Introduction to Statistical Methods in Pathology

Amir Momeni
Matthew Pincus
Jenny Libien

Springer

Introduction to Statistical Methods in Pathology

Amir Momeni • Matthew Pincus
Jenny Libien

Introduction to Statistical Methods in Pathology

Springer

Amir Momeni
Department of Pathology
State University of New York
Downstate Medical Center
Brooklyn, New York, USA

Matthew Pincus
Department of Pathology
State University of New York
Downstate Medical Center
Brooklyn, New York, USA

Jenny Libien
Department of Pathology
State University of New York
Downstate Medical Center
Brooklyn, New York, USA

ISBN 978-3-319-60542-5 ISBN 978-3-319-60543-2 (eBook)
DOI 10.1007/978-3-319-60543-2

Library of Congress Control Number: 2017944203

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

To an ever-increasing extent, pathologists are being required to use statistics in their practice. In clinical pathology or laboratory medicine, statistics are a fundamental requirement for the evaluation of the reliability of quantitative results for values of serum and, in general, body fluid analytes such as electrolytes, glucose, blood urea nitrogen (BUN), creatinine, critical enzymes, etc. and for the analysis of the correlation between the results generated on different analyzers, all of which are used for quantitative determination of the same analytes. Correlations between measurements of parameters that allow categorization of tumors such as correlation of nuclear grade with pathological stage require knowledge of statistical methods in anatomic pathology. Correlation of the staging of different cancers with survival involves another major use of statistics in both anatomic and clinical pathology.

Often, pathologists utilize statistical methods without knowledge of the physical and mathematical basis that underlies the particular statistics that they are using. This can give rise to erroneous conclusions. For example, many, but certainly not all, quantitative analyses for analytes follow so-called Gaussian statistics, an example of parametric statistics, with a known mathematical form for the distribution of values that gives rise to the "bell-shaped curve," called the normal distribution. This involves computation of means, standard deviations, confidence intervals for means, and a number of other parameters.

However, these methods cannot be used for analyte values that do not follow Gaussian statistics which requires that the distribution of values for a given analyte be distributed in what is termed a normal distribution as represented by the so-called bell-shaped curve. This can affect determinations such as the reference ranges for analytes based on values determined from presumed normal or well individuals. If the distribution of values is assumed to be Gaussian, the range would be computed as the mean of the values plus or minus two standard deviations from the mean. However, if the values actually do not follow a Gaussian distribution, serious errors can be made in establishing the reference range which may be too narrow or too wide. Not infrequently, the use of non-parametric statistics rather must be used in establishing reference ranges.

We are currently living in what has been termed "the age of metrology." This means that, to an increasing degree, statistics govern most aspects of laboratory medicine including whether or not values can be accepted as being "true" or

"reliable," and the criteria for acceptability are being made more stringent. This raises the question as to why statistics are considered essential in evaluation of clinical laboratory results.

Statistics provide a means for at least partially removing arbitrariness for making such critical decisions as whether results are acceptable or whether two sets of data are actually the same or are different. However, there are limitations to statistical analysis.

In all statistical analysis, there is some arbitrariness. For example, analysis of the concentration of an analyte in a control sample is said to be "acceptable" if the value lies between plus or minus two standard deviations of the mean determined for concentration of this analyte in the control sample on a clinical chemistry analyzer. The reason for this two standard deviation rule is that, for a Gaussian or normal distribution, the mean plus or minus two standard deviations encompass about 95% of the possible values. All other values are considered to be "outliers." This is an arbitrary number. One can inquire why some other number might be used such as 97% (allowed by using approximately three standard deviations from the mean) or some other number. Here, there is no definitive answer.

Given the current metrology requirements and their acceptance by federal and state regulatory agencies and by most laboratorians and given the necessity for use of statistics in analyzing specific clinical data, it is desirable to introduce pathologists to the statistical methods available to them so that they understand what methods to use in analyzing clinical data and how to use them. It is the purpose of this textbook to achieve this goal.

Our aim, therefore, is to impart to the reader how to evaluate different types of data using the appropriate statistical methods and why these methods are used, rather than to refer the reader to specific programs that analyze the data without explanation of the basis of the methods used. In this textbook, we present the most commonly used statistical methods in the field of pathology. Our presentation is based on three simple steps:

1. *Definition of the statistical problem*. For example, when a control is assayed, the statistical problem is to determine whether the result is acceptable or not acceptable.
2. *The mathematical form of the statistical distribution that solves the statistical problem*. Using the same example given above, since the assay is performed on the same control repeatedly, any deviation of the values from one another should be random, i.e., there is random error. Random error is described by the Gaussian distribution, i.e., when the probability of getting a particular value is plotted against the values themselves, a bell-shaped curve is obtained. The mathematical form for this bell-shaped probability distribution is the exponential form ae^{-bx2}, where x is any value determined experimentally and a and b are constants related to the standard deviation.
3. *How to compute the significance of results obtained from data obtained in the medical laboratory using the appropriate distribution*. The mean for the Gaussian distribution can be shown to be the most probable value on the bell-shaped

curve and equals the median value. From the Gaussian distribution, one standard deviation from the mean can be computed. It can further be shown that approximately 95% of all values lie within the width of the bell-shaped curve at two standard deviations. It happens that one standard deviation can also be computed as the square root of the sum of the squares of the differences between each value determined experimentally and the mean value divided by the number of values.

Thus, as we discussed above, if we wish to define acceptability of a value as any value that lies within two standard deviations of the mean value, then if a result is within this cutoff, i.e., plus or minus two standard deviations from the mean value, it is acceptable.

The textbook is arranged so that the most commonly used statistics in pathology are discussed first in Chap. 2 in which normal or Gaussian distributions are described; the concepts of accuracy and precision are discussed; the evaluation of test efficacy, i.e., sensitivity, specificity, and positive and negative predictive value, is presented; and the evaluation of so-called receiver operator curves is performed in deciding which of two or more tests has better or best diagnostic accuracy.

Chapter 3 then presents general probability analyses and discusses probability distributions that are not used as frequently in pathology but may be useful, especially the ones involving conditional probabilities.

Chapter 4 presents the underlying theory for analyzing correlations, e.g., when samples are analyzed on two or more analyzers, what criteria are to decide whether the values obtained on assaying samples can be considered to be the same or different. This chapter discusses how to fit straight lines to experimentally determined points, a process termed linear regression analysis, and how to decide how well the "best fit" line fits the points.

