_i)} \\ &\Leftrightarrow f(x_i, y_j) = f(x_i) \cdot f(y_j) \end{aligned}$$

Similarly, in the case of independence, the true joint probabilities also factor into a product of the marginal probabilities $p_{ij} = p(x_i; y_j) = p(x_i) \cdot p(x_j) = p_i \cdot p_j$, $(i = 1, \ldots, m;\ j = 1, \ldots, r)$.

We hardly ever use the covariance as autonomous characteristic. It is more of an auxiliary quantity that we can use to calculate other characteristics (see correlation in the following paragraph).

Note: A data set observed from independent variables may not exhibit an exact factorization of the joint relative frequency distribution into the product of the marginal frequency distributions of the respective variables. Similarly, the sample covariance of a data set drawn from independent variables may not be exactly zero. But one may conclude, as a result of further statistical tests, that the joint relative frequency distribution approximately factors and the sample covariance approximately equals zero thereby providing evidence that the variables are independent. However, keep in mind that a covariance of zero is only a necessary condition for independence, not a sufficient one.

More Information

If the variables X and Y are independent, then their covariance is equal to zero, that is, $Cov(X, Y) = 0$.

Proof

$$\begin{aligned}
Cov(X,Y) &= \sum_{i=1}^{m}\sum_{i=1}^{r}(x_i - E(x))(y_j - E(y)) \cdot p_{ij} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{r}(x_i - E(x))(y_j - E(y)) \cdot p_i \cdot p_j \\
&= \left\{\sum_{i=1}^{m}(x_i - E(x))p_i\right\}\left\{\sum_{j=1}^{r}(y_j - E(y))p_j\right\} \\
&= \left\{\sum_{i=1}^{m}x_i p_i - E(x)\sum_{i=1}^{m}p_i\right\}\left\{\sum_{j=1}^{r}y_j p_j - E(y)\sum_{j=1}^{r}p_j\right\} \\
&= \{E(x) - E(x)\}\{E(y) - E(y)\} = 0
\end{aligned}$$

Explained: How the Covariance Is Calculated

For $n = 15$ firms the variables Y—annual profit (in Mill. Euro) and X—annual rent for computer equipment (in 1,000 Euro) have been recorded. The possible realizations of these variables are given in columns 2 and 3 of Table 10.18.

For these 15 firms, how much common variation (about their respective means) exists between variables X and Y? The sample means (averages) of the variables are:

$$\bar{y} = 30 \text{ (Mill. Euro)}$$

$$\bar{x} = 200 \text{ (Tsd. Euro)}$$

Column 4 of the table contains the deviation of variable Y from its sample mean and those for variable X are contained in column 5.

The sample covariance is calculated according to the following formula:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{m}\sum_{j=1}^{r}(x_i - \bar{x})(y_j - \bar{y}) \cdot f_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})$$

Table 10.18 Annual profits (in Mill. Euro) and annual rent for computer equipment (in Tsd. Euro) for 15 firms

Firm i	Annual profit y_i	Annual rent x_i	$(y_i - \bar{y})$	$(x_i - \bar{x})$	$(y_i - \bar{y})(x_i - \bar{x})$
1	10	30	−20	−170	3,400
2	15	30	−15	−170	2,550
3	15	100	−15	−100	1,500
4	20	50	−10	−150	1,500
5	20	100	−10	−100	1,000
6	25	80	−5	−120	600
7	30	50	0	−150	0
8	30	100	0	−100	0
9	30	250	0	50	0
10	35	180	5	−20	−100
11	35	330	5	130	650
12	40	200	10	0	0
13	45	400	15	200	3,000
14	50	500	20	300	6,000
15	50	600	20	400	8,000

The product of the deviations for each firm is listed in column 6 of the table. The sample covariance is the sum of the elements of this column divided by $(n-1) = 14$.

$$s_{xy} = 28,100/14 = 2,007.143$$

10.6 Relation Between Continuous Variables (Correlation, Correlation Coefficients)

The common variation (covariation) of the two continuous variables X and Y determines the strength of the relation between the two variables. Variation is measured using the dispersion or deviation of the realizations from their mean. In the first step, we center the observations:

$$x_k^* = (x_k - \bar{x})$$
$$y_k^* = (y_k - \bar{y}), \quad k = 1, \ldots, n$$

The common variation of both variables is the product of the deviations of the observations of their mean (see the calculation of the covariance):

$$\sum_{k=1}^{n} x_k^* y_k^* = \sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})$$

The scale on which each of the variables is measured and the number of observations can have a large impact on the magnitude of the common variation.

Assume the mean of one variable is 8 and the observed value is 10, and the mean of another variable is 1,008 and the observed value is 1,260. Then the deviation of the mean in the first variable is 2 and the deviation of the second is 252, the relative deviation of the mean value is in both cases 25 %. This fact may not have been observed if we simply calculated the common variation for this observation 504.

Therefore, in order to get similar deviations of the variables, we perform a standardization: $(x_k - \bar{x})/s_x$ and $(y_k - \bar{y})/s_y$. Now, change the above equation into:

$$\sum_{k=1}^{n} \frac{(x_k - \bar{x})}{s_x} \frac{(y_k - \bar{y})}{s_y}$$

We subsequently divide this sum of products by the number of observations in order to eliminate its influence. Now we have obtained the Bravais-Pearson (sample) correlation coefficient which measures the strength of the linear relation between the two continuous variables X and Y:

$$r_{yx} = r_{xy} = \frac{\sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})}{n \cdot s_x \cdot s_y} = \frac{s_{xy}}{s_x \cdot s_y}$$

The final parts of the above equation shows that the Bravais-Pearson correlation coefficient is equal to the variation common to both variables X and Y (= covariance) standardized by the product of the standard deviations of each variable.

The Bravais-Pearson correlation coefficient can also be written as follows:

$$r_{yx} = \frac{\sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{n} (x_k - \bar{x})^2 \sum_{k=1}^{n} (y_k - \bar{y})^2}}$$

$$= \frac{n \sum_{k=1}^{n} x_k y_k - \sum_{k=1}^{n} x_k \sum_{k=1}^{n} y_k}{\sqrt{\left[n \sum_{k=1}^{n} x_k{}^2 - \left(\sum_{k=1}^{n} x_k \right)^2 \right] \left[n \sum_{k=1}^{n} y_k{}^2 - \left(\sum_{k=1}^{n} y_k \right)^2 \right]}}.$$

Properties of the Correlation Coefficient

- The correlation coefficient only takes on values between -1 and $+1$:

$$-1 \leq r_{xy} \leq 1$$

- The sign of the correlation coefficient tells us the direction of the linear relation
  - "+" corresponds to a positive correlation (proportional variation)
  - "−" corresponds to a negative correlation (inverse proportional variation)
- If all observations are exactly on a straight line, the correlation coefficient is equal to 1 or −1.

  The more the absolute value of the correlation coefficient approaches 1, the more pronounced is the linear relation between the variables X and Y (and the other way round).
- If the variables X and Y are independent, then the correlation coefficient is equal to 0.

  On the other hand, a correlation coefficient of 0 only means that there is no linear relation between the variables X and Y (linear independence). But it is very well possible that there exists a pronounced nonlinear relation between both variables.
- The correlation coefficient is symmetric: $r_{xy} = r_{yx}$

Relation of Correlation and the Scatterplot of X and Y Observations

- Perfect correlation ($|r_{xy}| = 1$) (Fig. 10.10)
- Strong correlation ($|r_{xy}| > 0.5$) (Fig. 10.11)
- Weak correlation ($|r_{xy}| < 0.5$) (Fig. 10.12)
- No correlation ($r_{xy} = 0$) (Fig. 10.13)

  A correlation of 0 corresponds "in general" to some kind of a circular scatterplot point cloud.

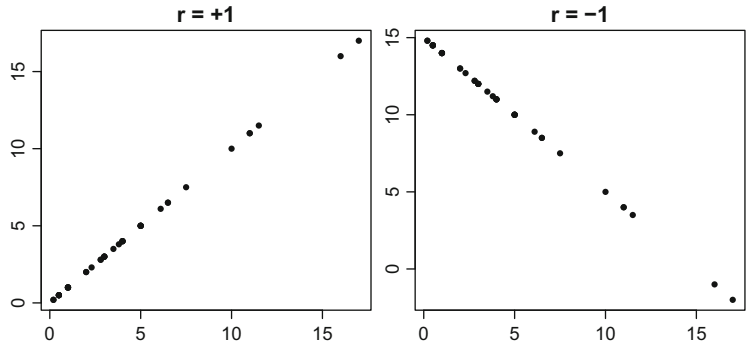


Fig. 10.10 Perfect correlation

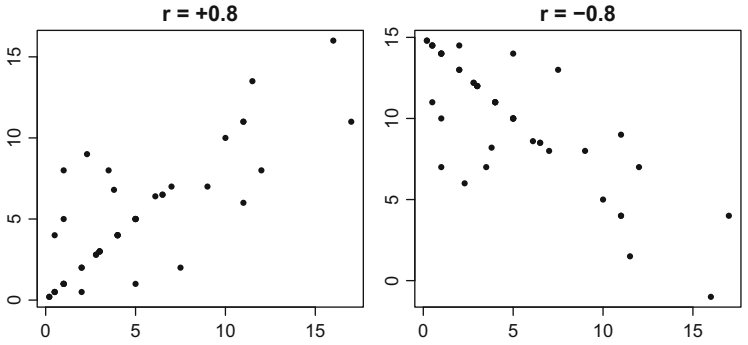


Fig. 10.11 Strong correlation

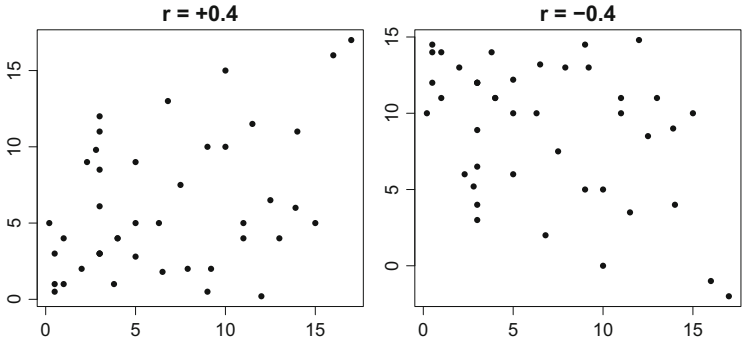


Fig. 10.12 Weak correlation

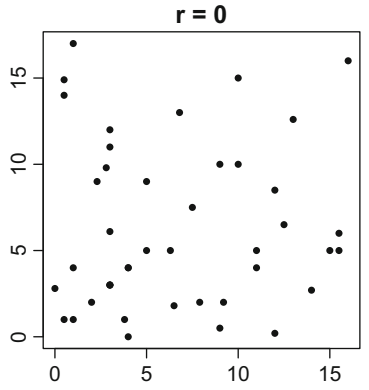


Fig. 10.13 No correlation

Example In $n = 15$ firms, we observed the variables Y—annual profit (in Mill. Euro) and X—annual rent for the computer facilities (in 1,000 Euro). You can

Table 10.19 Annual profits (in Mill. Euro) and annual rent for computer equipment (in 1,000 Euro) for 15 firms

Company k	Annual profit in Mill. EUR (y_k)	Annual rent in Tsd. EUR (x_k)
1	10	30
2	15	30
3	15	100
4	20	50
5	20	100
6	25	80
7	30	50
8	30	100
9	30	250
10	35	180
11	35	330
12	40	200
13	45	400
14	50	500
15	50	600

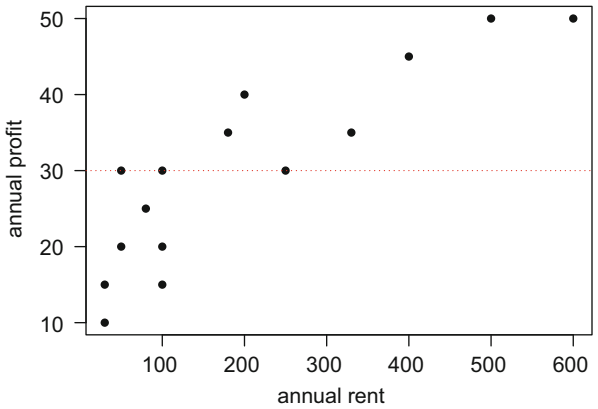


Fig. 10.14 Annual profits (in Mill. Euro) and annual rent for computer equipment (in Tsd. Euro) for 15 firms

see their variable values in Table 10.19. We also illustrate them graphically in a scatterplot as shown in Fig. 10.14.

From the observations, the following results can be obtained:

$$\bar{y} = 30, \quad \sum_{k=1}^{15} (y_k - \bar{y})^2 = 2{,}250$$
$$\bar{x} = 200, \quad \sum_{k=1}^{15} (x_k - \bar{x})^2 = 457{,}000$$

$$\sum_{k=1}^{15}(x_k - \bar{x})(y_k - \bar{y}) = 28{,}100$$

$$r_{xy} = \frac{28100}{\sqrt{(457000) \cdot (2250)}} = 0.8763$$

The sample correlation coefficient is in this example 0.8763. This points to a strong positive linear relation.

Explained: Relationship of Two Metrically Scaled Variables

In 1985, rates of criminal activity of the 50 states of the USA were recorded, among them murder rate. The relationship between the murder rate and the size of the population can be visualized by a scatterplot (Fig. 10.15).

The different sums of squared deviations (SSD) are calculated in the following way:

Sum of the products of deviations of "population" and "murder":

$$SSD(population \mid murder) = \sum(x_k - \bar{x})(y_k - \bar{y}) = 260,121.05$$

Sum of squared deviations for "population":

$$SSD(population) = \sum(x_k - \bar{x})^2 = 1,259,033,421.62$$

Sum of squared deviations for "murder":

$$SSD(murder) = \sum(y_k - \bar{y})^2 = 725.54$$

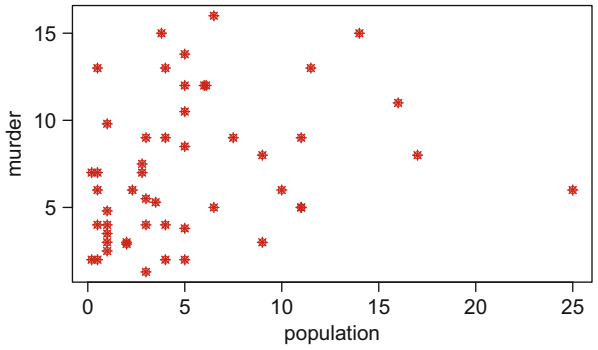


Fig. 10.15 Murder rate and size of population of 50 US states

The sample correlation coefficient is equal to

$$r = \frac{260,121.05}{\sqrt{(1,259,033,421.62) \cdot (725.54)}} = 0.27$$

The sample correlation coefficient of 0.27 points to a weak positive linear relationship.

Interactive: Correlation Coefficients

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please select a scatterplot type, e.g., scatterplot matrix. Moreover, choose which coefficient should be calculated and displayed.

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

Output

The interactive example allows us to display a two- or three-dimensional frequency distribution in the form of a 2D/3D-scatterplot or a scatterplot matrix. After choosing a set of variables (attention: three variables are required for the 3D-plot), the output window shows the corresponding scatterplot (see Fig. 10.16). In addition, we may choose to display a correlation coefficient.

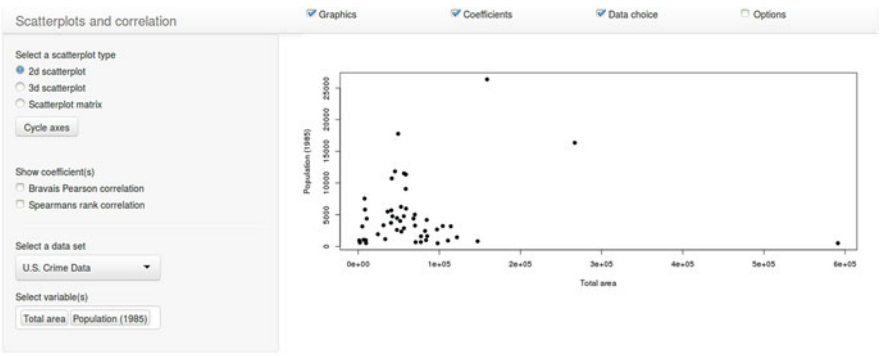


Fig. 10.16 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_corr

10.7 Relation Between Discrete Variables (Rank Correlation)

Spearman's Rank Correlation Coefficient

The starting point for the measurement of relationships of two discrete, or ordinal, variables X and Y are the ranks.

$$R(x_i), R(y_i), i = 1, \ldots, n$$

which are assigned to the observations x_i and y_j according to their rank. The ranks are defined so that $R(x_i)$ is equal to 1 for the x_i that takes on the largest value we have observed, is equal to 2 for the x_i that takes on the second largest value we have observed, and so on.

Spearman's rank correlation coefficient is computed from the pairs of ranks as follows:

$$r_s = 1 - \frac{6\sum_{i=1}^{n}[R(x_i) - R(y_i)]^2}{n(n^2-1)} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \quad d_i = R(x_i) - R(y_i)$$

Spearman's rank correlation coefficient amounts to applying the Bravais-Pearson correlation coefficient to the ranks (rather than the observations themselves).

It is true that:

$$\sum_{i=1}^{n} R(x_i) = \sum_{i=1}^{n} R(y_i) = \frac{n(n+1)}{2}$$

$$\sum_{i=1}^{n} R(x_i)^2 = \sum_{i=1}^{n} R(y_i)^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\sum_{i=1}^{n} R(x_i)R(y_i) = \frac{1}{2}\left[\sum_{i=1}^{n} R(x_i)^2 + \sum_{i=1}^{n} R(y_i)^2 - \sum_{i=1}^{n}(R(x_i) - R(y_i))^2\right]$$

The Bravais-Pearson Correlation Coefficient is calculated as:

$$r_{yx} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

If we use the corresponding ranks $R(x_i)$ and $R(y_i)$ instead of the observations x_i and y_i themselves, then we have derived Spearman's rank correlation coefficient:

$$r_{yx} = \frac{n\sum_{i=1}^{n} R(x_i)R(y_i) - \sum_{i=1}^{n} R(x_i)\sum_{i=1}^{n} R(y_i)}{\sqrt{\left[n\sum_{i=1}^{n} R(x_i)^2 - \left(\sum_{i=1}^{n} R(x_i)\right)^2\right]\left[n\sum_{i=1}^{n} R(y_i)^2 - \left(\sum_{i=1}^{n} R(y_i)\right)^2\right]}}$$

$$= \frac{n \cdot \frac{1}{2} \cdot 2\frac{n(n+1)(2n+1)}{6} - n \cdot \frac{1}{2}\sum_{i=1}^{n}[R(x_i) - R(y_i)]^2 - \frac{n^2(n+1)^2}{4}}{n \cdot \frac{n(n+1)(2n+1)}{6} - \frac{n^2(n+1)^2}{4}}$$

$$= 1 - \frac{6\sum_{i=1}^{n}[R(x_i) - R(y_i)]^2}{n(n+1)(n-1)} = r_s$$

Properties of Spearman's Rank Correlation Coefficient

- Spearman's rank correlation coefficient can only take on values between -1 and $+1$: $-1 <= r_s <= 1$.
- The rank correlation coefficient takes on the value +1 if the ranks behave exactly the same way, i.e., $R(x_i) = R(y_i)$ for all i.
- Spearman's rank correlation coefficient takes on the value -1, if the ranks are perfectly opposed to each other, i.e., $R(x_i) = n + 1 - R(y_i)$ for all i.

Example

- X—Ranking of an athlete in downhill skiing
- Y—Ranking of an athlete in slalom

Does there exist a relationship between the ranking in both disciplines?

The coefficient $r_s = 0.714$ points to a strong relationship between the ranking in both disciplines (Table 10.20).

Table 10.20 Ranking of athletes in downhill and slalom skiing

Athlete	1	2	3	4	5	6
Downhill $R(x_i)$	2	1	3	4	5	6
Slalom $R(y_i)$	2	3	1	5	4	6
d_i^2	0	4	4	1	1	0

Kendall's Rank Correlation Coefficient

Kendall's rank correlation coefficient is based on the comparison of the order relation for all possible pairs of observations of two variables. Concordant are the pairs of variables which show the same order relation, i.e., which show for both variables a low or high value. Discordant are the pairs which show a different order relation, that is which show in one of the variables a low and in the other variable a high value. Moreover, there can be pairs of variables, which are equal in terms of one value or both values. We call this bounding.

The number of concordant pairs P and discordant pairs Q can be calculated as follows:

- The variable pairs $R(x_i)$ a $R(y_i)$ are sorted in increasing order of $R(x_i)$.
- We call p_i the number of ranks prior to $R(y_i)$ which are larger than $R(y_i)$.
- We call q_i the number of the ranks subsequent to $R(y_i)$ which are smaller than $R(y_i)$.

Using the number of discordant and concordant variable pairs, we can calculate Kendall's rank correlation coefficient:

$$T = \frac{P - Q}{P + Q},$$

with $Q = \sum_i q_i$ and $P = \sum_i p_i$ The total number of all ranks to be compared is given by: $n(n-1)/2 = Q + P$. The correlation coefficient can only take on values between -1 and $+1$: $-1 <= \tau <= 1$.

An alternative way of calculating Kendall's rank correlation coefficient is given by:

$$T = 1 - \frac{4Q}{n(n-1)} = \frac{4P}{n(n-1)} - 1.$$

Example Ten employees have been ranked according to their managerial abilities (X) and their work ethic (Y). In order to make a statement about the relationship between both variables, we calculate both Spearmans' and Kendall's rank correlation coefficients (Tables 10.21 and 10.22).

Table 10.21 Ranking of employees according to managerial abilities (X) and their work ethic (Y) for Spearman's rank correlation coefficient

Employee	1	2	3	4	5	6	7	8	9	10
$R(X)$	7	3	9	10	1	5	4	6	2	8
$R(Y)$	3	9	10	8	7	1	5	4	2	6
$d_i{}^2$	16	36	1	4	36	16	1	4	0	4

Table 10.22 Ranking of employees according to managerial abilities (X) and their work ethic (Y) for Kendall's rank correlation coefficient

Employee	5	9	2	7	6	8	1	10	3	4
$R(X)$	1	2	3	4	5	6	7	8	9	10
$R(Y)$	7	2	9	5	1	4	3	6	10	8
q	6	1	6	3	0	1	0	0	1	0
p	3	7	1	3	5	3	3	2	0	0

Table 10.23 Standings of 20 athletes in the 100 m dash and 200 m dash

Athlete (i)	01	02	03	04	05	06	07	08	09	10
100 m	5	7	3	13	2	15	19	14	12	1
200 m	3	9	1	10	7	5	13	14	17	4
Athlete (i)	11	12	13	14	15	16	17	18	19	20
100 m	6	20	17	4	18	11	10	16	9	8
200 m	11	16	18	12	20	2	15	19	6	8

- **Spearman's rank correlation coefficient**

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

$$r_s = 1 - 6 \cdot 118/(10 \cdot 99) = 0.2848$$

- **Kendall's rank correlation coefficient**

$$Q = 18, \quad P = 27$$

$$Q + P = n(n-1)/2 = 10 \cdot 9/2 = 45$$

$$T = (27 - 18)/(27 + 18) = 9/45 = 0.200$$

Explained: Relationship Between Two Ordinally Scaled Variables

The standings of 20 athletes in the 100 m dash and 200 m dash are given in Table 10.23.

In what follows, the statistical relationship between the standings of the athletes in the two disciplines will be determined. Since the variables are ordinally scaled (discrete) we will use Spearman's and Kendall's rank correlation coefficients. Calculating both coefficients gives the following results:

```
Spearman's rank correlation coefficient:  0.6617


Kendall's rank correlation coefficient:  0.4526
    -concordant pairs          138
    -discordant pairs           52
    -identical wrt x             0
    -identical wrt y             0
    -identical wrt x & y         0
```

Spearman's coefficient is calculated as:

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

The information necessary to apply the formula can be obtained from the table—d is the difference between x_i and y_j, n is the number of athletes (= 20). The calculations produce a coefficient of 0.6617, which implies a positive relationship between the standings in the two disciplines—athletes doing well in the 100 m dash also tend to do well in the 200 m.

To calculate **Kendall's rank correlation coefficient**, one needs to determine the concordant and discordant pairs of athletes. A pair of observations (=athletes) is called concordant, if the same order relation applies to both variables and discordant if the order relations don't agree. For instance, athletes 1 and 2 are concordant: athlete 1 has a better standing than athlete 2 in both the 100 m dash and the 200 m dash. Athletes 1 and 5, however, are discordant: athlete 1 is behind in the 100 m but is ahead of athlete 5 in the standings of the 200 m dash. Overall, there are $\frac{n \cdot (n-1)}{2} = 190$ different pairs in this example, 138 of which are concordant while 52 are discordant. Using these numbers Kendall's rank correlation coefficient can be calculated:

$$\tau = \frac{P - Q}{P + Q},$$

where $Q = \sum_i q_i$ and $P = \sum_i p_i$.

Here, P is the number of concordant pairs and Q the number of discordant pairs. Kendall's rank correlation coefficient turns out to be 0.4526 in this example, which is an evidence for a positive relationship between the standings.

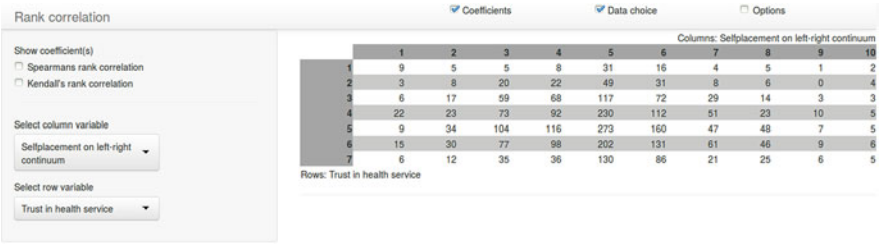


Fig. 10.17 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_rank

Interactive: Example for the Relationship Between Two Ordinally Scaled Variables

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please select which coefficient should be calculated and displayed.

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

Output

After choosing a set of variables, the output window in Fig. 10.17 shows the corresponding crosstable. In addition, this example allows us to calculate Spearmans' and Kendall's rank correlation coefficients for two series of ranks to be input by the user.

10.8 Relationship Between Nominal Variables (Contingency)

The starting point for the analysis of relationships between two nominal variables X and Y is the joint frequency distribution of X and Y put into a contingency table including the absolute frequencies $h_{ij} = h(x_i, y_j) \quad (i = 1, \ldots, m; j = 1, \ldots, r)$ or the relative frequencies $f_{ij} = f(x_i, y_j) = h(x_i, y_j)/n$ $(i = 1, \ldots, m; j = 1, \ldots, r)$.

As we showed in Sect. 10.5 the relative frequency for the joint appearance of realizations x_i and y_i $(i = 1, \ldots, m; j = 1, \ldots, r)$—in the case of independence—is equal to the product of the relative frequencies of the marginal distribution of both variables:

$$f_{ij} = f_{i\bullet} f_{\bullet j} \quad \text{and} \quad h_{ij} = \frac{h_{i\bullet} h_{\bullet j}}{n} = n f_{i\bullet} f_{\bullet j}$$

We can now calculate an auxiliary quantity—the squared contingency, represented by χ^2:

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{r} \frac{\left(h_{ij} - \frac{1}{n}h_{i\bullet}h_{\bullet j}\right)^2}{\frac{1}{n}h_{i\bullet}h_{\bullet j}} = n\sum_{i=1}^{m}\sum_{j=1}^{r} \frac{(f_{ij} - f_{i\bullet}f_{\bullet j})^2}{f_{i\bullet}f_{\bullet j}}$$

The numerator of the summands above form the squared deviations of the observed absolute (relative) frequencies from the expected absolute (relative) frequencies (if the variables are independent). Dividing by the expected absolute (relative) frequencies (if the variables are independent) we obtain a standardization.

We use the squared contingency to calculate the contingency coefficient as follows:

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

The contingency coefficient provides a measure of the strength of the relationship between nominal variables.

$$0 \leq C \leq \sqrt{\frac{C^* - 1}{C^*}}; \quad C^* = min(m, r).$$

If the contingency coefficient equals 0 we have statistical independence. The contingency coefficient almost never reaches 1 even when there is a perfect relationship between both variables because the sample size n is always larger than 0 and therefore the denominator is always larger than the numerator.

In order to solve this problem and to be able to reach the value 1 in case of a perfect relationship, we often use the corrected contingency coefficient which is calculated as follows:

$$C_{\text{corr}} = C \cdot \sqrt{\frac{C^*}{C^* - 1}} \qquad O \leq C_{\text{corr}} \leq 1$$

Example We want to analyze if there is a relationship between smoking and lung cancer. We use the contingency table given in Table 10.24.

Table 10.24 Smoking and lung cancer

	Lung cancer		
Smoker	Yes(y_1)	No(y_2)	MD X
Smoker yes (x_1)	10	15	$h_{1\cdot}$=25
Smoker no (x_2)	5	70	$h_{2\cdot}$=75
MD Y	$h_{\cdot 1}$=15	$h_{\cdot 2}$=85	n=100

$$\chi^2 = \frac{\left(10 - \frac{15\cdot(25)}{100}\right)^2}{\frac{15\cdot(25)}{100}} + \frac{\left(15 - \frac{85\cdot(25)}{100}\right)^2}{\frac{85\cdot(25)}{100}} + \frac{\left(5 - \frac{15\cdot(75)}{100}\right)^2}{\frac{15\cdot(75)}{100}} + \frac{\left(70 - \frac{85\cdot(75)}{100}\right)^2}{\frac{85\cdot(75)}{100}} = 16.34$$

$$C = \sqrt{\frac{16.34}{100 + 16.34}} = 0.375$$

$$C_{\text{corr}} = 0.375 \cdot \sqrt{\frac{2}{2-1}} = 0.53$$

The corrected contingency coefficient of 0.53 is evidence for a relationship between smoking and lung cancer.

Explained: Relationship Between Two Nominally Scaled Variables

The "department store" data set contains the following variables recorded for $n = 165$ randomly selected customers:

Variable	Possible realizations
X gender	1—male
	2—female
Y method of payment	1—cash
	2—ATM card
	3—credit card
Z residence	1—Berlin
	2—not in Berlin

Below, the three possible two-dimensional frequency distributions are shown that can be formed for the variables in this data set. The contingency coefficient is calculated each time.

The two-dimensional frequency distribution for the variables **gender** and **method of payment** is a 2×3 contingency table (Table 10.25).

χ^2 statistic	0.08
Contingency coefficient	0.02
Corrected contingency coefficient	0.03

The corrected contingency coefficient of 0.03 shows that there is only a very weak relationship between gender and method of payment.

Table 10.25 Two-dimensional frequency distribution for gender and method of payment

	Method of payment (Y)			
Gender (X)	(y_1)	(y_2)	(y_3)	MD X
Male (x_1)	31 (0.188)	32 (0.194)	23 (0.139)	86 (0.521)
Female (x_2)	30 (0.182)	29 (0.176)	20 (0.121)	79 (0.479)
MD Y	61 (0.370)	61 (0.370)	43 (0.260)	165 (1.000)

Table 10.26 Two-dimensional frequency distribution for gender and residence

	Residence (Z)		
Gender (X)	Berlin (z_1)	Not in Berlin (z_2)	MD X
Male (x_1)	50 (0.303)	36 (0.218)	86 (0.521)
Female (x_2)	37 (0.224)	42 (0.255)	79 (0.429)
MD Y	87 (0.527)	78 (0.473)	165 (1.000)

Table 10.27 Two-dimensional frequency distribution for residence and method of payment

	Method of payment (Y)			
Residence (Z)	(y_1)	(y_2)	(y_3)	MD X
Berlin (z_1)	44 (0.267)	22 (0.133)	21 (0.127)	87 (0.527)
Not in Berlin (z_2)	17 (0.103)	39 (0.237)	22 (0.133)	78 (0.473)
MD Y	62 (0.370)	61 (0.370)	43 (0.260)	165 (1.000)

The two-dimensional frequency distribution for the variables **gender** and **residence** is a 2×2 contingency table (Table 10.26).

χ^2 statistic	2.11
Contingency coefficient	0.11
Corrected contingency coefficient	0.16

The corrected contingency coefficient of 0.16 shows that there is only a weak relationship between gender and residence.

The two-dimensional frequency distribution for the variables **residence** and **method of payment** is a 2×3 contingency table (Table 10.27).

χ^2 statistic	16.27
Contingency coefficient	0.30
Corrected contingency coefficient	0.42

The corrected contingency coefficient of 0.42—being considerably larger than in the previous two cases—shows that there is a medium strength relationship between residence and method of payment.

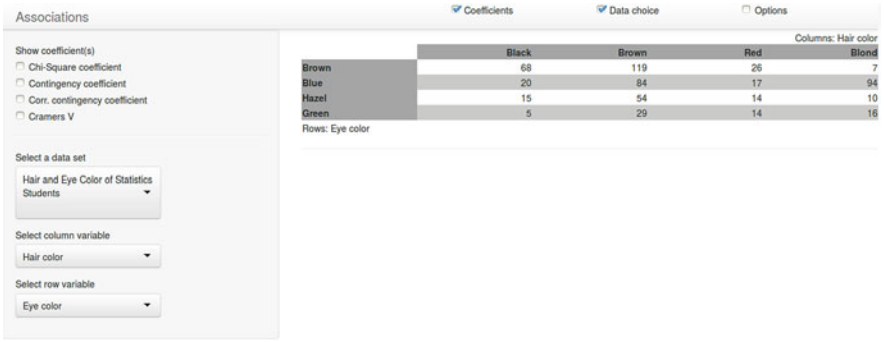


Fig. 10.18 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_asso

Interactive: Example for the Relationship Between Two Nominally Scaled Variables

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please select which coefficient you like to be calculated and displayed.

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

Output

After choosing a set of variables, the output window in Fig. 10.18 shows the corresponding crosstable. In addition, this example allows us to calculate the Chi-Square coefficient, Contingency coefficient, Corr. contingency coefficient, and Cramers V for the pre-selected variables.

Chapter 11
Regression

11.1 Regression Analysis

The Objectives of Regression Analysis

The main objective of regression analysis is to describe the expectation and dependence of a quantity Y on quantities $X_1, X_2, \ldots$. A one-directional dependence is assumed. This dependence can be expressed as a general regression function of the following form:

$$E(y|x) = f(x_1, x_2, \ldots).$$

The symbol $E(y|x)$ indicates that the regression function of observed values $x_1, x_2, \cdots$ does not correspond to an observed value y, but rather, it is the average value of y given the x_i's, which lies on the regression function.

The random variables $X_1, X_2, \ldots$ are referred to as **regressors, explanatory variables**, or **independent variables**.

The random variable Y is referred as **regressand** or **dependent variable**.

An example is the simple linear regression with a dependent variable "Time working" and one independent variable "Amount of production." Notice that this regression is referred to as simple because there is a single independent variable and is a linear regression since the function $f(\textit{Amount of production})$ is assumed to be linear (Fig. 11.1).

If the dependence of $\mathbf{Y}$ on $\mathbf{X}$ can be represented by a linear function, the regression value $E(y_i|x_i)$ does describe the expected value of Y given $X = x_i$. It follows that the value of any observation i can be decomposed as follows:

$$y_i = E(y_i|x_i) + u_i \quad i = 1, \ldots, n$$

W.K. Härdle et al., *Introduction to Statistics*, DOI 10.1007/978-3-319-17704-5_11

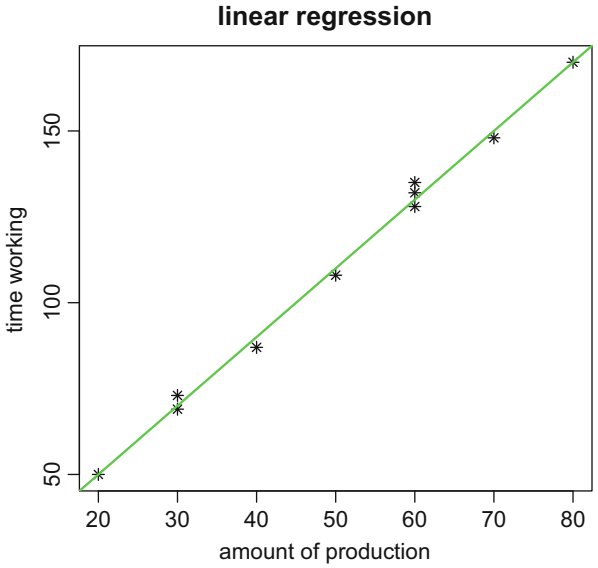


Fig. 11.1 Example of a simple linear regression

The difference between the observed values y_i and the value of the regression function $E(y_i|x_i)$ is called a residual u_i. It contains those influences on y_i that cannot be described by means of the regression function; alternatively, this means that deviations of the observed values from the regression function cannot be explained by the independent variables employed in the regression function.

$$u_i = y_i - E(y_i|x_i) \quad i = 1, \cdots, n$$

Regression Function

The regression function is a representation of average statistical dependence of a dependent variable on one or more independent variables. The dependence is described by a function based on n observations.