Chapter 5 provides the statistical basis for the all-important question as to whether a new test for diagnosis of a particular disease is valid. Generally, these tests provide a yes or no answer, i.e., the results are discrete and not continuous. It happens that the distribution that is most appropriate for answering this question of reliability of the test is given by the chi-squared distribution. A major point of this chapter is to illustrate that, although the results of this type of testing are discrete, it is possible to represent the probability distribution for right or wrong results as a continuous function so that cutoffs such as those used for the Gaussian distribution can be used and quantitative decisions can therefore be made.

Chapter 6 addresses the statistical basis for the comparison of two or more sets of data to determine whether they are the same or different. This involves specific tests on the mean values for the sets of data.

Chapter 7 discusses multivariate analysis, i.e., extension of the linear regression analysis discussed in Chap. 4 to linear regression with more than two variables.

Chapter 8 presents methods for inferring values omitted from datasets that are necessary for statistical analysis.

Chapter 9 presents the statistical solution to a problem that is common to all medical practice: survival analysis. Many readers may be familiar with Kaplan-Meier curves for survival of patients who carry specific diagnoses or who are being

treated for specific diseases. This chapter explains the statistical basis for this type of analysis and other approaches that achieve the same goal.

Chapters 10 and 11 deal with quality assurance. Chap. 10 addresses how methods are quantitatively validated and Chap. 11 discusses the rules for evaluating quality control.

Chapters 12 and 13 deal with the problems of how to evaluate quantitatively and how to design diagnostic studies.

Chapter 14 is an introduction to statistical analysis of large datasets. This type of analysis is now becoming of paramount importance as the amount of genetic information on patients has been increasing exponentially. In this chapter, the technique of clustering, which allows for data simplification, is discussed.

We hope that the readers of this textbook will find it helpful to them in evaluating data in clinical practice and/or in research.

Brooklyn, NY, USA

Matthew R. Pincus
Amir Momeni
Jenny Libien

Contents

Why Every Pathologist Needs to Know Statistics

1

Introduction

Statistics permeates our lives as pathologists. We use statistics in the interpretation of laboratory tests, in deciding whether to use a new immunohistochemical stain or diagnostic method, for our research projects, in our critical reading of scientific literature, and in our quality improvement and laboratory management activities [1]. Although we regularly use statistics as pathologists, we do not understand statistics as well as we would like. In a survey of pathologists to assess statistical literacy (Schmidt et al., Arch Pathol Lab Med, 2016) [2], the majority of pathologists surveyed expressed the desire to have a better understanding of statistics. This book aims to help pathologists achieve a higher level of statistical literacy and gain greater comfort in using statistical methods.

We start now with the basics – the definitions of keywords. The Merriam Webster definition of statistics is "a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data" or "a collection of quantitative data." For pathologists, statistics can be thought of as a way to make sense of our observations and measurements.

Outcomes and Variables in Pathology and Laboratory Medicine

The test result can have different formats. We can consider test results as variables. 'Variables' are data items which can be counted or measured. Variables may be categorical, which is also known as nominal, and have a limited number of data items such as positive staining vs. negative staining for a surgical pathology project examining the use of a new immunohistochemical stain. Alternatively, variables may be ordinal in which the variables are ordered, but there is still a limited number of data items (but often more than in categorical). An example of an ordinal scale could also use that same surgical pathology project examining the use of a new immunohistochemical stain; however, a semiquantitative assessment is performed,

A. Momeni et al., *Introduction to Statistical Methods in Pathology*,
DOI 10.1007/978-3-319-60543-2_1

and the percent of cells stained are ranked as 0 for none, 1+ for mild staining, 2+ for moderate staining, or 3+ for strong staining. Interval variables are quantitative, and the differences between the variables are equal. They can be continuous, with infinite subdivision, or discrete, with a set of fixed values such as age in years. Quantitative clinical laboratory values, with results such as 3.21 mmol, are examples of interval variables and are often continuous interval variables.

Laboratory test results may vary due to intra-subject differences such as physiologic changes due to circadian rhythms, intra-observer variation, and interobserver variation. Variation is also due to preanalytic (prior to specimen testing), analytic (during testing), and postanalytic (after testing such as during computer entry or data transmission) variables. Sometimes the variation occurs due to pure chance and randomness. Consequently, it is important that we can identify when a change in a test result reflects a random variation or a variation that occurred due to non-pathologic reasons or a variation that occurred because the patient is affected by a disease.

The true outcome of tests irrespective of their format is the clinical impact they have on diagnosis and/or prognosis of patients. Clinical applicability of tests is dependent on different elements: test characteristics, cost, ease of performance, and clinical accuracy.

True outcome studies evaluate the impact of a new laboratory test on overall health, healthcare costs, morbidity, and mortality. Although seldom performed, these studies would determine the true usefulness of a new laboratory test. Laboratory tests can be accurate and reliable but cause harm to a patient when the wrong test is ordered, or the test result is not used properly. Inaccurate test results, although rare, may occur.

Through the application of statistical theory and statistical analysis, we can determine whether test results can have a true effect on patient outcome or clinical practice; the value of a test relies on its contribution to the clinical management of the patients, and measurement of this contribution requires applying statistical tests.

The underlying concept of laboratory medicine and pathology is that we can identify, measure, or quantify variables in individuals that can help to diagnose them with different diseases. The aim of most test outcomes in pathology is to identify and distinguish a diseased individual from a person unaffected by that disease. These measurements are meaningless on their own; it is only through experimental studies and application of statistical concepts that we can define a test outcome relevant to the diagnosis of a disease. The basic principle for testing is that a test is applied to the population to obtain the population average. The next step is to apply the test to diseased and healthy individuals and determine whether the test can contrast between the two states using robust statistical methods. Thus, understanding of test characteristics (such as sensitivity and specificity) and implementation of tests require that the pathologists have a basic understanding of the statistical concept behind these characteristics.

Statistical tools are either descriptive or inferential. Descriptive statistics are used to summarize data – mean, standard deviation, range, and sample size are examples. Inferential statistics are used to compare results between groups or

populations. Throughout the course of this book, we will show how descriptive and inferential statistical tools can help to transform test results into meaningful patient relevant outcomes [3–5].

Components of a Useful Diagnostic Test

Quantitative clinical laboratory test results are more than just "interval variables" – we use the results to guide diagnostic and therapeutic decision making and to determine prognosis. Validity, accuracy, reliability, sensitivity, specificity, and usefulness all contribute to whether a new test becomes part of our arsenal. Validity refers to how well a test measures what it is supposed to measure. There can be analytical validity which indicates how well a test detects the presence or absence of a particular change, for example, a mutation, and there can be clinical validity which indicates how well a test detects the presence or absence of a disease. Accuracy is when a test result is near the absolute true value as determined by control specimens which have also been evaluated by the "gold standard" testing method. Accuracy can be calculated as true positives + true negatives/all individuals tested. Reliability indicates that the test is reproducible – that repeat tests will give similar (nearly identical) results. Sensitivity and specificity are measures of diagnostic accuracy. Sensitivity is a measure of a test's ability to detect true positives among all those individuals who have the disease and is calculated as true positive / (true positive + false negative). Specificity is a measure of a test's ability to detect true negatives among all the individuals without the disease and is calculated as true negative / (true negative + false positive) [6–8]. We will revisit these accuracy measures in Chap. 2.