In what follows, we assume only the case when a variable Y depends on a single variable X. The form of the regression function $f(x)$ always depends on the specific application and the purpose of an analysis.

Examples of possible regression functions include:

Linear function	$E(y\|x) = b_0 + b_1 x$
Quadratic function	$E(y\|x) = b_0 + b_1 x + b_2 x^2$
Power function	$E(y\|x) = ax^b$
Exponential function	$E(y\|x) = b_0 {b_1}^x$
Logarithmic function	$E(y\|x) = kl(1 + e^{a+bx})$

11.2 One-Dimensional Regression Analysis

One-Dimensional Linear Regression Function

A simple linear regression function has the following form:

$$E(y_i|x_i) = b_0 + b_1 x_i \quad i = 1, \ldots, n$$

In this equation, x_i represents the observed values of a random variable **X** (fixed) and b_0 and b_1 are unknown regression parameters.

The actual observed values $y_i (i = 1, \ldots, n)$ can be obtained by summing residual u_i and $E(y_i|x_i)$ (as you can see in Fig. 11.2):

$$y_i = E(y_i|x_i) + u_i = b_0 + b_1 x_i + u_i \quad i = 1, \ldots, n$$

Regression Parameters

Parameters of a simple linear regression function have the following meaning:

- b_0—intercept term (constant)
  It describes the intersection of the corresponding regression line and the y-axis and it has the same value as variable Y at this point.
- b_1—linear slope coefficient (also a constant)
  It characterizes the slope of the corresponding regression line. It tells us by how many units the expected value of random variable Y will change if the value of variable X is increased by one unit.

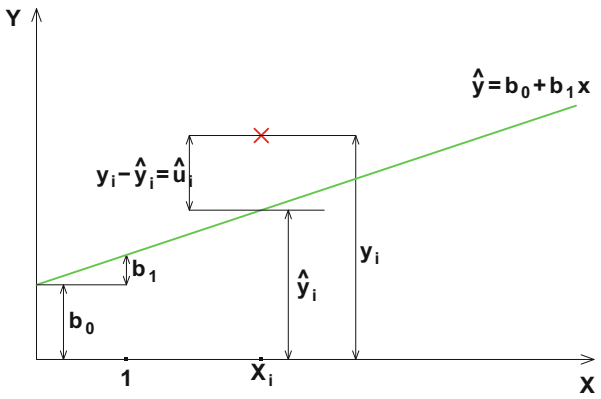


Fig. 11.2 Components in linear regression analysis

Estimation of Regression Parameters

To estimate regression parameters, two important conditions have to be satisfied.

1*st Condition*

The deviations of estimated regression values $\widehat{y_i}$ from observed values y_i should be on average equal to zero; that is

$$\sum_{i=1}^{n}(y_i - \widehat{y_i}) = \sum_{i=1}^{n}\widehat{u_i} = 0$$

$$\bar{\hat{u}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{u_i} = 0$$

However, this condition is satisfied for infinitely many regression lines, namely those that go through the point of sample means $\bar{x}, \bar{y}$ (Fig. 11.3). Notice that the above expressions imply that $y_i = \widehat{y_i} + \widehat{u_i}$. Therefore, for each observation i we have decomposed the observed y_i into two parts: (1) an estimated regression function $\widehat{y_i} = \widehat{E(y_i|x_i)}$ (i.e., an estimate of the conditional mean); and (2) an estimated residual $\widehat{u_i}$ (disturbance).

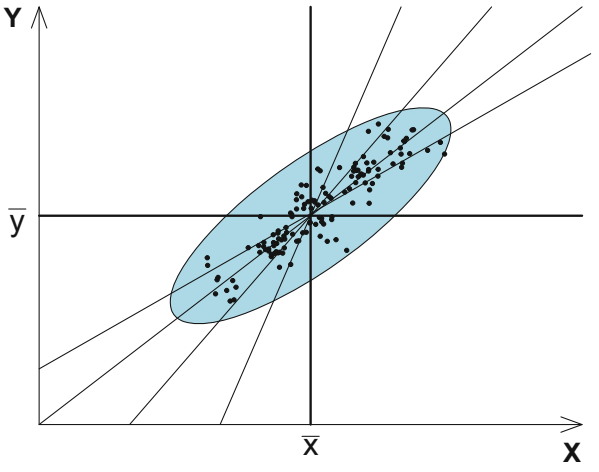


Fig. 11.3 Possible regression lines without 2^{nd} condition

2nd Condition

We search for a regression line such that the spread (variance) of the corresponding estimated residuals (called disturbances)

$$s^2_{\hat{u}} = \frac{1}{n-2}\sum_{i=1}^{n}(\widehat{u}_i - \bar{\hat{u}})^2$$

is minimal in comparison with all other possible regression lines.

The first condition

$$\bar{\hat{u}} = 0$$

implies

$$s^2_{\hat{u}} = \frac{1}{n-2}\sum_{i=1}^{n}(\widehat{u}_i - 0)^2 = \frac{1}{n-2}\sum_{i=1}^{n}\widehat{u}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i))^2.$$

The second condition is illustrated in Fig. 11.4.

The squares drawn in the figure correspond to the squared residuals and the total area of the squares should be minimized. Hence, the approach is called the least squares (LS) method.

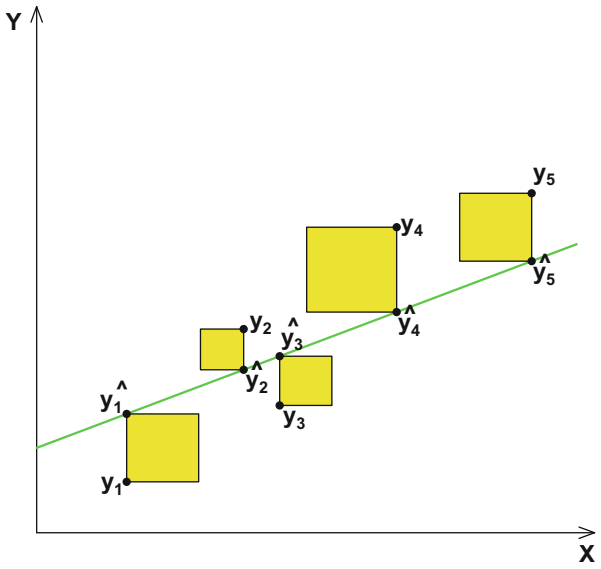


Fig. 11.4 Illustration of 2nd condition

The least squares method minimizes the sum of squared deviations of regression values from the observed values (residual sum of squares—RSS)

$$\sum_{i=1}^{n} (y_i - \widehat{y_i}))^2 \rightarrow min. \quad | \; E(y_i|x_i) = b_0 + b_1 x_i.$$

The minimized function has two unknown variables (b_0 and b_1).

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \rightarrow min. \quad E(y_i|x_i) = b_0 + b_1 x_i$$

To find a minimum, the first partial derivatives have to be equal to zero.

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \rightarrow min.$$

$$\frac{\partial S(b_0, b_1)}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) \doteq 0$$

$$\frac{\partial S(b_0, b_1)}{\partial b_1} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) x_i \doteq 0$$

To verify whether the solution is really a minimum, the second partial derivatives have to be evaluated.

$$\frac{\partial^2 S(b_0, b_1)}{\partial {b_0}^2} = 2n > 0$$

$$\frac{\partial^2 S(b_0, b_1)}{\partial {b_1}^2} = 2 \sum_{i=1}^{n} {x_i}^2 > 0$$

Since both of the second order derivatives are positive, the extremum will always be a minimum.

The first order derivatives (equal to zero) lead to the so-called **(least squares) normal equations**. By solving these equations, the estimated regression parameters ($\widehat{b_0}$ and $\widehat{b_1}$) can be computed.

$$n\widehat{b_0} + \widehat{b_1} \sum_{i=1}^{n} x_i = \sum_{i=1} n y_i$$

$$\widehat{b_0} \sum_{i=1}^{n} x_i + \widehat{b_1} \sum_{i=1}^{n} {x_i}^2 = \sum_{i=1}^{n} x_i y_i$$

The normal equations can be solved by means of linear algebra (Cramer's rule):

$$\widehat{b_0} = \frac{\begin{vmatrix} \sum y_i & \sum x_i \\ \sum x_i y_i & \sum {x_i}^2 \end{vmatrix}}{\begin{vmatrix} n & \sum x_i \\ \sum x_i & \sum {x_i}^2 \end{vmatrix}} = \frac{\sum y_i \sum {x_i}^2 - \sum x_i \sum x_i y_i}{n \sum {x_i}^2 - \sum x_i \sum x_i}$$

$$\widehat{b_1} = \frac{\begin{vmatrix} n & \sum y_i \\ \sum x_i & \sum x_i y_i \end{vmatrix}}{\begin{vmatrix} n & \sum x_i \\ \sum x_i & \sum {x_i}^2 \end{vmatrix}} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum {x_i}^2 - \sum x_i \sum x_i}$$

Dividing the original equations by n, we get a simplified formula suitable for the computation of regression parameters:

$$\widehat{b_0} + \widehat{b_1}\bar{x} = \bar{y}$$
$$\widehat{b_0}\bar{x} + \widehat{b_1}\bar{x^2} = \overline{xy}$$

For the estimated intercept $\widehat{b_0}$, we get:

$$\widehat{b_0} = \bar{y} - \widehat{b_1}\bar{x}$$

For the estimated linear slope coefficient $\widehat{b_1}$, we get:

$$\begin{aligned} (\bar{y} - \widehat{b_1}\bar{x})\bar{x} + \widehat{b_1}\bar{x^2} &= \overline{xy} \\ \widehat{b_1}(\bar{x^2} - \bar{x}^2) &= \bar{x}\bar{y} - \overline{xy} \\ \widehat{b_1} {S_X}^2 &= S_{XY} \\ \widehat{b_1} &= \frac{S_{XY}}{{S_X}^2} \end{aligned}$$

Properties

- The sample variance of X must be greater than zero: ${S_X}^2 > 0$
- From the simplified normal equations, you can see that: $(\bar{x}, \bar{y}) \rightarrow$ if $x_i = \bar{x}$ then $\widehat{y_i} = \bar{y}$

$$\widehat{y_i} = \widehat{b_0} + \widehat{b_1} x_i = \bar{y} + \widehat{b_1}(x_i - \bar{x}) = \bar{y}$$

Table 11.1 Production output and working time

i	x_i	y_i	$x_i y_i$	${x_i}^2$	${y_i}^2$	$\widehat{y_i}$	$\widehat{u_i}$
1	30	73	2,190	900	5,329	70	3
2	20	50	1,000	400	2,500	50	0
3	60	128	7,680	3,600	16,384	130	−2
4	80	170	1,360	6,400	28,900	170	0
5	40	87	3,480	1,600	7,569	90	−3
6	50	108	5,400	2,500	11,664	110	−2
7	60	135	8,100	3,600	18,225	130	5
8	30	69	2,070	900	4,761	70	−1
9	70	148	10,360	4,900	21,904	150	−2
10	60	132	72,920	3,600	17,424	130	2
$\sum$	500	1,100	61,800	28,400	134,660	1,100	0

- Combining results from correlation and regression analysis, it is possible to obtain the estimated linear slope coefficient $\widehat{b_1}$ as follows:

$$\widehat{b_1} = \frac{S_{xy}}{{S_x}^2}, \quad r_{xy} = \frac{S_{xy}}{S_x S_y}$$

$$\Rightarrow \widehat{b_1} = r_{xy}\frac{S_y}{S_x}$$

The regression $(y|x)$ of y on x does **not correspond** to the regression $(x|y)$ of x on y.

$$\widehat{b_0} = \bar{y} - \widehat{b_1}\bar{x} \qquad \widehat{b_0}^* = \bar{x} - \widehat{b_1}^*\bar{y}$$
$$\widehat{b_1} = \frac{S_{XY}}{{S_X}^2} \qquad \widehat{b_1}^* = \frac{S_{XY}}{{S_Y}^2}$$

Example

- X—production output
- Y—working time
- $n = 10$ production cycles in a firm

Computation of auxiliary variables (sample mean, sample variance, and sample standard deviation) (Table 11.1):

$$\bar{x} = 50 \qquad s_x^2 = 3,400/10 = 340 \qquad s_x = 18.44$$
$$\bar{y} = 110 \qquad s_x^2 = 13,660/10 = 13,366 \qquad s_y = 36.96$$

sample Covariance and sample correlation coefficient equal:

$$s_{xy} = 6,800/10 = 680 \quad r_{xy} = 680/(18.44 \cdot 36.96) = 0.9977$$

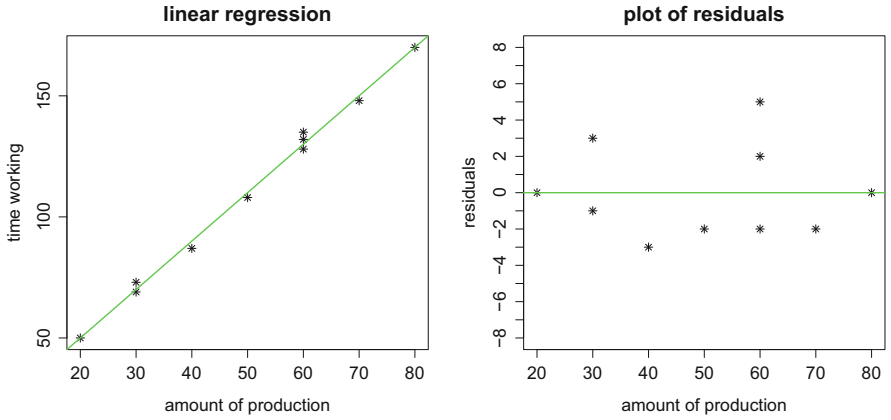


Fig. 11.5 Linear regression line and residual plot for production and working time

From these values, we can compute the estimated regression coefficients $\widehat{b_0}$ and $\widehat{b_1}$:

$$\widehat{b_1} = 680/340 = 2$$
$$\widehat{b_0} = 110 - 2 \cdot (50) = 10$$

As a result, we obtain the following estimated regression line (Fig. 11.5):

$$\widehat{y_i} = 10 + 2x_i$$

Quality (Fit) of the Regression Line

Once the regression line is estimated, it is useful to know how well the regression line approximates the observed data, that is, how good the representation of the data by means of the regression line is.

A measure that can describe the quality of representation is called the coefficient of determination (or R-Squared R^2). Its computation is based on a decomposition of the variance of the dependent variable Y.

The smaller is the sum of squared estimated residuals, the better is the quality (fit) of the regression line. Since the least squares approach minimizes the variance of the estimated residuals, it also maximizes the R^2 by construction.

$$\sum (y_i - \widehat{y_i})^2 = \sum \hat{u_i}^2 \rightarrow min.$$

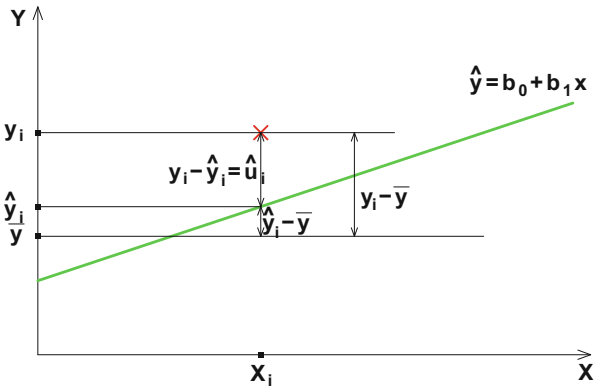


Fig. 11.6 Decomposition of observed values

The sample variance of Y is:

$$s_y{}^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}$$

The deviation of the observed values y_i from the arithmetic mean $\bar{y}$ can be decomposed to two parts: the deviation of the observed values y_i from the estimated regression values and the deviation of the estimated regression values from the sample mean.

$$y_i - \bar{y} = [(y_i - \widehat{y_i}) + (\widehat{y_i} - \bar{y})], \quad i = 1, \cdots, n$$

This decomposition is depicted in Fig. 11.6.
Analogously, the sum of the squared deviations can be decomposed:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} [(y_i - \widehat{y_i}) + (\widehat{y_i} - \bar{y})]^2$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 + \sum_{i=1}^{n} (\widehat{y_i} - \bar{y})^2$$

We were able to derive the second equation above by noting that $\sum_{i=1}^{n} (y_i - \widehat{y_i})(\widehat{y_i} - \bar{y}) = 0$. The reader is urged to prove this using the second least square first order condition above along with the definition of $\widehat{y_i}$.

Dividing both sides of the second equation by n, it follows:

$$\frac{\sum_i^n (y_i - \bar{y})^2}{n} = \frac{\sum_{i=1}^n (y_i - \widehat{y_i})^2}{n} + \frac{\sum_{i=1}^n (\widehat{y_i} - \bar{y})^2}{n}$$

$$\frac{\sum_i^n (y_i - \bar{y})^2}{n} = \frac{\sum_{i=1}^n \widehat{u_i}^2}{n} + \frac{\sum_{i=1}^n (\widehat{y_i} - \bar{y})^2}{n}$$

$$S_y{}^2 = S_{\hat{u}}{}^2 + S_{\hat{y}}{}^2$$

The total sample variance of Y is equal to (can be decomposed into) the sum of the sample variance of the estimated residuals (the unexplained part of the variance of Y) and the part of the variance of Y that is explained by the regression function (the sample variance of the regression function).

It holds:

- The larger the portion of the sample variance y as explained by the model is (i.e., $S_{\hat{y}}{}^2$), the better the fit of the regression function.
- On the other hand, the larger the residual variance $\widehat{S_u{}^2}$ as a percentage of the sample variance of y, alternatively the larger the outside influences unexplained by the regression function are, the worse the regression function fits.

The Coefficient of Determination

The coefficient of determination is defined as the ratio of the (sample) variance Y explained by the regression function and the total (sample) variance of Y. That is, it represents the proportion of the sample variance in y "explained" by the estimated regression function.

$$R_{yx}^2 = \frac{\sum_{i=1}^n (\widehat{y_i} - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{S_{\hat{y}}{}^2}{S_y{}^2}$$

An alternative way for computing the coefficient of determination is:

$$R_{yx}^2 = \frac{[\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})]^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}{}^2}{S_y{}^2 S_x{}^2}$$

$$R_{xy}^2 = \frac{(n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i)^2}{[n \sum_{i=1}^n x_i{}^2 - (\sum_{i=1}^n x_i)^2][n \sum_{i=1}^n y_i{}^2 - (\sum_{i=1}^n y_i)^2]}$$

Characteristics:

- The coefficient of determination has the following domain: $0 \leq R_{yx}^2 \leq 1$

  The higher the coefficient of determination is, the better the regression function explains the observed values.

  If all observed values lie on the regression line, the coefficient of determination is equal to 1. The total variance of Y can be explained by the variable X. Y depends completely (and linearly) on X.

  If the coefficient of determination is zero, the total variance of Y is identical with the unexplained variance (the residual variance). The random variable X does not have any linear influence on Y.
- $R_{xy}^2 = R_{yx}^2$ Symmetry (the fit of the regression of y on x is identical to the fit of the regression of x on y)
- For a linear regression function, the coefficient of determination corresponds to the square of the correlation coefficient: $R_{yx}^2 = r_{yx}^2$.

Example For the above described dependence between the working time and the production output, the sample correlation coefficient and the coefficient of determination are:

$$r_{yx} = 0.9977$$
$$R_{yx}{}^2 = 0.9954$$

One-Dimensional Nonlinear Regression Function

Example

- $n = 8$ comparable towns
- X—the number of the public-transportation maps that are distributed for free among citizens of the city at the beginning of the analyzed time period.
- Y—increase in the number of citizens using public transport during the analyzed time period (Table 11.2).

Linear Regression

$$\widehat{y_i} = \widehat{b_0} + \widehat{b_1} x_i = -1.82 + 0.0435 x_i$$
$$R_{yx}{}^2 = 0.875$$

Table 11.2 Data on X and Y

Town i	Increase Y (in 1,000)	Public-transportation maps X (in 1,000)
1	0.60	80
2	6.70	220
3	5.30	140
4	4.00	120
5	6.55	180
6	2.15	100
7	6.60	200
8	5.75	160

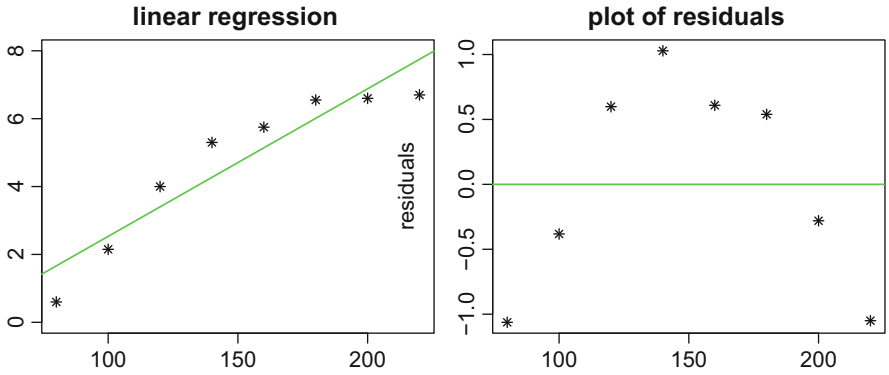


Fig. 11.7 Linear model for X and Y

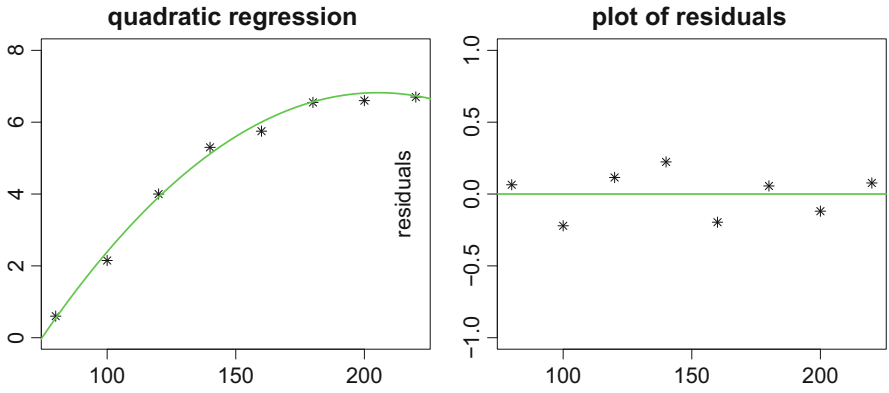


Fig. 11.8 Nonlinear model for X and Y

As we see from Fig. 11.7 the estimated residuals are not randomly dispersed around zero, but instead they have a rather clear nonlinear pattern. Hence, it can be beneficial to use a nonlinear regression model instead of the linear one (Fig. 11.8).

Quadratic Regression: Second-Order Polynomial

$$\widehat{y_i} = \widehat{b_0} + \widehat{b_1} x_i + \widehat{b_2} {x_i}^2 = -10.03 + 0.1642 x_i - 0.0004 {x_i}^2$$
$${R_{yx}}^2 = 0.995$$

Explained: One-Dimensional Linear Regression

Now, we examine the monthly net income and monthly expenditures on living of 10 two-person households (Table 11.3).

These observations are drawn in the following scatterplot. You can see that the net income of a household has a positive influence of the household's expenditures and that this dependence can be estimated by means of a linear regression function.

We want to estimate a linear regression function describing expenditures of a household as a function of the household's net income (Fig. 11.9).

To estimate the linear regression model, some auxiliary calculations are needed (Table 11.4).

Using the derived formulas, the estimated regression parameters $\widehat{b_0}$ and $\widehat{b_1}$ are computed as follows:

$$\begin{aligned}
\widehat{b_0} &= \frac{\sum y_i \sum {x_i}^2 - \sum x_i \sum x_i y_i}{n \sum {x_i}^2 - \sum x_i \sum x_i} \\
&= \frac{(25,400 \cdot 179,330,000) - (39,700 \cdot 112,420,000)}{(10 \cdot 179,330,000) - (39,700 \cdot 39,700)} \\
&= 423.13 \\
\widehat{b_1} &= \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum {x_i}^2 - \sum x_i \sum x_i} \\
&= \frac{(10 \cdot 112,420,000) - (39,700 \cdot 25,400)}{(10 \cdot 179,330,000) - (39,700 \cdot 39,700)} \\
&= 0.5332
\end{aligned}$$

Table 11.3 Data on monthly net income and monthly expenditures for 10 two-person households

Household	1	2	3	4	5
Net income in EUR x_i	3,500	5,000	4,300	6,100	1,000
Expenditures in EUR y_i	2,000	3,500	3,100	3,900	900
Household	6	7	8	9	10
Net income in EUR x_i	4,800	2,900	2,400	5,600	4,100
Expenditures in EUR y_i	3,000	2,100	1,900	2,900	2,100

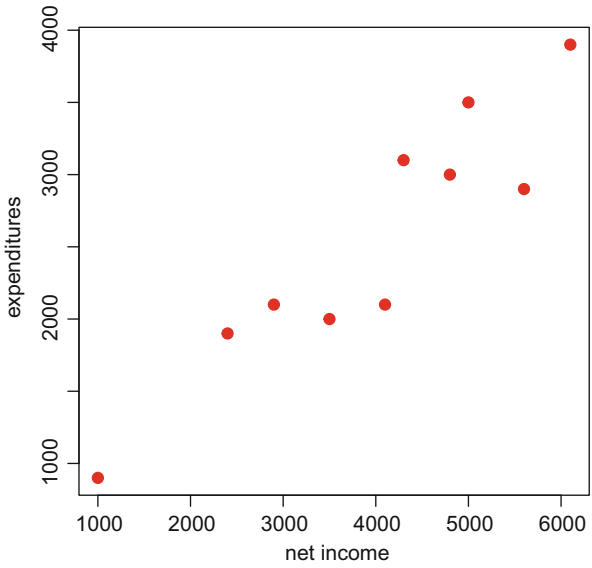


Fig. 11.9 Scatterplot of monthly net income and monthly expenditures

Table 11.4 Auxiliary calculations for linear regression analysis

HH	x_i	y_i	$x_i \cdot y_i$	${x_i}^2$	${y_i}^2$
1	3,500	2,000	7,000,000	12,250,000	4,000,000
2	5,000	3,500	17,500,000	25,000,000	12,250,000
3	4,300	3,100	13,330,000	18,490,000	9,610,000
4	6,100	3,900	23,790,000	37,210,000	15,210,000
5	1,000	900	900,000	1,000,000	810,000
6	4,800	3,000	14,400,000	23,040,000	9,000,000
7	2,900	2,100	6,090,000	8,410,000	4,410,000
8	2,400	1,900	4,560,000	5,760,000	3,610,000
9	5,600	2,900	16,240,000	31,360,000	8,410,000
10	4,100	2,100	8,610,000	16,810,000	4,410,000
$\sum$	39,700	25,400	112,420,000	179,330,000	71,720,000

Thus, the estimated regression function is

$$\widehat{y_i} = 423.13 + 0.5332 \cdot x_i$$

Expenditures = 423.13 + 0.5332 · Net income

The estimated regression line can be drawn in the scatterplot as shown in Fig. 11.10.

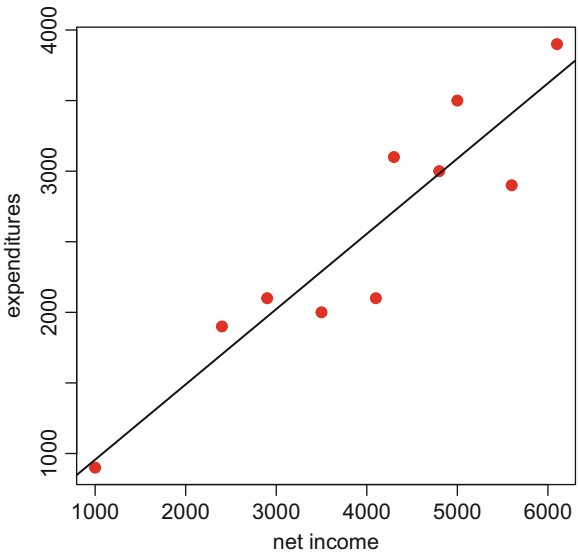


Fig. 11.10 Estimated regression line for monthly net income and monthly expenditures

The slope of the line corresponds to the marginal propensity to consume: an increase in the net income by one Mark (1 EUR) translates on average to 0.53 EUR increase in expenditures for the observed households.

Once sample standard deviations of x and y and their sample covariance are computed, we can readily obtain the sample correlation coefficient:

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{1,286,900}{1,553.5 \cdot 894.68} = 0.926$$

It hints to a strong (positive) dependence between households' net incomes and living expenditures.

The quality of the fit of the regression function can be evaluated via the coefficient of determination. It is a ratio of the variance explained by the regression function and the total sample variance of expenditures Y:

$$R^2 = \frac{\sum (\widehat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{6,175,715.85}{7,204,000.00} = 0.857$$

The coefficient of determination shows that 86% of the variation in households' expenditures can be explained by a linear dependence on the household's net incomes.

Enhanced: Crime Rates in the US

In year 1985, information about various crimes in each of 50 states of the USA was collected, including data on:

$X1$—land area
$X2$—population
$X3$—murder
$X4$—rape
$X5$—robbery
$X6$—assault
$X7$—burglary
$X8$—larceny
$X9$—auto-theft
$X10$—US states region number
$X11$—US states division number

The dependence of **robbery (X5)** on the **population (X2)** of a state can be depicted in a scatterplot. Every state is represented in the diagram by a single point $(X2, X5)$. Moreover, an estimated **regression line** is added in Fig. 11.11 (it is drawn in black).

The regression analysis provides the following results:

- The estimated regression intercept is 48.1134. In this case, it does not make sense to interpret this number; $\widehat{b_0}$ is a kind of correction parameter.
- The increase in the population of a state by one unit (i.e., by 1,000 citizens) leads to the increase in the number of robberies by $\widehat{b_1} = 0.0112$.

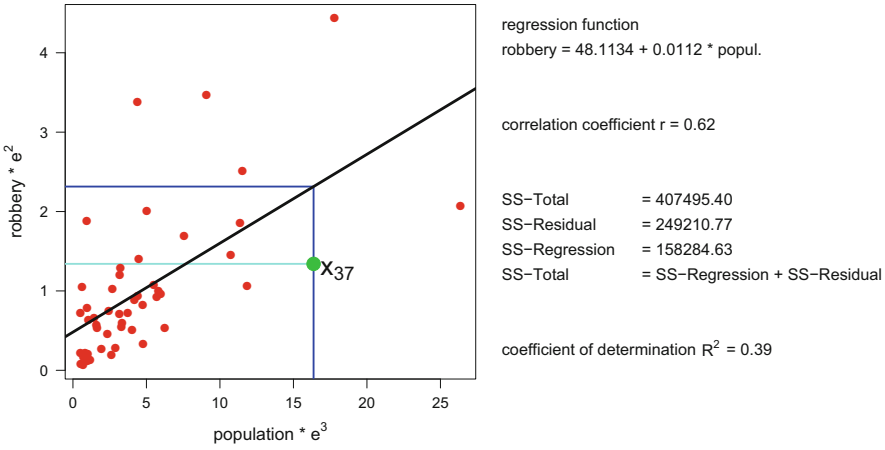


Fig. 11.11 Linear regression analysis of robbery (X5) and population (X2)

- The sample correlation coefficient is 0.62—this implies a (positive) dependence of the population and the number of robberies.
- To estimate the fit of the estimated regression function, the coefficient of determination can be used. Its calculation is based on the decomposition of the sample variance of the dependent variable. For the calculation, we can use the total sample variance (SS-Total), the unexplained (residual) variance (SS-Residual), and the explained variance (SS-Regression). Using the formula

$$R^2 = \frac{SS - Regression}{SS - Total} = \frac{\sum (\widehat{y_i} - \bar{y})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{SS - Residual}{SS - Total},$$

  we get that the coefficient of determination equals 0.39. The regression line does not characterize the observed values very well, the explanatory power of the model is weak.

The observation $x(37)$ corresponds to the population of 16,370 thousands and the number of robberies 134.1. The estimated regression function for such a state predicts the number of robberies to be equal to 231.66.
Note: The interactive example allows you to display (graphically) the pairwise dependence of other variables as well.

Enhanced: Linear Regression for the Car Data

The following measures were collected for 74 different types of cars:

$X1$—price
$X2$—mpg (miles per gallon)
$X3$—headroom (in inches)
$X4$—rear seat clearance
(distance from front seat back to the rear seat, in inches)
$X5$—trunk space (in cubic feet)
$X6$—weight (in pound)
$X7$—length (in inches)
$X8$—turning diameter (clearance required to make a U-turn, in feet)
$X9$—displacement (in cubic inches)

The dependence of **turning diameter (X8)** on the **length (X7)** of a car can be depicted in a scatterplot. Every car is represented in the diagram by a single point $(X7, X8)$. Moreover, an estimated **regression line** is added in Fig. 11.12 (it is drawn in black).

The regression analysis provides the following results:

- The estimated regression intercept is 7.1739. In this case, it may not make sense to interpret this number; $\widehat{b_0}$ is a kind of correction parameter.

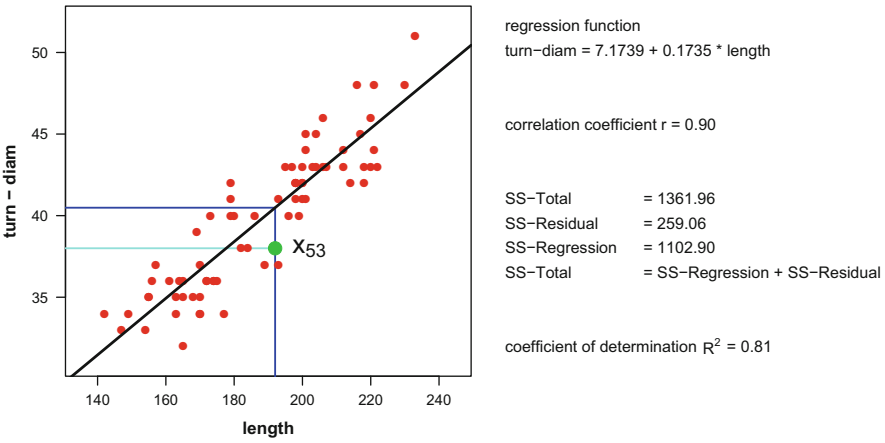


Fig. 11.12 Linear regression analysis of turning diameter (X8) and length (X7)

- The increase in the length of a car by one unit (i.e., by one inch in this case) leads to the increase in the turning diameter by $\widehat{b_1} = 0.1735$ feet.
- The sample correlation coefficient is 0.90—this implies a strong (positive) dependence of the turning diameter and the length.
- To estimate the fit of the estimated regression function, the coefficient of determination can be used. Its calculation is based on the decomposition of the variance of the dependent variable. For the calculation, the total sample variance (SS-Total), the unexplained (residual) variance (SS-Residual), and the explained variance (SS-Regression) are available. Using the formula

$$R^2 = \frac{SS - Regression}{SS - Total} = \frac{\sum (\widehat{y_i} - \bar{y})^2}{\sum (y_i - \bar{y})^2},$$

  we get that the coefficient of determination equals 0.81. The regression line characterizes (explains) the observed values quite well.

The observation $x(53)$ corresponds to the length of a car of 192 inches and a turning diameter 38 feet. The estimated regression function for a car of this length predicts the turning diameter to be equal to 40.49 feet.
Note: The interactive example allows you to display (graphically) the pairwise dependence of other variables as well (Fig. 11.13).

Interactive: Simple Linear Regression

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

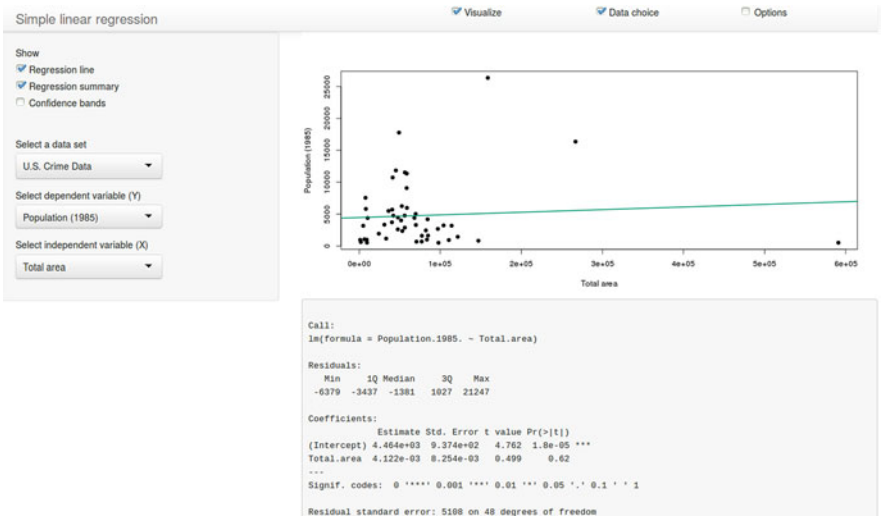


Fig. 11.13 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_regr

Please choose if you like

- the linear regression line to be included in the graphic
- a summary of regression results to be displayed below the graphic
- the confidence bands to be shown

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

Output

Using this interactive example, you can estimate a one-dimensional regression function for any two variables X, Y from three available data sets. The program generates a scatterplot, adds an estimated regression line and confidence bounds to the plot.