Examples

Sensitivity
A new laboratory test is used to detect a malignant brain tumor in 100 patients who are known to have malignant brain tumors using the gold standard of brain biopsy. The new lab test is positive in 95 of the 100 patients. The number of true positives is 95 and the sensitivity is calculated as 95/100 = 95%. The false negative rate is 5%. That is, five of the individuals who tested negative by the new lab test actually had the disease.

$$\text{Sensitivity} = \text{number of true positives}/(\text{Number of true positives} + \text{number of false negatives})$$

Specificity
This new laboratory test for malignant brain tumors is also used on 100 individuals who do not have brain neoplasms by neuroimaging. The test is negative for 95 of the individuals and is positive in 5 of the individuals. The specificity is 95%.

$$\text{Specificity} = \text{number of true negatives}/(\text{Number of true negatives} + \text{number of false positives})$$

The usefulness of a test is related to the accuracy of a test as determined by the sensitivity and specificity and by the prevalence. Positive predictive value is the probability that a positive test indicates disease and is calculated as true positive / (true positive + false positive). It indicates the reliability of a positive result. The negative predictive value is the probability that a negative test indicates no disease and is calculated a true negative / (true negative + false negative). It indicates the reliability of a negative result. Predictive value depends on the prevalence of the disease, with higher prevalence leading to higher predictive values. Correspondingly, if a test is of low prevalence, the positive predictive value will be low, and there will be more false positives. Remember that high prevalence means high pretest probability, and low prevalence means low pretest probability. Therefore, the utility or usefulness of a test depends on the prevalence of the disease being tested for in the population.

Example Malignant brain tumors are present in 10 of 1000 individuals based on using the gold standard for diagnosis. The positive predictive value of the new laboratory test can be calculated using the known sensitivity and specificity. Ninety-five percent of 10 patients (= 9.5) who have the brain tumor test are positive with the new test. The test is also falsely positive in 5% of the 990 (= 49.5) individuals who do not have the disease. The specificity is also 95% and 940.5 individuals test as true negatives.

$$\begin{aligned}\text{Positive predictive value} &= \text{True positives}/(\text{True positives} + \text{false positives}) \\ &= 9.5/(9.5 + 49.5) = 16.1\%\end{aligned}$$

$$\begin{aligned}\text{Negative predictive value} &= \text{True negatives}/(\text{True negatives} + \text{false negatives}) \\ &= 940.5/(940.5 + 0.5) = 99.9\%\end{aligned}$$

Defining Test Objectives

When evaluating tests, it is imperative to know the objective for that test. Different objectives require the test to have different characteristics. A common example in pathology is when a test is used for screening versus diagnosis.

Screening tests are meant to identify disease when asymptomatic (preclinical) or when the disease can be more easily treated. In contrast, diagnostic clinical laboratory testing is performed to identify the presence or absence of disease when an individual shows clinical signs and/or symptoms of the disease. Screening is performed as part of preventive medicine and is used for early detection. Therefore, screening tests are designed to have many true positives, few false negatives, and more false positives than in a diagnostic test. Diagnostic confirmatory testing needs

to be performed after a positive screening test. Screening tests are not meant for hospitalized, ill individuals.

Summary

Even as end users, pathologists are required to understand the evidence base that supports a diagnostic test; this requires them to understand the statistical tools that were employed in the studies to prove the utility of a test. Furthermore, and in order to integrate a test into their laboratory and practice, they are required to understand the statistical concepts that allow them to use that test for clinical decision making. Starting from the next chapter, we will explain how we can assess diagnostic tests and define their characteristics.

References

1. Marchevsky AM, Walts AE, Wick MR. Evidence-based pathology in its second decade: toward probabilistic cognitive computing. Human Pathology. 2017;61:1–8.
2. Cohen MB, Schmidt RL. More on Evidence-Based Medicine. Archives of pathology & laboratory medicine. 2016;140(1):10.
3. Marchevsky AM, Wick MR. Evidence-based pathology: systematic literature reviews as the basis for guidelines and best practices. Archives of Pathology and Laboratory Medicine. 2015;139(3):394–9.
4. McPherson RA, Pincus MR. Henry's clinical diagnosis and management by laboratory methods. Elsevier Health Sciences; 2016.
5. Burtis CA, Ashwood ER, Bruns DE. Tietz textbook of clinical chemistry and molecular diagnostics. Elsevier Health Sciences; 2012.
6. Pencina MJ, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Statistics in medicine. USA. 2008;27(2):157–72.
7. Laskowitz DT, Kasner SE, Saver J, Remmel KS, Jauch EC, BRAIN Study Group. Clinical usefulness of a biomarker-based diagnostic test for acute stroke. Stroke. USA. 2009;40 (1):77–85.
8. Bonini P, Plebani M, Ceriotti F, Rubboli F. Errors in laboratory medicine. Clinical chemistry. 2002;48(5):691–8.

Assessing Diagnostic Tests

2

The most important question in diagnostic medicine is "Does this individual have a disease?". The entire field of laboratory medicine and pathology has been developed to aid in answering that question. To be clinically relevant, a diagnostic test needs to be able to differentiate between the diseased and healthy state, and it also needs to be accurate and precise. Furthermore, a good diagnostic test should be clinically applicable, it should not cause harm, and in the face of ever-increasing constraints on healthcare finances, it needs to be cost-effective. In this chapter, we will address different aspects of assessing a diagnostic test [1].

The first step in assessing a diagnostic test is to examine the theoretical concept of the test and establish a causal linkage between the test and the condition of interest. The test methodology and instrument also need to be scrutinized. The precision and accuracy of the measurement instrument and test should be determined. The measurement error needs to be quantified and minimized if possible. These concepts are collectively called technical accuracy and precision.

The next step in assessing a diagnostic test is to establish the discrimination power of the test, the ability of the test to differentiate those affected by a condition from the unaffected. This step requires carefully designed clinical trials to determine the diagnostic metrics of the test and establish its accuracy. The level of diagnostic accuracy and the accuracy metrics that are used are dependent on the possible application of the test; a screening tool needs high sensitivity, while testing for a very rare condition requires high specificity. These evaluations fall under the umbrella of diagnostic accuracy [2–5].