11.3 Multi-Dimensional Regression Analysis

Multi-Dimensional Regression Analysis

If a variable Y, which is to be modeled, depends on more than one variable X, we refer to the regression relationship as **multi-dimensional** or a **multiple linear regression**.

Let us write down a multi-dimensional linear regression function with m independent variables $X_1, X_2, \cdots, X_m (m < n)$:

$$E(y_i|X) = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_m x_{mi}$$

The estimation of unknown regression parameters can be done in the same way as in the case of one-dimensional linear regression—via the least squares method (LS).

More detailed discussion of multi-dimensional regression is omitted here, as it is one of the main topics of Econometrics.

Chapter 12
Time Series Analysis

12.1 Time Series Analysis

Definition

A time series is the vector of realizations of a random variable X over the time.

Graphical Representation

Scatterplots show the development of the realizations of the underlying random variable over time. The horizontal axis represents the time t (days, months, years) while the vertical axis shows the corresponding value x_t of X. In the following there are some examples from various fields of interest (Figs. 12.1, 12.2, and 12.3).

The Objectives of Time Series Analysis

The above examples illustrate how different the behavior of a time dependent random variable can be. The understanding of these different temporal attributes in any application is the aim of time series analysis. Descriptive time series models are chosen so that they explain the characteristics of the series. A time series could be interpreted as the realization of a stochastic process hence one tries to find a stochastic model that could have generated the observed data. An important issue is the identification of influence factors, which may be time series themselves. Stochastic time series modeling can help to understand such observed process. Also, assuming that the model remains valid in the future, it is possible to forecast

W.K. Härdle et al., *Introduction to Statistics*, DOI 10.1007/978-3-319-17704-5_12

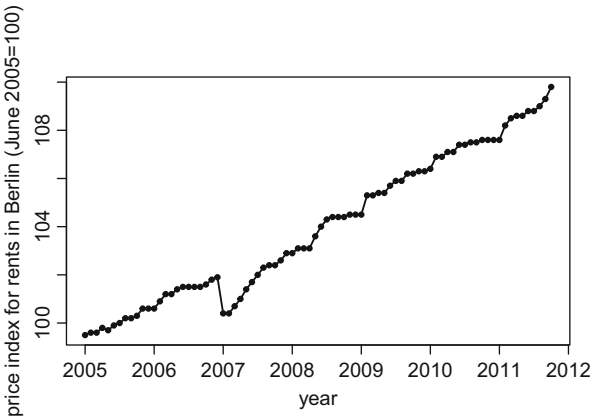


Fig. 12.1 Price index for rents in Berlin, 2005–2012

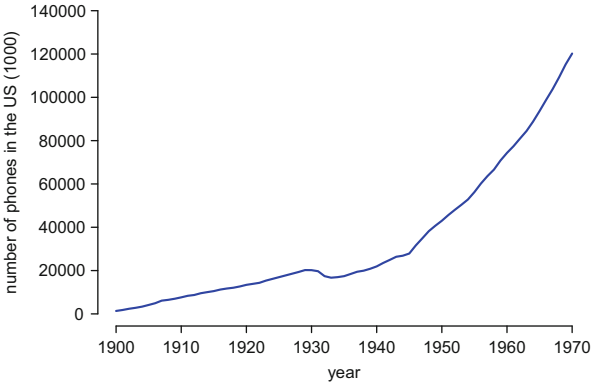


Fig. 12.2 Number of phones in the US (measured in $1,000s$), 1900–1970

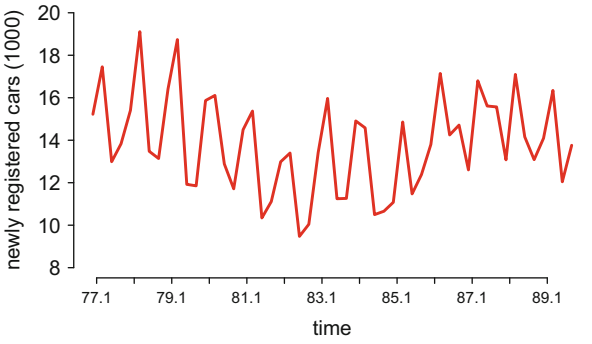


Fig. 12.3 Number of newly registered cars in Berlin, 1977:1–1989:4

future observations (predictive time series models). In the following we consider descriptive time series models only.

Components of Time Series

Time series are decomposed into its underlying driving components to show its characteristics:

- **Trend**
  General long-run trend of the series.
- **Periodic variation**
  Short-run influences, which overlap the long-run development corresponding to a rigid model. If the period is one year, periodic variations are called **seasonal variations**.
- **Iregular variation**

Trend and **periodic (seasonal) variation** are the systematic components.

12.2 Trend of Time Series

The analysis of time series starts with the extraction of the long-run behavior or trend from the observed values. There are a variety of different methods, leading to different trend lines for one and the same series. The choice of a particular method requires a comparison of advantages and disadvantages.

In this section we will present the moving average and least squares methods.

Method of Moving Average

In this method the estimated trend at every point in time is a weighted average of the original observed data:

$$T(t) = \sum_{i=-a}^{b} \lambda_i x_{t+i} \,,$$

with

$$\sum_{i=-a}^{b} \lambda_i = 1$$

The set of weights λ_i is called the filter.

The selection of the filter depends on periodic/seasonal variations and the desired smoothness. We will usually employ symmetric filters ($a = b$). They include future as well as past periods.

If the weights λ_i of a filter are equal for all i, the filter is called a simple moving average, if not, we call it a weighted moving average.

Support area The weighted average will be calculated in a window (area) of the original data. The choice of a and b determines the length of the window of data that are used for the support area. As a matter of principle the series of estimated trends can only be as long as the original series (equality, if $a = b = 0$). The longer the support area selected, the smaller the number of trend values that can be calculated and the smoother is the resulting trend series.

Frequently Used Filters for Time Series with Seasonal Variations

Symmetric filters ($a = b$) are often specified so that the $2a + 1$ weights are in square brackets. For the smoothing of seasonal time series the following filter can be applied. The reason is that they filter (smooth) out the periodic variations from original data for the trend calculation.

- six-month data

$$[1/4,\ 1/2,\ 1/4] \qquad (a = 1)$$
$$[1/8,\ 1/4,\ 1/4,\ 1/4,\ 1/8] \qquad (a = 2)$$

- quarterly data

$$[1/8,\ 1/4,\ 1/4,\ 1/4,\ 1/8] \qquad (a = 2)$$
$$[1/16,\ 1/8,\ 1/8,\ 1/8,\ 1/8,\ 1/8,\ 1/8,\ 1/8,\ 1/16] \qquad (a = 4)$$

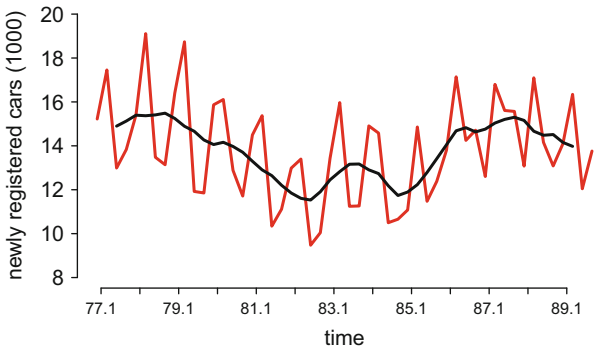


Fig. 12.4 Example for smoothing a time series

- monthly data

$$[1/24,\ 1/12,\ 1/12,\ 1/12,\ 1/12,\ 1/12,\ 1/12,\ 1/12,\ 1/12,\ 1/12,\ 1/12,\ 1/24] \qquad (a = 6)$$

Example (Quarterly Data)

- Number of newly registered cars in Berlin, 1977:1–1989:4 (Fig. 12.4)
- Filter: $[1/8,\ 1/4,\ 1/4,\ 1/4,\ 1/8]$
- red: original series
- black: smoothed series (trend)

Least-Squares Method

The Least-squares method is a second approach to estimate the trend component of a time series. The method was presented in the regression analysis chapter. We select a set of functions, which describe the trend as a function of time t and estimate the parameters of these functions. These parameter values minimize the sum of squared variations of the trend from the original data.

$$\sum_{t=1}^{T}(x_t - \hat{x}_t)^2 \rightarrow \text{ min.}$$

In the following we derive expressions for the least squares estimates of a simple linear trend and exponential trend functions.

Linear Trend Function

Suppose that the variable X depends linearly on time t

$$\hat{x}_t = a + b \cdot t$$

The sum of the squared residuals clearly depends on the parameters a and b as

$$S(a, b) = \sum_{t=1}^{T}(x_t - \hat{x}_t)^2 = \sum_{t=1}^{T}(x_t - a - b \cdot t)^2 \rightarrow \text{ min.}$$

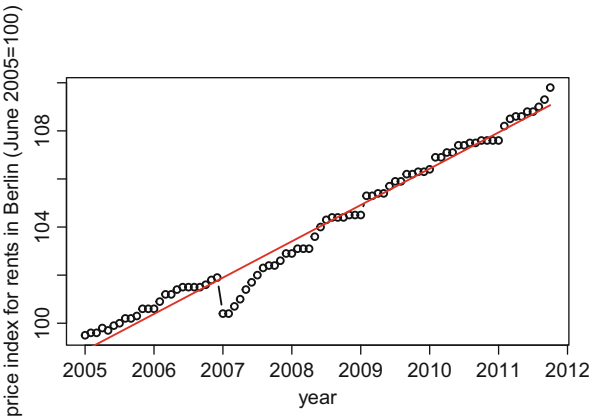


Fig. 12.5 Time series with a linear trend

Minimization results in the following estimators of the parameters (as introduced in the previous chapter on linear regression).

$$a = \frac{\sum_{t=1}^{T} x_t \sum_{t=1}^{T} t^2 - \sum_{t=1}^{T} t \sum_{t=1}^{T} x_t t}{T \sum_{t=1}^{T} t^2 - \left(\sum_{t=1}^{T} t\right)^2}$$

$$b = \frac{T \sum_{t=1}^{T} x_t t - \sum_{t=1}^{T} x_t \sum_{t=1}^{T} t}{T \sum_{t=1}^{T} t^2 - \left(\sum_{t=1}^{T} t\right)^2}$$

Example Price index for rents in Berlin, 2005–2012 (monthly data):

$$\hat{x}_t = 98.748 + 0.126 \cdot t, \qquad R^2 = 0.974,$$

where $t = 1$ corresponds to January 2005 (Fig. 12.5).

Exponential Trend

Suppose that the variable X exhibits an exponential dependence on time t of the form

$$\hat{x}_t = ab^t,$$

or, similarly, in logarithmic form

$$\log(\hat{x}_t) = \log(a) + t \log(b)$$

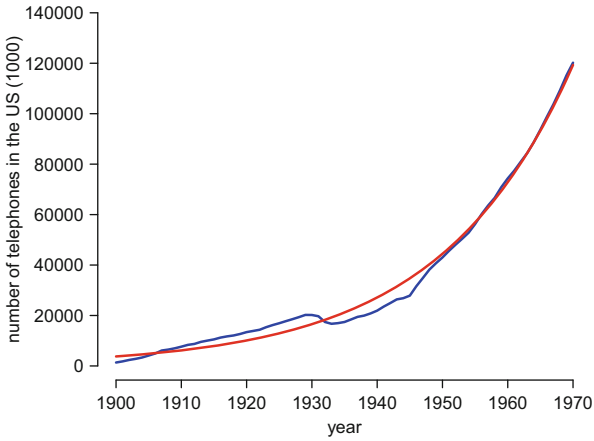


Fig. 12.6 Time series with an exponential trend

Least squares minimization results in the following estimators of the parameters.

$$\log a = \frac{\sum_{t=1}^{T} \log x_t \sum_{t=1}^{T} t^2 - \sum_{t=1}^{T} t \sum_{t=1}^{T} t \log x_t}{T \sum_{t=1}^{T} t^2 - \left(\sum_{t=1}^{T} t\right)^2}$$

$$\log b = \frac{T \sum_{t=1}^{T} t \log x_t - \sum_{t=1}^{T} \log x_t \sum_{t=1}^{T} t}{T \sum_{t=1}^{T} t^2 - \left(\sum_{t=1}^{T} t\right)^2}$$

Example Number of phones in the US (measured in $1,000s$), 1900–1970

$$\log \hat{x}_t = 3.553645 + 0.021448 \cdot t,$$

$$R^2 = 0.9923,$$

where $t = 0$ corresponds to the year 1899 (Fig. 12.6).

$$\hat{x}_t = 3,578.04 \cdot (1.051)^t$$

More Information: Simple Moving Average

Order of Moving Average

Domain: number ($k = a = b$) of past observations used for the calculation of the average.

- odd order $2k+1$:

$$X_t^\star = \frac{1}{2k+1} \sum_{i=t-k}^{t+k} X_i \qquad t = k+1, \ldots, T-k$$

Example

k	1	2
order	$2k+1=3$	$2k+1=5$
x_1	$x_1^\star$ n.a.	$x_1^\star$ n.a.
x_2	$x_2^\star = \frac{1}{3} \cdot \sum_{i=1}^{3} x_i$	$x_2^\star$ n.a.
x_3	$x_3^\star = \frac{1}{3} \cdot \sum_{i=2}^{4} x_i$	$x_3^\star = \frac{1}{5} \cdot \sum_{i=1}^{5} x_i$
x_4	$x_4^\star = \frac{1}{3} \cdot \sum_{i=3}^{5} x_i$	$x_4^\star = \frac{1}{5} \cdot \sum_{i=2}^{6} x_i$
$\vdots$	$\vdots$	$\vdots$
x_{T-2}	$x_{T-2}^\star = \frac{1}{3} \cdot \sum_{i=T-3}^{T-1} x_i$	$x_{T-2}^\star = \frac{1}{5} \cdot \sum_{i=T-4}^{T} x_i$
x_{T-1}	$x_{T-1}^\star = \frac{1}{3} \cdot \sum_{i=T-2}^{T} x_i$	$x_4^\star$ n.a.
x_T	$x_T^\star$ n.a.	$x_T^\star$ n.a.

Where **n.a.,** in the table, means that it is not feasible to estimate the trend for this particular point in time given our data and weighting structure.

- even order $2k$:

$$X_t^\star = \frac{1}{2k} \left[\frac{1}{2} X_{t-k} + \frac{1}{2} X_{t+k} + \sum_{i=t-(k-1)}^{t+(k-1)} X_i \right] \qquad t = k+1, \ldots, T-k$$

Example

k	1	2
order	$2k=2$	$2k=4$
x_1	$x_1^\star$ n.a.	$x_1^\star$ n.a.
x_2	$x_2^\star = \frac{1}{2}\left[\frac{1}{2}x_1 + \frac{1}{2}x_3 + x_2\right]$	$x_2^\star$ n.a.
x_3	$x_3^\star = \frac{1}{2}\left[\frac{1}{2}x_2 + \frac{1}{2}x_4 + x_3\right]$	$x_3^\star = \frac{1}{4}\left[\frac{1}{2}x_1 + \frac{1}{2}x_5 + \sum_{i=2}^{4} x_i\right]$
x_4	$x_4^\star = \frac{1}{2}\left[\frac{1}{2}x_3 + \frac{1}{2}x_5 + x_4\right]$	$x_3^\star = \frac{1}{4}\left[\frac{1}{2}x_2 + \frac{1}{2}x_6 + \sum_{i=3}^{5} x_i\right]$
$\vdots$	$\vdots$	$\vdots$
x_{T-2}	$x_{T-2}^\star = \frac{1}{2}\left[\frac{1}{2}x_{T-3} + \frac{1}{2}x_{T-1} + x_{T-2}\right]$	$x_{T-2}^\star = \frac{1}{4}[\frac{1}{2}x_{T-4} + \frac{1}{2}x_T + + \sum_{i=T-3}^{T-1} x_i]$
x_{T-1}	$x_{T-1}^\star = \frac{1}{2}\left[\frac{1}{2}x_{T-2} + \frac{1}{2}x_T + x_{T-1}\right]$	$x_{T-1}^\star$ n.a.
x_T	$x_T^\star$ n.a.	$x_T^\star$ n.a.

Explained: Calculation of Moving Averages

The following time series describes the development of the balance of payments (in Millions of DM) of Germany in the years 1977–1995.

The trend of these time series is estimated by the moving average method. Recall, this approach uses the formula

$$T(t) = \sum_{i=-a}^{b} \lambda_i x_{t+i}\,, \text{ with } \sum_{i=-a}^{b} \lambda_i = 1\,.$$

Since past and future values should have equal weights for the trend estimation in t, we choose $a = b$. For the smoothing of yearly data a simple moving average is applied, where all weights are identical. The weights must add to 1 over the entire supporting area, that means:

$$\lambda_i = \frac{1}{2a+1} \quad \forall i\,.$$

In Table 12.1 the moving average $T(t)$ was calculated for $a = 1$, $a = 2$ and $a = 3$.

Table 12.1 Calculation of moving averages

Year	t	Balance of payments	$T(t)$ $a = 1$	$T(t)$ $a = 2$	$T(t)$ $a = 3$
1977	1	9478			
1978	2	18003	5483.3		
1979	3	−11031	−7169.3	−4754.2	
1980	4	−28480	−17084.0	−4676.6	−476.0
1981	5	−11741	−10118.3	−6162.6	2161.4
1982	6	9866	2899.3	1631.6	6493.4
1983	7	10573	16126.3	16993.0	20325.4
1984	8	27940	28946.7	36499.8	36122.1
1985	9	48327	54020.0	50946.0	50418.9
1986	10	85793	72072.3	66498.6	63874.7
1987	11	82097	85408.7	81722.0	64551.3
1988	12	88336	91496.7	75118.4	56000.4
1989	13	104057	69234.0	51576.6	44779.3
1990	14	15309	29150.0	29113.0	29186.0
1991	15	−31916	−15609.3	6774.4	12573.9
1992	16	−30221	−28498.0	−20875.2	−4876.7
1993	17	−23357	−29256.3	−30700.6	
1994	18	−34191	−30455.3		
1995	19	−33818			

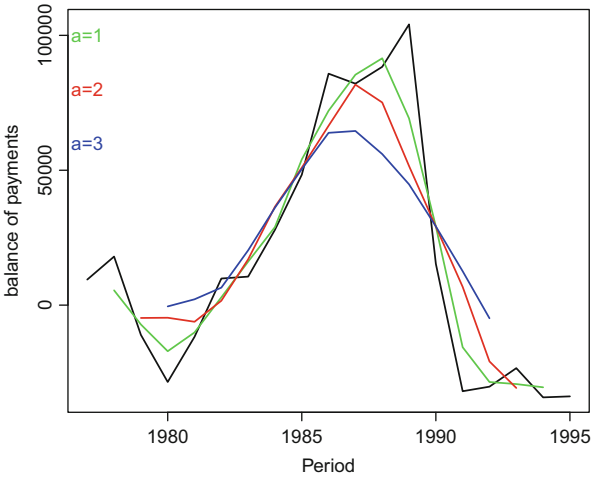


Fig. 12.7 Three alternative estimations of the long-run trend of the original series

If $a = 1$, one cannot estimate a trend for the period $t = 1$, because the value of the time series is unknown in $t = 0$. For $t = 2$ the estimated trend is $(9,478)/3 + (18,003)/3 + (-11,031)/3 = 5,483.3$.

In Fig. 12.7 the three alternative estimations of the long-run trend and the original series are compared.

One detects two important characteristics of the procedure:

- The larger the supporting area, over which the trend is estimated, the fewer the number of values of the trend that can be estimated.
- The estimated trend becomes smoother with increased supporting area (i.e., the larger is $b + a$).

Interactive: Test of Different Filters for Trend Calculation

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right (Fig. 12.8).

Please choose the type of trend to be calculated and included in the graphic.

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

Output:

Using this interactive example, you can select different filters and observe the effects of your selection on the estimated trend. The program generates a lineplot of the time series and adds an estimated regression trend.

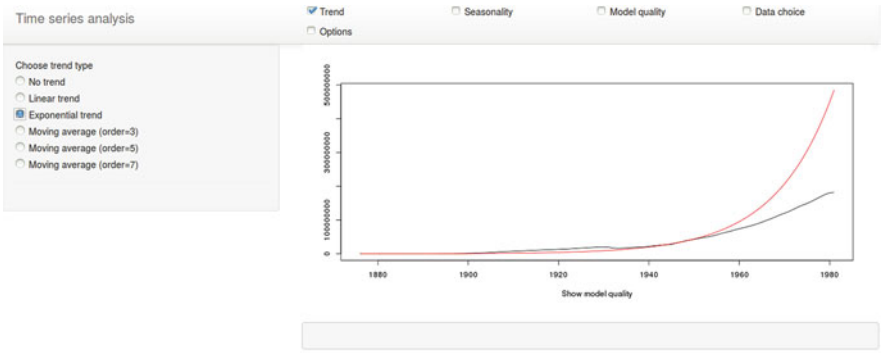


Fig. 12.8 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_time1

12.3 Periodic Fluctuations

So far from the original observed time series only the trend has been estimated. Information about seasonal attributes was dealt with, smoothed out, by the selection of a suitable filter. Now the season components are also to be estimated. For a better understanding, we introduce some useful definitions first.

- **Periods:** p_i**,** $i = 1, \ldots, P$ Number of repetitions of one season.

  Example Quarterly data over 10 years: $P = 10$

- **Time subintervals:** k_j**,** $j = 1, \ldots, k$ Number of observations in a seasonal cycle.

  Example Quarterly data: $k = 4$

- **Total number of observations:** $T = k \cdot P$
- **(Estimated) Trend values:** $\hat{x}_{i,j}$
- **Observed values:** $x_{i,j}$
- **(Estimated) Seasonal fluctuation components:** $s_{i,j}$

One must distinguish between additive and multiplicative time series models: An additive relationship between trend, seasonal component, and residuals is considered in the additive model while this relationship is multiplicative in multiplicative model. Accordingly the calculations of the estimated seasonal components differ.

- **Additive time series model**

$$s_{i,j} = x_{i,j} - \hat{x}_{i,j}\,, \qquad \bar{s}_j = \frac{1}{P}\sum_{i=1}^{P} s_{i,j}$$

$$\hat{x}_{i,j}^{\mathrm{ZRM}} = \hat{x}_{i,j} + \bar{s}_j \quad \text{for } i = 1, \ldots, P \quad j = 1, \ldots, k$$

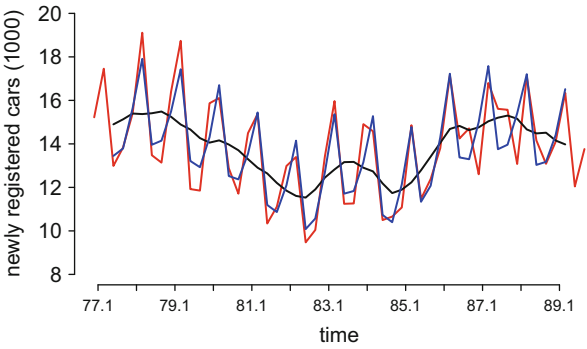


Fig. 12.9 Original time series (*red*); smoothed series (*black*); trend and seasonal component (*blue*)

Table 12.2 Seasonal components

j	Sum	$\bar{s}_j$	P
1	2.934	0.244	12
2	30.424	2.535	12
3	−17.434	−1.453	12
4	−16.120	−1.343	12

The forecasted value of the variable X from the time series model (ZRM) consists of the estimated trend value $\hat{x}_{i,j}$ *added* to the mean (estimated) seasonal coefficient $\bar{s}_j$.

- **Multiplicative time series model**

$$s_{i,j} = \frac{x_{i,j}}{\hat{x}_{i,j}}, \qquad \bar{s}_j = \frac{1}{P}\sum_{i=1}^{P} s_{i,j}$$

$$\hat{x}_{i,j}^{\text{ZRM}} = \hat{x}_{i,j} \cdot \bar{s}_j \quad \textit{for} \quad i = 1, \ldots, P \quad j = 1, \ldots, k$$

The forecasted value of the variable X from to the time series model (ZRM) consists of the estimated trend value $\hat{x}_{i,j}$ *multiplied* by the mean (estimated) seasonal coefficient $\bar{s}_j$.

Example

- Number of newly registered cars in Berlin—1977:1–1989:4 Additive time series model (Fig. 12.9, Table 12.2):
- Filter: [1/8, 1/4, 1/4, 1/4, 1/8]
- Red: Original time series
- Black: Smoothed series (estimated trend)
- Blue: Trend and seasonal component (estimated time series)

Explained: Decomposition of a Seasonal Series

This example shows how one decomposes an observed time series $x(t)$ into estimates of the trend $T(t)$, the seasonal component $S(t)$, and a residual vector $e(t)$. The model considered has the additive form $x(t) = T(t)+S(t)+e(t)$. For illustration we apply the method to data on newly registered cars in Berlin.

Trend

Two different procedures for estimation of the trend component were introduced above: The least squares and the moving average methods. Here the latter is used, where the trend is calculated according to

$$T(t) = \sum_{i=-a}^{b} \lambda_i x_{t+i}\,, \text{ with } \sum_{i=-a}^{b} \lambda_i = 1\,.$$

In order to remove all seasonal variation, one applies the filter $[1/8,\ 1/4,\ 1/4,\ 1/4,\ 1/8]$ to the observed quarterly data. It gives an even consideration of past and future data ($a = b = 2$) and the same weighting of all seasons.

Example

$$T(3) = 1/8 \cdot x(1) + 1/4 \cdot x(2) + 1/4 \cdot x(3) + 1/4 \cdot x(4) + 1/8 \cdot x(5)$$

Seasonal Variation

From the model $x(t) = T(t) + S(t) + e(t)$ it follows $x(t) - T(t) = S(t) + e(t)$. The left-hand side of this equation is an estimated detrended series. Assuming that the seasonal variation in the respective quarters has the same value (thus e.g.: $S(3) = S(7) = \ldots = S(51)$), an obvious procedure for the seasonal adjustment is the computation of the arithmetic means over all differences $x(t) - T(t)$, which belong to one season.

Example

$$\begin{aligned} S(3) = S(7) = \cdots = S(51) = \\ = [(x(3) - T(3)) + (x(7) - T(7)) + \cdots + (x(51) - T(51))]/12 \end{aligned}$$

For this procedure it is not important which method was used to estimate the trend.

Residuals

One calculates the estimated residuals via $e(t) = x(t) - T(t) - S(t)$.

Results of the Decomposition of Car Registration Time Series

You should check on the basis of the results for at least one period whether you can reconstruct the procedure described above or not (Tables 12.3 and 12.4).

The result of the decomposition is graphically illustrated. Note that the estimated trend series $T(t)$ (the green series) actually contains no more seasonal variation. This acknowledges the adequacy of selecting the filter $[1/8,\ 1/4,\ 1/4,\ 1/4,\ 1/8]$ for smoothing time series with quarterly data.

Table 12.3 Decomposition of car registration time series—part 1

Quarter	t	$x(t)$	$T(t)$	$S(t)$	$e(t)$	
1977.1	1	15222				
1977.2	2	17456				
1977.3	3	12988	14897.9	−1909.9	−1452.8	−457.1
1977.4	4	13833	15127.8	−1294.8	−1343.3	48.5
1978.1	5	15407	15395.9	11.1	244.5	−233.4
1978.2	6	19110	15370.5	3739.5	2535.4	1204.1
1978.3	7	13479	15408.8	−1929.8	−1452.8	−477.0
1978.4	8	13139	15487.3	−2348.3	−1343.3	−1005.0
1979.1	9	16407	15246.3	1160.7	244.5	916.2
1979.2	10	18738	14891.0	3847.0	2535.4	1311.6
1979.3	11	11923	14663.0	−2740.0	−1452.8	−1287.2
1979.4	12	11853	14267.1	−2414.1	−1343.3	−1070.8
1980.1	13	15869	14058.5	1810.5	244.5	1566.0
1980.2	14	16109	14160.9	1948.1	2535.4	−587.3
1980.3	15	12883	13971.5	−1088.5	−1452.8	364.3
1980.4	16	11712	13707.8	−1995.8	−1343.3	−652.5
1981.1	17	14495	13298.0	1197.0	244.5	952.5
1981.2	18	15373	12905.1	2467.9	2535.4	−67.5
1981.3	19	10341	12641.3	−2300.3	−1452.8	−847.5
1981.4	20	11111	12205.5	−1094.5	−1343.3	248.8
1982.1	21	12985	11850.1	1134.9	244.5	890.4
1982.2	22	13397	11608.3	1788.7	2535.4	−746.7
1982.3	23	9474	11530.5	−2056.5	−1452.8	−603.7
1982.4	24	10043	11907.6	−1864.6	−1343.3	−521.3
1983.1	25	13431	12450.5	980.5	244.5	736.0

Table 12.4 Decomposition of car registration time series—part 2

Quarter	t	$x(t)$	$T(t)$	$S(t)$	$e(t)$	
1983.2	26	15968	12824.3	3143.7	2535.4	608.3
1983.3	27	11246	13161.1	−1915.1	−1452.8	−462.3
1983.4	28	11261	13172.4	−1911.4	−1343.3	−568.1
1984.1	29	14908	12905.5	2002.5	244.5	1758.0
1984.2	30	14581	12736.5	1844.5	2535.4	−690.9
1984.3	31	10498	12182.3	−1684.3	−1452.8	−231.5
1984.4	32	10657	11738.1	−1081.1	−1343.3	262.2
1985.1	33	11078	11894.6	−816.6	244.5	−1061.1
1985.2	34	14858	12232.4	2625.6	2535.4	90.2
1985.3	35	11473	12788.6	−1315.6	−1452.8	137.2
1985.4	36	12384	13414.6	−1030.6	−1343.3	312.7
1986.1	37	13801	14047.3	−246.3	244.5	−490.8
1986.2	38	17143	14685.3	2457.7	2535.4	−77.7
1986.3	39	14249	14826.5	−577.5	−1452.8	875.3
1986.4	40	14712	14633.8	78.2	−1343.3	1421.5
1987.1	41	12603	14761.0	−2158.0	244.5	−2402.5
1987.2	42	16799	15038.3	1760.7	2535.4	−774.7
1987.3	43	15611	15204.5	406.5	−1452.8	1859.3
1987.4	44	15568	15301.1	266.9	−1343.3	1610.2
1988.1	45	13077	15157.0	−2080.0	244.5	−2324.5
1988.2	46	17098	14665.1	2432.9	2535.4	−102.5
1988.3	47	14159	14481.8	−322.8	−1452.8	1130.0
1988.4	48	13085	14514.5	−1429.5	−1343.3	−86.2
1989.1	49	14093	14155.9	−62.9	244.5	−307.4
1989.2	50	16344	13976.1	2367.9	2535.4	−167.5
1989.3	51	12044				
1989.4	52	13762				

Note: In Fig. 12.10 the black line represents our actual observed data series, green is the estimated trend component, the blue line is the estimated seasonal component, and the red line is the estimated residuals.

Interactive: Decomposition of Time Series

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please choose

- the trend type, e.g., exponential trend
- the seasonality type

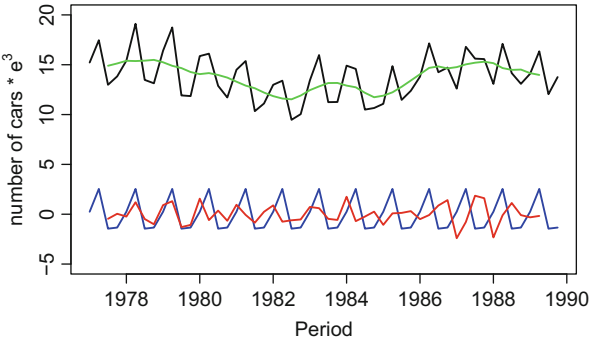


Fig. 12.10 Decomposition of car registration time series; trend (*green*), observed data (*black*), seasonal component (*blue*), residuals (*red*)

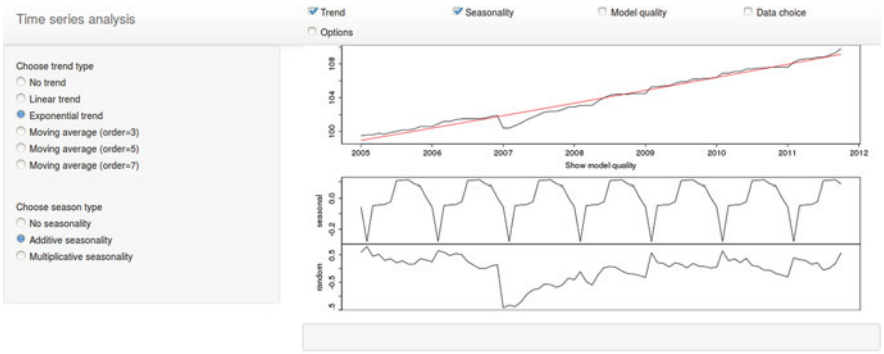


Fig. 12.11 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_time2

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

Output:

Using this interactive example, you can decompose a time series into its trend and seasonality components. The program generates a lineplot of the time series, adds an estimated regression trend, and displays seasonality (Fig. 12.11, Fig. 12.12).

12.4 Quality of the Time Series Model

In the preceding paragraph it likely became clear that, a priori, there is no best time series model. In particular, there are different methods for the estimation of the trend which do not differ in the parameters only, but follow different methodologies.

In order to select one model from the variety of possible models, one needs a criteria to justify a decision. How well a model describes (fits) the available data can be seen from the structure and the fluctuation of the residuals. The following

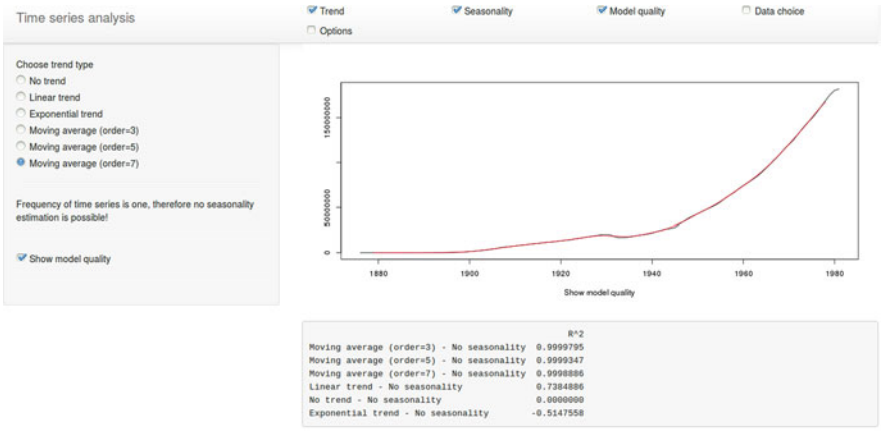


Fig. 12.12 Screenshot of the interactive example, available at http://u.hu-berlin.de/men_time3

measures, which offer information about the fluctuation of the residuals, have already been studied (Fig. 12.12).

Mean Squared Dispersion (Estimated Standard Deviation)

$$s_{\text{ZRM}} = \sqrt{\frac{1}{T}\sum_{i=1}^{P}\sum_{j=1}^{k}(x_{i,j} - \hat{x}_{i,j}^{\text{ZRM}})^2}$$

the coefficient of variation

$$v = \frac{S_{\text{ZRM}}}{\bar{x}}$$

coefficient of determination (applicable only if the trend was calculated with the least squares method.)

$$R^2 = 1 - \frac{s_{\text{ZRM}}^2}{s_x^2}$$

$$s_x^2 = \frac{1}{T}\sum_{i=1}^{P}\sum_{j=1}^{k}(x_{i,j} - \bar{x})^2 \qquad 0 \leq \frac{s_{\text{ZRM}}^2}{s_x^2} \leq 1$$

Interactive: Comparison of Time Series Models

The interactive example includes a number of sidebar panels. You can access the panels by setting a mark at the corresponding check box on the upper right.