The next critical question in assessing a diagnostic test involves determining the possible effect of the test on patient outcomes. This step involves the crucial tradeoff of benefit versus harm; the benefit of the diagnostic test should be weighed against possible adverse outcomes for the patient or the population. Also in this step, questions of applicability and feasibility should be addressed.

Finally, the cost of the new diagnostic test should also be addressed. Cost can potentially be the single most prohibitive step in adopting a new test, and to justify possible additional expenses, the cost-effectiveness of the test should be determined.

A. Momeni et al., *Introduction to Statistical Methods in Pathology*,
DOI 10.1007/978-3-319-60543-2_2

Some of the questions relating to appraisal of new diagnostic tests will be covered in Chap. 13, where we will discuss an evidence-based approach to appraisal of diagnostic studies which establish the scientific basis for new tests. In this chapter, we will start with the concept of technical accuracy and precision focusing on measurement error and statistical analysis of error. Next will be the concept of diagnostic accuracy focusing on discriminative and predictive powers of the test. Clinical impact or clinical applicability is the next step in assessing a diagnostic test. The final part of this chapter provides a brief introduction of cost-effectiveness analysis [6–8].

Technical Accuracy and Precision

Technical accuracy is the ability of a test to produce valid and usable information. Precision is essentially the reproducibility of the test, the ability to obtain very similar results if the test is repeated multiple times. Technical accuracy and precision should be determined for every new diagnostic test that is being developed, and subsequently every time a laboratory adds a test to its repertoire, it must ensure that the test is technically accurate and precise under its laboratory conditions. Evaluation of technical accuracy and precision should be an ongoing effort. Technical accuracy and precision are essentially about minimizing measurement error.

Error

Every measurement in laboratory medicine has a degree of uncertainty; this uncertainty is called "error" and refers to imprecisions and inaccuracies in measurement. This measurement error refers to the difference between the true value of the measured sample and the measured value. Effectively, the results we report are best estimates of the true value.

Understanding the nature of error and quantifying it is of utmost importance in laboratory medicine as the results can have direct clinical impact on patients. High precision instruments have limited the measurement error in recent years, yet we still should estimate the error of our measurements and take corrective actions when the error surpasses an acceptable threshold.

Two main important forms of error are "random error" and "systematic error" (Fig. 2.1). The effects of systematic error and random error are additive. Random errors are caused by unpredictable changes in the measurement which may be related to instrument, sample, or environment. Addressing the causes of random error is usually difficult, and there is always a degree of random error present for every measurement.

For example, if you measure the sodium concentration of a solution with a sodium content of 140 mEq/l five separate times with results being 140 mEq/l, 141 mEq/l, 139 mEq/l, 138 mEq/l, and 142 mEq/l, then you are witnessing a

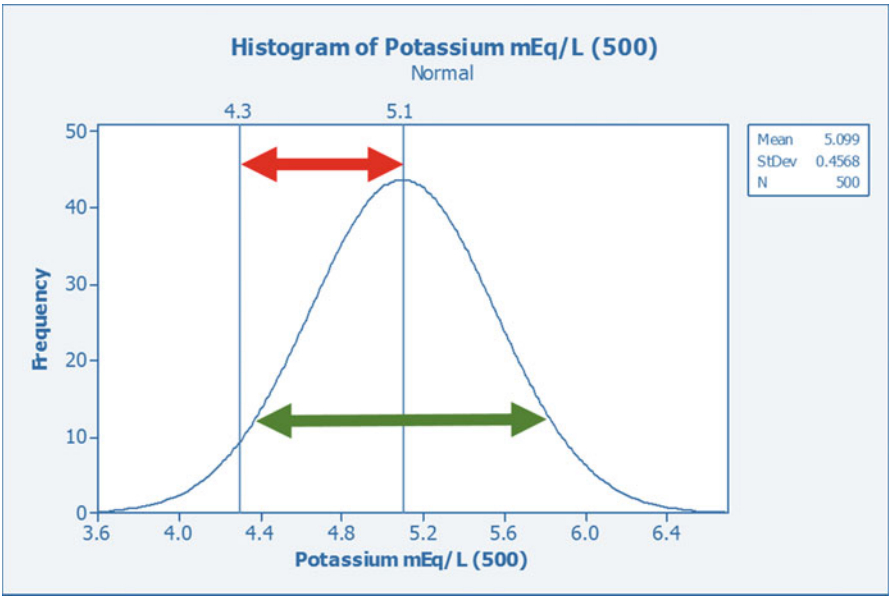


Fig. 2.1 These are the results of 500 repeated measurements of a sample with potassium concentration of 4.3 mEq/L. The *red line* shows the inaccuracy of results (systematic error), and the *green line* shows the imprecision of results (random error)

random error. The variations in results are random and cannot be predicted. However, random errors, as driven by chance, follow a Gaussian normal distribution; this allows us to use statistical analysis to quantify and address random error in our measurements. The degree of random error determines the "precision" of a test/instrument. Random error can be minimized by increasing the number of measurements. Averaging repeated measurement results is one way to report a more precise estimate of the expected value. Mean or average $(\bar{x})$ is the sum of measurement results divided by the number of measurements.

$$\text{Average } (\bar{x}) = \frac{x1 + x2 + x3 + x4 + x5 + \ldots + xN}{N} \tag{2.1}$$

Theoretically, with infinite measurements, the mean of measurement results will be the true value. With a finite number of measurements, the true value will be within a range of the measurement mean. The range is mainly determined by the number of measurements (with more measurements the range will be narrower). This range is called the "confidence interval" and will be addressed later in this chapter.

The simplest form of random error is called "scale error." Scale error refers to the precision of an instrument that makes a measurement as well as the precision of the reporting of the result. The measurement and reporting can be as integer numbers (i.e., 1, 2, 3, 4, ...), and a true sample value of 4.2 will be reported as

4. The 0.2-unit imprecision is due to scale error. The scale errors for instruments are determined by the resolution of measurement; higher resolution instruments will provide a more precise result. While scale error can be minimized, it can never be totally rectified as there is always a limit to the resolution of the instrument. Different tests require different resolutions and as such the scale precision differs between them. For example, in measuring cardiac troponins, a much better resolution is needed compared to measuring sodium levels. In laboratory medicine, test scales are determined by the nature of the test as well as the clinical significance of the scale. As such, the scale imprecision of tests is usually clinically insignificant. For example, a sodium level of 133 mEq/l versus 132.987 mEq/l is considered as clinical equivalents.

Systematic error is a nonzero error; averaging or repeating the results will not minimize the error. Systematic errors are reproducible and skew the results consistently in the same direction. Systematic error is otherwise known as bias. Bias can be difficult to identify and address. In laboratory medicine, the most common method of addressing bias is to use calibration. In calibration, a standard sample at different concentrations is measured, and the difference between the results and the expected value (bias) is reduced by using a correction factor. Systematic error has different causes including environmental factors, calibration problems, instrument drift, confounding factors, and lag time errors. Systematic error determines the accuracy of the results [9, 10].

Accuracy refers to the proximity of the measured value to the expected (true) value. Precision, on the other hand, deals with repeatability of the results and refers to consistency of results from repeated independent measurements. Precision is a measure of reliability and reproducibility. For each test, both accuracy and precision need to be addressed. These concepts are shown in Fig. 2.2.

There are different ways of reporting precision including fractional uncertainty and confidence interval. Fractional uncertainty is the ratio of uncertainty to the measured value. The confidence interval will be discussed later in this chapter.

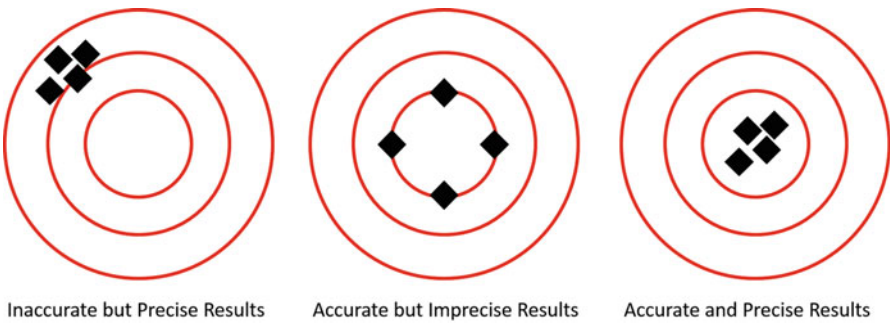


Fig. 2.2 This figure depicts the concepts of accuracy and precision

$$\text{Fractional uncertainty} = \frac{\text{Uncertainty}}{\text{Measured value}} \tag{2.2}$$

Accuracy can be reported as relative error and shows the ratio of drift to the true value. Note that the relative error is directional (can be positive or negative) with a minus relative error signifying a systematic error that underestimates the result.

$$\text{Relative error} = \frac{(\text{Measured value} - \text{True value})}{\text{True value}} \tag{2.3}$$

Standard Deviation

Standard deviation (σ) or (SD) is the uncertainty associated with each single measurement. In other words, standard deviation shows the degree of variation (spreading) of measurement results. Standard deviation is a useful measure in variables that follow a Gaussian normal distribution. Higher standard deviations signify a wider spread of data (Fig. 2.3).

Standard deviation is the square root of the variance (σ^2). Variance is the sum of squared deviation of every measurement from the mean:

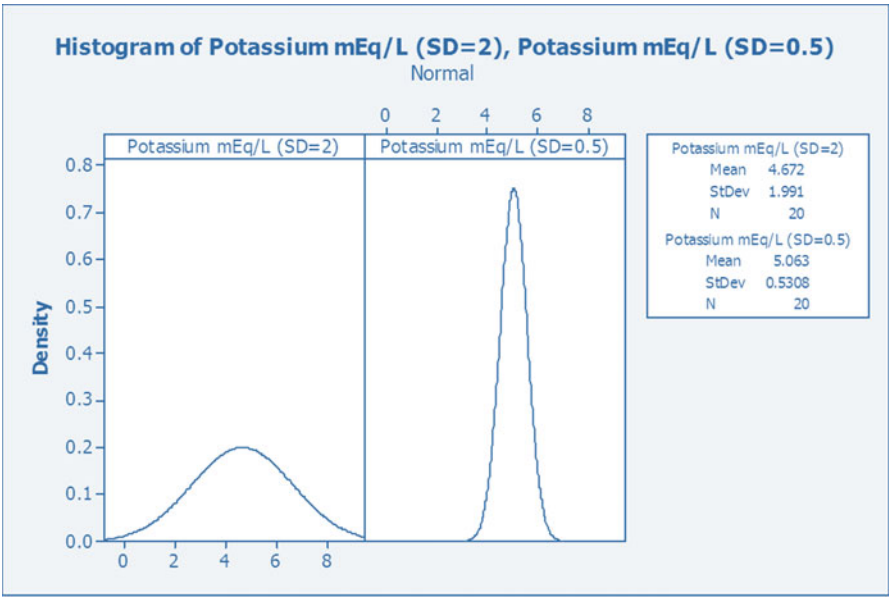


Fig. 2.3 Histogram plots of potassium measurements with a SD of 2 and 0.5. Note that as the SD increases the Gaussian bell curve becomes wider (wider spreading)

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N} \tag{2.4}$$

where N is the number of measurements (or size of the sample) and $(\bar{x})$ is the sample mean.

Thus, standard deviation can be written as

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}} \tag{2.5}$$

As standard deviation shows the spreading of uncertainty of a measurement, it is used in calculating standard deviation of mean (also known as standard error) which is in turn used to calculate the confidence interval.

Confidence Interval

Confidence interval (CI) is the range of values, estimated from sample data, which is likely to include a population parameter (Θ). In epidemiologic studies, the population parameter is usually the population mean (μ), and the sample mean $(\bar{x})$ is used to measure the confidence interval. In laboratory medicine, the parameter is usually the result of a test with the confidence interval being a range which is likely to include the actual measurement.

Confidence interval is centered around the measured parameter (either sample mean or test result) with the range defined by level of confidence (C). Level of confidence refers to the probability that the range contains the actual value. As the level of confidence increases, the range becomes narrower, or, conversely, the lower the level of confidence, the broader the range will be. Confidence levels are usually set at 90%, 95%, or 99%. A confidence interval of 95% means that there is a 0.95 probability that the actual value is within the range provided. For most measurements, a level of confidence of 95% is considered acceptable. Figure 2.4 shows the confidence interval on a normal density curve. For the purposes of this

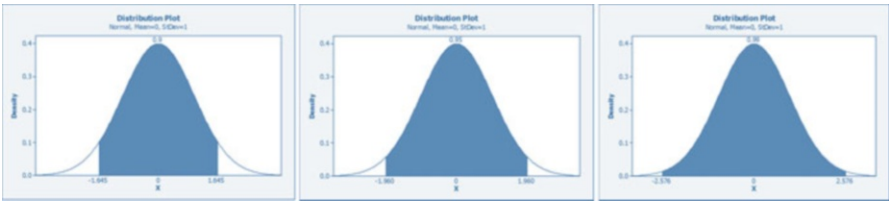


Fig. 2.4 Normal density curves depicting 90%, 95%, and 99% confidence interval of a normally distributed sample with a mean of 0 and standard deviation of 1

chapter, we assume that the measured value follows a normal distribution. Data from large sample size that does not follow a normal distribution can be approximated to a normal distribution using the central limit theorem which is discussed in the next chapter.