Please choose

- the trend type, e.g., exponential trend
- the seasonality type
- if you like the model fit to be displayed

The last two panels allow you to choose a dataset or variable and to change the font size. For a detailed explanation on datasets and variables, please refer to Appendix A.

Output

Like in the preceding paragraph you can select a time series, which will be decomposed into estimated trend, seasonal component, and residuals. The program generates a lineplot of the time series, adds an estimated regression trend, and displays seasonality. Furthermore, in the result window you now also find measurements of the quality of the fit of your model to the data.

Appendix A
Data Sets in the Interactive Examples

A.1 ALLBUS Data

The German General Social Survey (ALLBUS) is a biennial survey that has been conducted since 1980 on the attitudes, behavior, and social structure of persons residing in Germany. A set of questions is asked in every ALLBUS on background information about respondents and their socio-economic context; detailed sets of questions on one or two topics per ALLBUS are replicated about every 10 years. More specific information can be found at the GESIS website: http://www.gesis.org/en/allbus.

The data used in the interactive examples has been taken from the English language version of ALLBUS-Cumulation 1980–2012. Note that not all questions were asked in biennial survey, therefore we have selected several years. Observations with missing values have been deleted from the data.

A.1.1 ALLBUS1992, ALLBUS2002, and ALLBUS2012: Economics

Variable	Type	Values and labels
Current economic situation in Germany	Ordered	1=very good,. . . ,5=very bad
Current economic situation in federal state[a]	Ordered	As before
Respondents own current financial situation	Ordered	As before
Economic situation in Germany in one year	Ordered	As before
Economic situation in federal state in one year[a]	Ordered	As before
Respondents own financial situation in one year	Ordered	As before

(continued)

W.K. Härdle et al., *Introduction to Statistics*, DOI 10.1007/978-3-319-17704-5

Satisfaction with performance of federal government	Ordered	1=very satisfied,. . . ,5=very dissatisfied[a]
Respondents monthly net income (categorized)	Ordered	0=no income, . . . , 22=more than 7500 EUR
Household net income (categorized)	Ordered	0=no income, . . . , 22=more than 7500 EUR

[a]Only for 1992

A.1.2 ALLBUS1994, ALLBUS2002, and ALLBUS2012: Trust

Variable	Type	Values and labels
Self placement on left right continuum	Ordered	1=extreme left, . . . , 10=extreme right
Trust in health service	Ordered	1=no trust at all, . . . , 7=great deal of trust
Trust in federal constitutional court	Ordered	As before
Trust in federal parliament (Bundestag)	Ordered	As before
Trust in municipal administration[a]	Ordered	As before
Trust in army[a]	Ordered	As before
Trust in catholic church	Ordered	As before
Trust in protestant church	Ordered	As before
Trust in judicial system	Ordered	As before
Trust in television	Ordered	As before
Trust in newspaper	Ordered	As before
Trust in universities/higher education	Ordered	As before
Trust in federal government	Ordered	As before
Trust in trade unions[a]	Ordered	As before
Trust in police	Ordered	As before
Trust in job centers[a]	Ordered	As before
Trust in state pension system[a]	Ordered	As before
Trust in employer association[a]	Ordered	As before
Trust in political parties[b]	Ordered	As before

[a]Only for 1994
[b]Not for 1994

A.1.3 ALLBUS2002, ALLBUS2004, and ALLBUS2012: General

Variable	Type	Values and labels
East West	Binary	1=Old federal states, 1=New federal states
Interview type	Binary	1=Paper-and-Pencil Interview, 2=Computer-Assisted Personal Interview
Sex	Binary	1=male, 2=female
Member in trade union	Binary	1=yes, 2=no
Support political party[a]	Binary	1=yes, 2=no
Eligible for voting in last federal elections[c]	Binary	1=yes, 2=no
Age	Numeric	in years
Body mass index[b]	Numeric	
Height [b]	Numeric	in cm
Weight [b]	Numeric	in kg
Respondents monthly net income	Numeric	in EUR
Household net income	Numeric	in EUR

[a]Only for 2002
[b]Not for 2002
[c]Not for 2012

Source: Wikimedia Commons, the free media repository[1]
New federal states: Berlin (east), Brandenburg, Mecklenburg-Vorpommern, Sachsen, Sachsen-Anhalt, and Thüringen

[1] http://commons.wikimedia.org/wiki/File:Deutschland_politisch_2010.png.

A.2 Boston Housing Data

Housing data for 506 census tracts of Boston from the 1970 census and were first published by D. Harrison and D.L. Rubinfeld (1978) in the article *Hedonic prices and the demand for clean air* in the Journal of Environmental Economics and Management, no. 5, pages 81–102.

Variable	Type	Values and labels
Per capita crime rate by town	Numeric	
Proportion of residential land zoned for lots over 25,000 sq.ft	Numeric	
Proportion of non-retail business acres per town	Numeric	
Charles River dummy variable	Binary	1=tract bounds river, 0=otherwise
Nitric oxides concentration	Numeric	in parts per 10 million
Average number of rooms per dwelling	Numeric	
Proportion of owner-occupied units built prior to 1940	Numeric	
Weighted distances to five Boston employment centers	Numeric	
Index of accessibility to radial highways	Ordered	
Full-value property-tax rate per USD 10,000	Numeric	
Pupil-teacher ratio by town	Numeric	
Transformed proportion of blacks B by town	Numeric	$1000(B - 0.63)^2$
Percentage of lower status of the population	Numeric	
Median value of owner-occupied homes	Numeric	in 1000 US$

A.3 Car Data

The car dataset was taken from the book *Graphical Methods for Data Analysis* by J.M. Chambers, W.S. Cleveland, B. Kleiner, and P.A. Tukey (1983). It consists of 13 variables measured for 74 car types.

Variable	Type	Values and labels
Car model	Factor	
Price	Numeric	in US$
Mileage	Numeric	miles per gallon
Repair record 1977	Ordered	1=worst,. . . , 5=best
Repair record 1978	Ordered	1=worst,. . . , 5=best
Headroom	Numeric	in inches
Rear seat clearance	Numeric	in inches
Trunk space	Numeric	in cubic feet
Weight	Numeric	in pound
Length	Numeric	in inches
Turning diameter	Numeric	in feet
Displacement	Numeric	in cubic inches
Gear ratio for high gear	Numeric	

A.4 Credit Data

The data are taken from Datasets at the Department of Statistics, University of Munich, and the SFB386[2] are used in the books by Fahrmeir et al. The dataset consists of 1000 consumer credits from a German bank. We modified the data to achieve more binary variables.

Variable	Type	Values and labels
Creditability	Binary	1=not credit-worthy, 2=credit-worthy
Running account	Binary	1=no, 2=yes
Duration	Numeric	in month
Payment of previous credits	Binary	1=problems, 2=no problems
Purpose of credit	Factor	1=other, 2=new car, 3=used car, 4=furniture, 5=radio/television, 6=household appliances, 7=repair, 8=vacation, 9=repair, 10=business
Amount	Numeric	in Deutsche Mark
Savings or stocks	Binary	1=yes, 2=no
Has been employed by current employer for more than one year	Binary	0=yes, 1=no
Marital Status / Sex	Binary	1=other, 2=male: divorced / living apart
Guarantor	Binary	1=yes, 2=no

(continued)

[2] http://www.statistik.lmu.de/service/datenarchiv/kredit/kredit_e.html.

Living in current household for	Factor	1=less than 1 year, 2=between 1 and 4 years, 3=between 4 and 7 years, 4=7 or more years
Most valuable available assets	Binary	1=house/land, 2=other/none
Age	Numeric	in years
Other running credits	Binary	1=no, 2=yes
Type of apartment	Binary	1=other, 2=rented
Number of previous credits at this bank	Binary	1=1 or more, 2=none
Occupation	Binary	1=other, 2=unskilled or unemployed
Number of persons entitled to maintenance	Binary	1=3 and more persons, 2=between 0 and 2 persons
Telephone	Binary	1=no, 2=yes
Foreign worker	Binary	1=yes, 2=no

A.5 Decathlon Data

Data are from 33 decathlon participants in the Olympic games 1988. The data are taken from Hand, Daly, Lunn McConway, and Ostrowski (1994) *A handbook of small data sets*, Chapman & Hall, London.

Variable	Type	Values and labels
100 m run	Numeric	in seconds
Long jump	Numeric	in meter
Shot	Numeric	in meter
High jump	Numeric	in meter
400 m run	Numeric	in seconds
110 m hurdles	Numeric	in seconds
Discus throw	Numeric	in meter
Pole vault	Numeric	in meter
Javelin throw	Numeric	in meter
1500 m run	Numeric	in seconds
Score	Numeric	in points

A.6 Hair and Eye Color of Statistics Students

Distribution of hair and eye color and sex in 592 statistics students. The data are taken from the software R (R Core Team, 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/).

Variable	Type	Values and labels
Hair color	Factor	1=black, 2=brown, 3=red, 4=blond
Eye color	Factor	1=brown, 2=blue, 3=hazel, 4=green
Sex	Binary	1=male, 2=female

A.7 Index of Basic Rent

The dataset contains the monthly index for the basic rent for apartments in Berlin from January 2005 till October 2011. The basic rent is measured in EUR/m^2 and the index contains the basic rent compared to the base year 2005 (Laspeyres price index). The data are published by the Statistical Office for Berlin-Brandenburg (Amt für Statistik Berlin-Brandenburg—Verbraucherpreisindex im Land Berlin[3]).

Variable	Type	Values and labels
Month, Year	Numeric	
Index	Numeric	

[3] https://www.statistik-berlin-brandenburg.de/Statistiken/statistik_SB.asp?Ptyp=700&Sageb=61001&creg=BBB&anzwer=4.

A.8 Normally Distributed Data

The dataset consist of 1000 observations simulated from 12 different normal distributions.

Variable	Type	Values and labels
NORM_0_1	Numeric	Simulated from $N(0; 1)$
NORM_0_2	Numeric	Simulated from $N(0; 2)$
NORM_0_5	Numeric	Simulated from $N(0; 5)$
NORM_1_1	Numeric	Simulated from $N(1; 1)$
NORM_1_2	Numeric	Simulated from $N(1; 2)$
NORM_1_5	Numeric	Simulated from $N(1; 5)$
NORM_5_1	Numeric	Simulated from $N(5; 1)$
NORM_5_2	Numeric	Simulated from $N(5; 2)$
NORM_5_5	Numeric	Simulated from $N(5; 5)$
NORM_10_1	Numeric	Simulated from $N(10; 1)$
NORM_10_2	Numeric	Simulated from $N(10; 2)$
NORM_10_5	Numeric	Simulated from $N(10; 5)$

A.9 Telephone Data

This time series contains yearly data of telephones in the US from 1871 to 1981 and has been taken from Douglas Galbi's personal website.[4] He compiled the time series from several sources, mainly

1876–1944: Federal Communications Commission, Statistics of the Communications Industry, 1944, Table 6, and

1945–1981: Federal Communications Commission, Statistics of Communications Common Carriers, 1982, Table 5.

Variable	Type	Values and labels
Year	Numeric	
Telephones	Numeric	

[4] http://galbithink.org/.

A.10 Titanic Data

This dataset provides information on the fate of passengers on the fatal maiden voyage of the ocean liner *Titanic* and taken from the software R (R Core Team, 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/).

Variable	Type	Values and labels
Class	Factor	1=1st class, 2=2nd class, 3=3rd class, 4=crew
Sex	Binary	1=male, 2=female
Age	Binary	1=child, 2=adult
Survived	Binary	1=no, 2=yes

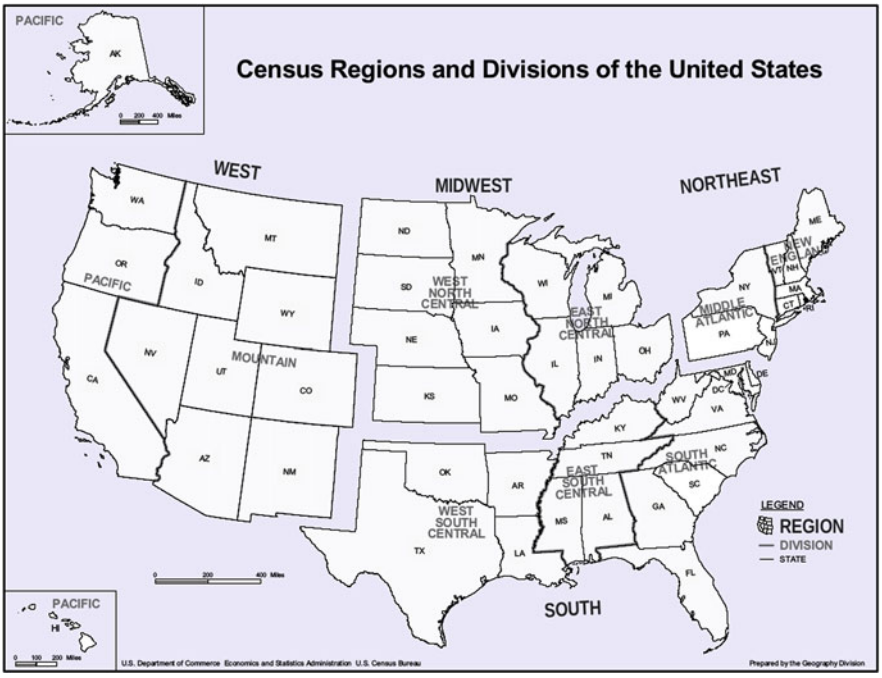
A.11 US Crime Data

The data contains crime rates per 100,000 population for some violent and property crimes for each state in the US. The data are taken from the Uniform Crime Reporting Statistics website (www.ucrdatatool.gov) and *Statistical abstract of the United States* 1987, table 25: resident population by states,[5] table 263: crime rates by state and by type[6], and table 316: area of states and other areas.[7]

[5] http://www2.census.gov/prod2/statcomp/documents/1987-02.pdf.

[6] http://www2.census.gov/prod2/statcomp/documents/1987-03.pdf.

[7] http://www2.census.gov/prod2/statcomp/documents/1987-03.pdf.



Variable	Type	Values and labels
State	Factor	Abbreviation
Total area	Numeric	in square miles
Population	Numeric	in thousands
Murder rate	Numeric	
Robbery rate	Numeric	
Aggravated assault rate	Numeric	
Burglary rate	Numeric	
Larceny-theft rate	Numeric	
Motor vehicle theft rate	Numeric	
Census region	Factor	1=Midwest, 2=Northeast, 3=South, 4=West
Census divisions	Factor	1=East North Central, 2=East South Central, 3=Middle Atlantic, 4=Mountain, 5=New England, 6=Pacific, 7=South Atlantic, 8=West North Central, 9=West South Central
State name	Factor	

Glossary

Absolute frequency - number of occurrences of certain value or combination of certain values of the investigated variable. 17, 21, 209, 396, 420, 422

Absolute scale - containing natural unit of measurement and natural zero point. See also metric scale. 12, 13

Alternative hypothesis - the hypothesis opposing to the null hypothesis (hypothesis testing). 313, 315, 321, 323, 324, 331, 332, 338–341, 343, 350, 355, 356, 366–368, 371, 372, 375, 381, 385–387, 390, 401, 408

Approximation - Under some assumptions, we are allowed to substitute some well-known simple distribution (typically Normal) for the true and complicated one. 235, 236

Arithmetic average - This value is obtained by spreading the sum of all observed realizations uniformly across all statistical elements. The arithmetic average makes sense only for metrically scaled variables. 46, 47, 52, 54, 56, 60, 65, 66, 120, 438, 462, 489

Asymptotic unbiasedness - property of an estimator. With increasing number of observations, the expected value of the estimator converges towards the true value of the estimated parameter. 256

Bar graph - is a graphical representation with rectangular bars whose lengths represent the values that they represent. 22

Binary variable - (also **dichotomous variable**) Random variable whose result is always one of two distinct values, most often “0",“1" or “true",“false". 12

Binomial distribution - distribution of a discrete random variable: “number of occurrences of an event in n repetitions of the experiment if the probability of occurrence of the event in one trial is p. The Binomial distribution has parameters n and p. 162, 199, 234, 264, 288, 362, 396

Boxplot - graphical display of selected summary statistics containing information about the frequency distribution of a metrically distributed random variable. It provides an idea about the shape of the distribution and about the structure of observed data. 64

W.K. Härdle et al., *Introduction to Statistics*, DOI 10.1007/978-3-319-17704-5

Bravais-Pearson correlation coefficient - It measures the strength and the direction of the linear relationship between two metrically distributed random variables. It is the ratio of covariance (common variability) and the product of standard deviations (variability of the variables). Its value lies between -1 and 1. 439, 445, 462, 473

Census - sampling and investigation of all elements of the sample space. 8, 210

Central Limit Theorem - The theorem concerns the approximation of the sum of random variables by Normal distribution for a sufficiently large number of summands. 220, 235, 289, 298, 332, 357, 377, 380, 384

Chebyshev inequality - bounds the probability that a random variable falls outside an interval around its expected value. 221

Chi-square "Goodness-of-Fit" test - statistical test. The null hypothesis states that the true distribution function of the observed data is equal to a given distribution function. The test statistic has Chi-square distribution. xvii, 397, 400

Chi-square distribution - a distribution of a sum of n independent, identically distributed random variables with standard Normal distributions. The parameter n is called the degrees of freedom. 204, 293

Chi-square test of independence - statistical test. The null hypothesis says that two random variables are independent. The test statistic has Chi-square distribution. 404, 405

Class boundaries - the boundary of a class of a metrically scaled variable is a value which bounds a given class from above (upper bound) or from bellow (lower bound). The difference between upper and lower bound is called the **width** of the group. 15, 37

Class midpoint - the value representing the group which is obtained as arithmetic average of its upper and lower boundary. 15

Class width - see boundaries of a class. 15, 37

Coefficient of determination - the coefficient of determination measures the quality and suitability of the chosen regression function for given data. It is defined as a ratio of the variability explained by the regression function and the total variability of the regressors, i.e., it can be interpreted as a proportion of variability explained by the regression model. Its values lie between 0 and 1, higher values mean that the model explains the data better. In linear regression, the coefficient of determination is equal to the square of the correlation coefficient. 463, 470, 472, 473, 493

Combination - choice of k elements out of total n elements if the order is not important is called combination of k-th class out of n elements. We distinguish combinations with and without replacement. See also combinatorics. 97, 102

Combinatorics - investigates various ways of sorting and/or grouping of certain elements. It is very important for the probability theory. See also permutation, variation, combination. 97

Complementary event - see event. 122, 154

Components of a time series - We distinguish the systematic components (trend, periodic fluctuations) and irregular random residual fluctuations. 479