As the number of measurements increase (in population statistics as the sample size increases), the confidence interval will become narrower. Repeated measurements reduce the effect of random error on the mean test result thus leading to increased precision. As you will see below, confidence interval is a function of the mean $(\bar{x})$, confidence level, standard deviation (σ), and sample size (n). The larger the sample size, the less effect will standard deviation have on the measurement (i.e., the smaller the standard error will be). Figure 2.5 shows the effect of repeated measurements on increased accuracy of prediction.

In cases where the mean (μ) is unknown but the standard deviation (σ) is known, the formula for confidence interval (CI) is:

$$\mathrm{CI} = \bar{x} \pm z\frac{\sigma}{\sqrt{n}} \tag{2.6}$$

The z-score is used for data that follow a normal distribution. The z-scores for 90%, 95%, and 99% level of confidence are 1.645, 1.96, and 2.576, respectively. In the potassium sample with 500 measurements, we have a sample mean of 5.099 and sample size of 500. If the standard deviation was known to be 0.5, then we can calculate the 95% CI:

$$95\%\mathrm{CI}_{\mathrm{potassium}(500)} = 5.099 \pm 1.96\frac{0.5}{\sqrt{500}} = 5.055 - 5.142 \tag{2.7}$$

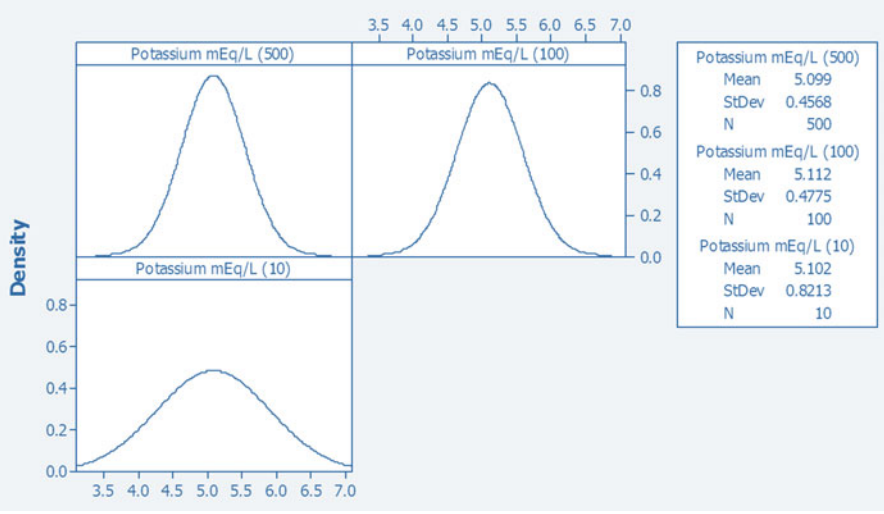


Fig. 2.5 Histograms showing repeated measurements of potassium in a 5.1 mEq/L potassium solution. As the number of measurements increases, the confidence interval becomes narrower

$\frac{\sigma}{\sqrt{n}}$ is also known as standard error of mean ($\sigma_{(\bar{x})}$). Thus, the confidence interval can be simplified as

$$\mathrm{CI} = \bar{x} \pm z^* \sigma_{\bar{x}} \tag{2.8}$$

If the mean and standard deviation are both unknown, or if the sample size is too small (<30) for *z*-scores to be used, then an alternative formula is used for calculating confidence intervals. If the standard deviation is unknown, then the sample standard deviation (*s*) is used as an estimate of population standard deviation.

$$\mathrm{CI} = \bar{x} \pm t\frac{s}{\sqrt{n}} \tag{2.9}$$

In these cases, the confidence level is determined by *t* distribution with *n*−*1* degree of freedom. Thus, in the potassium sample with 10 measurements, we have a sample mean of 5.102, sample standard deviation of 0.8213, and sample size of 10. The *n*−*1* *t*-score (10–1 = 9) for a 95% CI is 2.262. Subsequently, the 95%CI for the sample is:

$$95\%\mathrm{CI}_{\mathrm{potassium}(10)} = 5.102 \pm 2.262\frac{0.8213}{\sqrt{10}} = 5.689 - 4.514 \tag{2.10}$$

Note that the *t*-scores provide a wider confidence interval compared to *z*-scores.

As sample size increases, the *t*-scores will become closer to *z*-scores. Also as the sample size increases, the standard deviation of the sample will be closer to the standard deviation of the population. When measuring potassium 500 times with a known standard deviation of 0.5, we observe that the sample standard deviation is 0.456, while measuring the potassium 10 times provides a sample standard deviation of 0.821. In other words, repeated measures can lead to increased accuracy [11–19].

Calculating Reference Intervals

Reference interval refers to the range of values of a measurement in healthy individuals (e.g., the range of potassium in healthy adults is 3.5–5.1 mEq/L). Reference interval is a very important information that should be provided with every quantitative test to allow the clinicians to interpret the results and determine if a patient's results are abnormal. In the next section, where we introduce diagnostic accuracy, you will see that reference interval and its overlap with values from diseased individuals has a big role in discriminative power of a diagnostic test.

If a result is within the reference range, then these results are within a certain distance of the population mean and part of normal distribution. These results are alternatively known as within normal limit (WNL). The upper and lower limits of the normal distribution of the mean are determined using the population standard

deviation. It is generally accepted that the normal limit is ±2SD of the mean (with 95% of the healthy individual results falling within the normal limit). If a reference range is calculated in this manner, then it is called standard range. The measurements used in calculating reference ranges come from a population of healthy individuals. However, if characteristics of subgroups of the population affect the measurement, then a different reference interval should be calculated and used for each subgroup (e.g., creatinine reference range is different based on gender).

The most straightforward way to calculate a reference interval is to measure the values in a reference group of healthy individuals and sort the values from the least to the most. In this method, results that are at the 2.5–97.5% percentile (or any arbitrary cutoff) will be considered as the lower and upper limit of the reference interval, respectively. This method, despite simplicity, is not adequately reliable, and it is generally preferred that the reference interval is calculated using an arithmetic normal distribution or log-normal distribution method (discussed in Chap. 3). However, in instances where the data does not follow a Gaussian or log-normal distribution, this method can be employed.

In calculating the reference range for a variable, the assumption is that the measurements of the variable in the population follow a normal Gaussian distribution. As the population mean and standard deviation are usually unknown, then they must be estimated using a sample of the population. Using these estimates, then the 95% prediction interval (95%PI) is calculated as

$$95\%\text{PI} = \bar{x} \pm t_{0.975,n-1}\sqrt{\frac{n+1}{n}}\sigma \tag{2.11}$$

In cases where the sample size is greater than 30, the t distribution is considered as equaling 2.

For example, using a sample of 30 patients with an average potassium level of 4.5 mEq/L and standard deviation of 0.5, we can calculate the reference range for potassium as follows:

$$95\%\text{PI} = 4.5 \pm 2 \times \sqrt{\frac{31}{30} \times 0.5} = 4.5 \pm 1.01 \tag{2.12}$$

with the upper limit of reference range being 5.51 mEq/L and the lower limit of reference range being 3.49 mEq/L.

The reference interval can have its own confidence interval. This confidence interval is dependent on the standard deviation of the standard reference interval. The size of the standard deviation is a logarithmic function of the size of the sample with larger sample leading to smaller standard deviation (Fig. 2.6).

In our previous example, the standard deviation of the standard reference interval for a sample size of 20 is 0.4 of the primary value or in other words $0.4 \times 1 = 0.4$ mEq/L. Consequently, we can estimate the 95% confidence interval of the reference range limits as

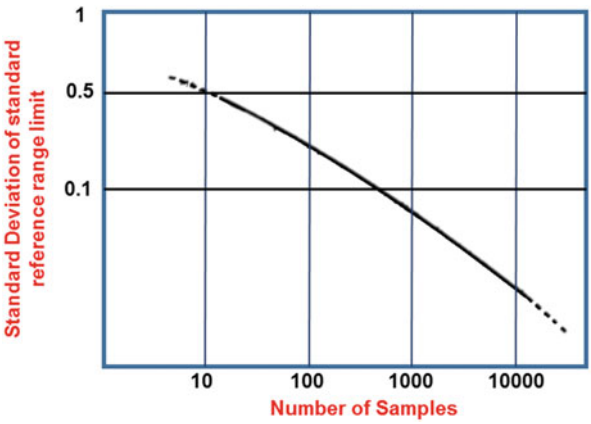


Fig. 2.6 This log-log graph shows the standard deviation of standard reference range limit versus the number of samples

$$95\%\text{CI}(\text{upper reference limit}) = 6.55 \pm 2 \times \sqrt{\frac{31}{30} \times 0.4} = 6.53 \pm 0.81 \quad (2.13)$$

$$95\%\text{CI}(\text{lower reference limit}) = 2.45 \pm 2 \times \sqrt{\frac{31}{30} \times 0.4} = 2.47 \pm 0.81 \quad (2.14)$$

These calculations are correct for all measurements that follow a Gaussian normal distribution. Goodness of fit tests such as Kolomogorov-Smirnov or Shapiro-Wilk can be employed to determine if the data has a Gaussian normal distribution.

Many laboratory tests, however, follow a log-normal distribution. One of the main reasons for this is the fact that most physiologic parameters that are measured can only assume nonnegative numbers (i.e., the results are always positively skewed). In these tests, unless the standard deviation is small compared to the mean, the Gaussian normal distribution cannot be used, and instead a log-normal distribution should be used. In other words, in measurements where the standard deviation is small compared to the mean, even if the sample measurements are positively skewed, the abovementioned calculation can still be used. As the standard deviation increases, however, log-normal distribution should be employed.

The simplest way for calculating the reference interval for a test with log-normal distribution is to calculate the natural logarithmic values of all the measurements. Consequently, arithmetic normal distribution reference interval calculations can be used to determine the lower and upper limits of the logarithmized values. The exponentiated values of these upper and lower limits will form the upper and lower limits of the reference value.

The switch to a log-normal distribution is made based on a difference ratio for the lower and upper limits. The difference ratio can be calculated as

$$\text{Difference ratio} = \frac{|\ \text{Limit}_{\text{Log-normal}} - \text{Limit}_{\text{Normal}}\ |}{\text{Limit}_{\text{Log-normal}}} \tag{2.15}$$

This difference ratio should be calculated separately for the lower limit and the upper limit. A difference ratio of more than 0.1 is considered as indicative of the need to use the log-normal distribution. The calculation of difference ratio, however, can be a cumbersome task, and thus a measure known as coefficient of variation can be used as a proxy for difference ratio. Coefficient of variation (CV) is the ratio of standard deviation to the mean.

$$\text{CV} = \frac{\sigma}{\bar{x}} \tag{2.16}$$

The lower limit of reference range is more sensitive to coefficient of variation, and a CV of 0.213 is the threshold for using a log-normal distribution for the lower limit. For the upper limit, due to positive skewedness of data, a higher CV of 0.413 is considered as the critical threshold.

In general, it is always a good idea to provide a histogram of values used for determining the reference value to the clinicians. This will allow them to better understand the reference interval [20].

Calculating Sample Size for Reference Interval Estimation

In calculating the sample size, the desired quantile of reference data (p), the desired quantile of confidence interval (α), and the desired quantile of reference interval (β) should be decided. Quantiles are defined intervals of the data; usually the data is divided into 100 quantiles each with equal number of the values. The reference interval is constructed to include the middle β% of the population. Usually in these calculations, α and β are set equally. After deciding these values, the corresponding z-values of Z_p, $Z_{(1-\alpha/2)}$, and $Z_{(1-B/2)}$ should be used to calculate the sample size. Another parameter of the formula is the relative margin of error (Δ) which is the percentage ratio of the width of the confidence interval for the reference limit to the width of the reference limit. Ideally, this margin of error should be small (i.e., the width of the CI for the reference interval limits should be small compared to the reference interval width), and usually the Δ is set at 10%. The formula for sample size calculation is as follows:

$$n \geq \frac{z^2_{\left(1-\frac{\alpha}{2}\right)}\left(D + \frac{z^2_p}{2}\right)}{z^2_{\left(1-\frac{\beta}{2}\right)}\left(\frac{\Delta}{100}\right)^2} \tag{2.17}$$

where (n) is the sample size and D is a constant which is equal to 1 if there are no subgroups in the sample (i.e., the same reference interval is used for all patients). If the test and the reference interval are dependent on a covariate, then the D value is

determined based on the nature of the covariate; with a uniformly distributed sample, D is 4. For normally distributed covariates, D is 5. For instances where the covariate can be grouped into three groups, D is 5/2.