Conditional distribution - in the framework of two-dimensional frequency distribution, it is the distribution of a variable X (resp. Y) for a fixed value (outcome) of variable Y (resp. X). 430, 432–434
Conditional probability - probability of an occurrence of a certain event under the condition that some other event also occurs. 82, 87, 89
Confidence interval - random interval, result of an interval estimate of some unknown parameter. 273, 275, 277, 279, 288, 305
Confidence level - probability that the confidence interval calculated from our data covers the true unknown value of the estimated parameter. 273, 297, 305
Consistency - a property of an estimator of some unknown parameter. With increasing number of observations, the expected value of the consistent estimator converges towards the true value of the unknown parameter and its variance converges to zero. 257, 289, 298
Contingency coefficient - It measures the intensity of a relation between two nominal random variables. It is calculated using quadratic contingency and its value lies between 0 and 1, where 0 means statistical independence. The contingency coefficient is practically never equal to 1 (complete dependency). Therefore, the adjusted contingency coefficient was introduced. 451
Contingency table - two-dimensional contingency table (or cross-table) is used to display the joint frequency distribution of two nominal or ordinal random variables. 125, 419, 421, 422, 432–434, 450, 452
Continuous variable - metrically scaled random variable which could return any of infinitely many values in any arbitrarily small interval. 14, 27, 40, 120, 125, 132, 150, 176, 181, 200, 264, 438
Covariance - a measure of joint variability of a pair of metrically scaled variables. It measure both the strength and the direction of the dependency. The correlation coefficient can be used to compare different covariances. 435, 437, 438, 462, 470
Critical region - values of the test statistic that lead to rejection of the null hypothesis. 315
Critical value - value(s) of the test statistics separating the critical and acceptance regions of the null hypothesis. It depends on the probability distribution of the test statistic and on the chosen level of significance. 315–317, 327, 328, 333, 334, 345, 349, 354, 356, 357, 362, 366, 375, 382, 385–387, 412, 415
Cross-section data - data collected at the same point in time or for the same period of time on different elements. 17
Cumulated frequency - frequency of observations smaller or equal to a given value or, for grouped variables, the upper bound of the class in which this value lies. It is defined for at least ordinal variables. We can have absolute or relative cumulated frequency. 34
Decile - any of the nine values that devide the sorted data into ten equal parts. 42
Density - for a continous random variable the density function describes the relative likelihood of taking on values withiin a given interval. 119
Descriptive statistics - statistical methods oriented towards the collection of data and its basic description. The results concern only the investigated set of data. 1, 3

Dichotomous variable - see binary variable. 12

Discrete variable - we say that metrically scaled random variable is discrete if the set of its possible values is finite or if it contains countably many elements. 14, 21, 40, 120, 132, 149, 154, 170, 200, 264

Disjoint events - see intersection. 73

Distribution function - The distribution function $F(x)$ of a random variable X is equal to the probability that the random variable is smaller or equal to x. 119, 149, 154, 170, 176, 182, 197, 311, 313, 333, 335, 362, 364, 366, 368, 371, 372, 391, 393, 407

Dotplot - two-dimensional graphical display of one-dimensional data. On the horizontal axis, you find the observed value. The value on the vertical axis is arbitrary (usually randomly chosen). 29

Efficiency - is a property of unbiased estimators. An estimator is called efficient if its variance is smaller than the variance of any other unbiased estimator of the same parameter. 256

Equivalence - Equivalence of events means their equality. It means that whenever event A happens, event B happens too and the other way around. In this case is A a subset of B and B is a subset of A. See also implication. 70

Error of type I - rejection of null hypothesis if it is true. 320, 323, 343, 360, 365, 372

Error of type II - acceptance of null hypothesis if it is false. 320, 321, 323, 324, 338, 340, 343, 346, 353, 360, 368, 370, 374, 388, 394, 404, 408

Estimate - realization of the estimator. 330

Estimator - function of the sample variables which is suitable for estimating some unknown parameter of the investigated distribution. 253, 293, 332, 348, 357, 371, 378, 406

Event - An event is any possible outcome of a random experiment. An **elementary** event is an event which cannot be split to some partial events; elementary events are disjoint. A **complementary** event is a set of all elementary events of the sample space S which are not contained in the investigated event. Events are subsets of the sample space and therefore we can use here common set relations and operations (see also implication, equivalence, union, intersection, logical difference). 69, 70, 75, 79, 82, 87, 107, 131, 154, 163

Expected value - the value of the random variable which we expect to obtain before the random experiment is carried out. It corresponds to the arithmetic average of the frequency distribution. 139, 150, 155, 163, 170, 177, 182, 196, 199, 206, 218, 229, 231, 233, 234, 253, 266, 275, 322, 398

Exponential distribution - a distribution of a continuous random variable. It has the parameter λ and it represents the probability distribution of the distance of two subsequent events in a Poisson process. 176

F-distribution - a distribution of a continuous random variable which is a ratio of two independent random variables with Chi-square distribution with f_1 and f_2 degrees of freedom. The distribution has two parameters, the above mentioned degrees of freedom f_1 and f_2. 207

Filter - a set of weights which are used to calculated moving averages for a given time series. The choice of the filter depends on the type of seasonal fluctuations and on the desired level of smoothing. Symmetric filters are often used. 479, 489

Frequency distribution - sorted results of an experiment together with their absolute frequencies are called the frequency distribution of the investigated variable. Depending on the number of variables, we distinguish one- and more-dimensional frequency distributions. The **Frequency table** provides systematic and accessible information about the data. 18, 21, 49, 63, 68, 120, 421, 423, 429, 444, 452

Frequency table - see frequency distribution. 21, 30, 37

Geometric average - It can be used to calculate the mean for (at least) ratio scaled random variables with positive values, which are multiplicatively interrelated. The logarithm of the geometric average is equal to the arithmetic average of the logarithms of the observed values. 52, 54

Grouping - joining of equal or similar observations of some variable into one group or class. See also class boundaries. 14, 36, 40

Harmonic average - a special type of arithmetic average for ratio scaled variables. It is used whenever we calculate an average from ratios and we have an additional information g_j which is related to the numerator of the ratio x_j. 50, 52

Histogram - graphical display of the frequencies of grouped continuous by the area of rectangles whose height corresponds to the relative frequency of the groups. The histograms are useful also for displaying the frequencies of discrete variables. 27, 31, 50

Hypergeometric distribution - discrete distribution with parameters M, N, and n. It describes the probability of occurrence of an event in n repetitions of random experiment under assumptions of independence and constant probability of success in a single trial. 163, 200, 236

Identification variables - characteristic which clearly defines the sample space and which identifies statistical elements (so that we know if they belong to the sample space under investigation). Its value is the same for all statistical elements in the sample space and it doesn't change during the investigation. 10

Inductive statistics - 1.) Statistical methods allowing to draw conclusions concerning parameters of some population based on a random sample from this population. 2.) According to the Theory of Probability, these are the methods which allow to make, with given accuracy, statements on the population based on the information from random samples. 2

Interpolation - method of calculating unknown function value from known "close" values of that function. 36

Interquartile range - It is the difference between the upper and the lower quartile. It is the width (size) of a region in which lies 50% of the central observed values. 57, 60, 65

Intersection (of events) - a set of all elementary events which belong to all events under consideration (i.e., the events involved in the intersection). Two events with an empty intersection are called **disjoint events**. 71

Interval estimate - The unknown parameter is estimated by an interval which covers the true value of the parameter with prescribed probability. 273, 275

Interval scale - We can measure and interpret differences between the values of random variables which are measured on the interval scale. Such variables do not have any natural zero and any natural unit of measurement (see also scale). 12, 13

Kendall correlation coefficient - The Kendall rank correlation coefficient is based on the comparison of the order of all possible pairs of the observed values. The pairs of observations with the same (or opposite) order are called **concordant** (resp. **discordant**). Apart of this, some pairs can have equal values. Kendall correlation coefficient is the ratio of the difference between the number of concordant and discordant pairs and the sum of concordant and discordant pairs. 447, 448, 450

Least squares - 1.) method for calculating estimators of the regression coefficients in linear regression. The estimators are defined as the numbers minimizing the sum of squared residuals (RSS - Residual Sum of Squares) of the fitted values from the observed values. 2.) Principle for the construction of estimators of an unknown parameter based on the minimization of the sum of squared differences between sample values and some function of the parameter. 266, 270, 459, 475, 481, 489, 493

Level of significance - probability that the test statistic falls into the critical region if the null hypothesis is true. 353, 357–360, 362, 364, 365, 367, 370, 376, 377, 382, 387–389, 393, 394, 397, 401, 406, 411, 413, 414, 417

Likelihood function - function which assigns, with respect to the observations, values (probability or density) to all possible values of the estimated parameter. 265

Logical difference - a logical difference of two events A and B is an event when we observe A and do not observe B. 73

Marginal distribution - for two-dimensional frequency distribution, the marginal distribution is the one-dimensional distribution of the variable X (or Y) which does not contain any information about the distribution of the other random variables Y (or X). 421, 429, 432–434, 436

Maximum likelihood - general principle for the construction of estimators of unknown parameters. The estimator is the value which maximizes the probability (or density) of the realized sample. 264

Mean absolute deviation - arithmetic average of the absolute deviations of the observations from a fixed point which is usually chosen as some mean value (most often median or arithmetic average). 57, 60

Mean squared error (MSE) - 1.) Arithmetic average of the squared deviations of the observed values from certain mean values. The MSE from the arithmetic average of the observations is called the variance. 2.) Expected value of the squared deviation of the estimator and the true value of the estimated parameter. 58, 212, 255

Median - the value which splits the sorted realizations of (at least ordinal) random variable into two equal parts. It is robust with respect to outliers and it corresponds to the second quartile. 42, 47, 60, 65, 66, 253

Mode - It is the most often observed realization of the variable. It can be determined for any scale. For the nominal variables it represents the only reasonable mean value. The mode is not sensitive with respect to outlying observations. 40

Nominal scale - We say that the scale is nominal if only the equivalence of the results can be determined, i.e., the various results of the experiment cannot be sorted (see also scale). 11, 12, 450

Normal distribution - a bell-shaped distribution of a continuous random variable with parameters μ and σ. The parameters μ determined the expected value and the parameter σ the standard deviation of the normally distributed random variable. 181, 196, 199, 219, 235, 264, 290, 306, 374

Null hypothesis - statistical formulation of some statement concerning the sample space which can be tested (and rejected) by a statistical test. 313–316, 319–321, 323, 324, 327–329, 335, 338, 340, 341, 343, 346, 350, 353, 356, 358, 365, 367, 371, 372, 374, 381, 386, 390, 392–394, 396–398, 401, 404–406, 408, 412, 415, 417, 418

Observation - the actual values assumed by statistical variables. 10, 419

Ordinal scale - the scale is ordinal if the outcomes of the experiment could be represented by natural numbers, we can determine equivalence of two elements and the results can be naturally sorted. Attention: using ordinal scale, you cannot interpret the size of differences between the classes (see also scale). 12

Permutation - each sorting of all n elements contained in some set is called a permutation. We distinguish permutations with repeating, permutations without repeating and permutations involving more groups of identical elements (see also combinatorics). 97, 98

Pictograph - graphical representation in which the size of some object or the number of depicted objects represent a numerical value. 23

Pie chart - is a circular chart devided into sectors which represent numerical proportions.. 23

Poisson distribution - distribution of a discrete random variable describing number of occurrences of an event; the event occurs repeatedly, but randomly and independently in a fixed time period. The Poisson distribution has parameter λ. 170, 200, 255, 402

Population - set of all statistical elements relevant for the statistical investigation of at least one chosen characteristic. 8, 218, 251

Power function - function which gives the dependency of the probability of rejecting the null hypothesis on the true value of the tested parameter. 338, 339, 341, 345, 346, 351, 355, 364, 367

Probability - measure P which quantifies certainty and uncertainty of events in the random experiment. 75, 80, 81, 86, 90, 98, 107, 109, 124, 131, 149, 154, 163, 184

Probability density function - function giving the probability that random variable X equals to the value x_j. 149, 163, 170

Probability distribution - It is obtained by assigning probabilities to the sorted values of random variable (discrete probability distribution). 119, 124, 149

Probability theory - theory concentrated on quantitative models of experiments with random outcomes (random experiments). 69, 98

Quantile - quantile x_p is the value which splits the upwards sorted realizations of the (at least ordinal) variable in the ratio $p : (1 - p)$, where p lies between 0 and 1. Special cases are quartiles, quintiles, and deciles. 41

Quartile - special case of the quantile for $p = 0.25$, $p = 0.5$, and $p = 0.75$. The sorted observations are split by the quartiles into four parts of equal size. The quartile $x_{0,25}$ is the lower quartile, $x_{0,75}$ is the upper quartile, and $x_{0,5}$ is the median. 42, 65

Quintile - special case of the quantile for $p = 0.2$, 0.4, 0.6, and 0.8. The sorted observations are split by the quintiles into five equally large parts. 42

Random experiment - This is a real or constructed experiment which can be repeated arbitrarily many times under the same conditions and whose result cannot be determined in advance. 69, 109, 120, 154

Random sampling - method of choosing elements of the sample space. Each element has nonzero probability of being selected. The probabilities do not have to be equal. 251, 354

Random variable - random variable is the (real) number which is assigned to every elementary event. 107, 113, 119, 124, 131, 139, 163, 170, 196, 199, 204, 207

Range - parameter of scale, it is the difference of the highest and smallest observation (for classified data it is the difference between the highest and smallest bound of the groups). 56, 60

Ratio scale - The ratio scale is characterized by the fact that ratios of our observations have natural interpretation. Variables with a ratio scale have natural zero, but they do not have natural measurement units. 12, 13, 50, 52

Regression function - description of a dependency of the explained variable (dependent variable) on one or more explanatory variables (independent variables, regressors) via a (usually linear) function based on n observations. The regression function assigns to the values of the explanatory variable some average value (fitted value) which can be very different from the value which was really observed. The difference between the fitted value and the observations is called the **residual**. 455, 457, 468, 472, 473, 475

Rejection region - . 315–318, 327, 329, 333, 334, 336, 344, 345, 355, 374, 385–387, 394, 399, 408

Relative frequency - the ratio of absolute frequency and the total number of observations. 17, 21, 25, 35, 76, 84, 209, 420, 422

Sample - subset of the sample space; the elements which have been chosen for the statistical investigation. 288, 326, 332, 336, 337

Sample mean - arithmetic average of the sample variables $X_1, \ldots, X_n$. 212, 218, 242, 263

Sample space - a set of all possible events of a random experiment. Each event is thus a subset of the sample space. The impossible event is empty set, the sure event is the complete sample space. 10, 69, 70

Sample survey - subset of the sample space; the elements which have been chosen for the statistical investigation. 8

Sample variable - random variable X_i which is defined as the value of the random variable X which will be observed on the i-th element of the sample space. 314, 331, 332, 336, 361, 378

Sample variance - empirical variance of the sample variables $X_1, \ldots, X_n$. 212

Sampling with replacement - sampling procedure. Each selected element is returned before next element is chosen. It corresponds to the simple random sample. 218, 222, 234

Sampling without replacement - sampling procedure. The selected elements are taken out of the sample space before the choice of next element. It corresponds to the representative random sample. 219, 222, 235

Scale - projection of some numerical set (scale) onto the set of investigated statistical elements, such that the relations are preserved. See also nominal scale, ordinal scale, metric scale, interval scale, ratio scale and absolute scale. 435

Scatterplot - graphical display of observed values of a pair of metrically scaled random variables. The values are displayed as a point in the cartesian system of coordinates. It allows to visualize the dependency between the variables. 3D scatterplot can be used for 3 variables. 419, 424, 427, 443, 469, 471, 472, 477

Scatterplot matrix - It is used for graphical display of more than two metrically scaled variables. It contains scatterplots of all pairs of the variables. Attention: with large number of variables, the scatterplot matrix becomes too complex to interpret. 419, 426

Seasonal component - see periodic fluctuations. 489, 494

Spearman correlation coefficient - It measures the strength of linear dependency between two ordinal random variables. It corresponds to the Bravais-Pearson correlation coefficient and its value lies always between -1 and 1. 445, 448, 450

Stacked bar chart - bar chart in which more than one quantity is captured in each bar. 22

Standard deviation - positive square root of the variance. 58, 60, 121, 140, 182, 229, 231, 233, 439, 462, 470, 493

Standard Normal distribution - normal distribution of a continuous random variable with expected value $\mu = 0$ and the variance $\sigma^2 = 1$. 235, 333

Statistical element - one object of the statistical investigation. It carries the information of interest in the experiment. 8, 10, 18

Statistical sequence - the series of observed values (data). The series can be sorted or unsorted. 16

Statistical variable - property of the statistical element. We distinguish identifying and investigated characteristics. 10

Statistics - science allowing to investigate objective empirical information obtained from (random) experiments and questionnaires, to build theoretical models for this information, and to analyze and interpret it. 1, 6, 234

Stem-and-leaf display - half-graphical display of the values of the observed series of a metrically scaled random variable. 28

Support area - space of all possible values that a random variable can assume. 480

T-distribution - distribution of a continuous random variable with parameter f (degrees of freedom). A random variable with t-distribution can be obtained as the ratio of two independent random variables with standard Normal and Chi-square distribution. 206, 220, 298

Target variables - variables of interest in the statistical investigation and whose (varying) values are observed on all statistical elements of the sample space. 10

Test statistic - function of the observed values which is used in the statistical test. 314, 317, 319–321, 327, 332, 334, 344, 349, 352, 356, 362–364, 366, 378, 384, 392, 394, 396, 398, 400, 403, 406–411, 413, 414

Time series - statistical sequence whose values were obtained in a sequence in different time points or time periods. See also components of a time series. 17, 477, 479, 485, 489, 494

Trend - the long-time development of the observed time series. The trend is usually estimated by the method of moving averages of by the Least Squares method (see also filter). 479, 485, 487, 489, 492, 494

Unbiasedness - property of an estimator. The expected value of the estimator is equal to the true value of the estimated parameter. 255, 289, 298

Union of events - The union of two events A and B is a set of all elementary events which belong to A or to B or to both A and B. 71, 79

Variance - the variance is the mean squared error of the observed values from their arithmetic average. 58, 60, 121, 139, 150, 155, 163, 170, 177, 196, 199, 206, 218, 229, 231, 233, 234, 462

Variations - each selection of k elements out of total n elements, where we take the ordering of the elements into account, is called variations of k-th class out of n elements. We distinguish variations with and without repetition (see also combinatorics). 97, 100, 102