For example, if you want to determine the reference interval for sodium concentration, consider that you need 80th quantile of the range included in the reference range (P), and you want the alpha to be 0.05 and beta to be 0.20, and then you can calculate the sample size using the above equation (z-scores for 0.9725, 0.95, and 0.80 are approximately 1.9, 1.64, and 0.84, respectively) [21].

$$n \geq \frac{z^2_{\left(1-\frac{0.5}{2}\right)} + (1 + 0.84^2/2)}{z^2_{\left(1-\frac{0.2}{2}\right)} \times 0.1^2} \cong 156 \tag{2.18}$$

Diagnostic Accuracy and Testing for Accuracy

Diagnostic accuracy refers to a test's discrimination power that allows it to identify presence of a disease or condition in an individual. Different measures such as sensitivity and specificity are considered as proxies for diagnostic accuracy. It is important to know that diagnostic accuracy measures cover different aspects, and depending on the clinical question, a specific set of measures should be used. Furthermore, these measures are dynamic and can change per different parameters mainly population characteristics; for example, disease prevalence can affect diagnostic accuracy. Diagnostic accuracy also suffers from the "gold standard" problem, where inaccuracies in the gold standard test can confound interpretation of the diagnostic accuracy studies. Diagnostic accuracy studies usually lack statistical power or fail to follow standard procedures further complicating the issue of diagnostic accuracy. Nonetheless, clinical utility of diagnostic tests is dependent on the diagnostic accuracy of the tests. Here we will address different indicators and measures of diagnostic accuracy.

It is important to know that some accuracy measures are more concerned with discriminative power and the ability of the test to discriminate between the diseased and healthy states. Other measures are more concerned with probability estimation and provide a likelihood of diseased state based on the test result. The most important discriminative measures are sensitivity and specificity. The most commonly used probabilistic indices are positive and negative predictive values and likelihood ratio. These latter measures are highly sensitive to disease prevalence (pretest probability (see Chap. 3)). Sensitivity and specificity, however, are not affected by disease prevalence and can be carried over to different populations.

Sensitivity and Specificity

Diagnostic tests, ideally, should be able to correctly set diseased individuals apart from healthy individuals; each diagnostic test should have a discrimination power that allows for such distinction. For tests that are binary, with a distinct positive or negative outcome, the measurement of this discrimination power is straightforward with positive outcome identifying disease state (true positive) and a negative test outcome highlighting a disease-free state (true negative). For tests that return a range of values, cutoffs should be determined that will distinguish the healthy from diseased. In the most ideal setting, the results of the test will not misdiagnose an individual. However, there is always an overlap of test outcomes between healthy and diseased individuals leading to incorrect assignment of a health state to individuals. If a healthy individual is labeled as diseased by error, this is called a "false-positive outcome." On the other hand, if a diseased individual is misdiagnosed as healthy based on the test outcome, this is called a "false-negative outcome." These four outcomes can be displayed in a 2×2 contingency table (shown in Table 2.1).

Sensitivity is a measure that shows the proportion of individuals with a positive test outcome who are correctly determined to be diseased. In other words, sensitivity is the proportion of "true positives" to all diseased individuals:

$$\text{Sensitivity } (\%) = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \tag{2.19}$$

Sensitivity is usually expressed as a percentage. Sensitivity is a measure of the diagnostic test's ability to screen for a condition. Increases in the sensitivity essentially mean that the number of "false negatives" has decreased. A test with a high sensitivity will identify a significant proportion of diseased individuals. A sensitive test can thus be used to screen individuals with a condition as any "negative" test outcome is more likely to be a "true negative." Alternatively, it can be stated that sensitive tests can be used to *rule out* the disease or condition of interest.

Specificity in contrast is a measure that determines the proportion of "true negative" individuals who are correctly determined as not having the disease, i.e., the proportion of "true negatives" to all individuals without the condition.

Table 2.1 2×2 contingency table showing the outcomes of the test in columns and the disease condition status in rows

	Test outcome	
	Positive	Negative
Condition positive	True positive (TP)	False negative (FN)
Condition negative	False positive (FP)	True negative (TN)

$$\text{Specificity } (\%) = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100 \tag{2.20}$$

Specific tests are useful for *ruling in* individuals with the condition or disease of interest. As specificity increases, the proportion of healthy individuals with a "false-positive" test outcome decreases which means that a "positive" test outcome is likely to be a "true positive."

Highly specific tests are used as confirmatory tests in a two-step diagnostic model: the first step is to employ a population-based screening test with high sensitivity followed by a highly specific test to confirm the diagnosis in individuals with a positive screening. This two-step model is preferred as tests are unlikely to be both very sensitive and very specific. Furthermore, tests with high sensitivity tend to be more affordable than highly specific tests and are thus more suited for population-level utilization. An example of the two-model is alkaloid testing where a primary test (screening) is performed using Marquis reagent spot test and the positive test results are confirmed by gas chromatography (confirmatory test). The Marquis test is fast, affordable, and easy to perform; furthermore, it has high sensitivity; all of these characteristics make it an ideal screening tool. Gas chromatography is a cumbersome and expensive test with high specificity which makes it a good confirmatory tool. Another example is the application of VDRL and RPR test for screening of syphilis infection followed by a confirmatory FTA-ABS or TP-PA test.

There usually exists a tradeoff between sensitivity and specificity. An ideal test will have a sensitivity and specificity of 100%, but such a level of accuracy is unattainable due to multiple factors including the Bayes error rate which states that there is always an irreducible error inherent in any measurement. As a test gains in sensitivity, it tends to lose specificity and vice versa. This tradeoff between sensitivity and specificity will be explained more as part of receiver operating curves (see below).

Overall accuracy is another measure that is useful and can be extracted from Table 2.1. Accuracy is a summative measure that shows the ratio of overall correct calls by the test to all the measurements made. Accuracy is one of the measures of agreement used in validating a diagnostic test and will be covered in Chap. 11.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2.21}$$

Predictive Values

Predictive values refer to the probability of having or not having the condition of interest based on the outcome of the test. Two predictive values can be extracted from the 2 × 2 table. First is the "positive predictive value" (PPV) which measures the probability of having the condition of interest (TP) in individuals with a positive test outcome (TP + FP).

$$\text{Positive predictive value (PPV)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100 \tag{2.22}$$

"Negative predictive value" (NPV) is the probability of not having the condition of interest (TN) in individuals with a negative test outcome (TN + FN).

$$\text{Negative predictive value (NPV)} = \frac{\text{TN}}{\text{TN} + \text{FN}} \times 100 \tag{2.23}$$

Predictive values are more useful clinical measures than sensitivity and specificity as they can directly provide an estimation to the clinician of the likelihood of their patient having or not having a condition based on a positive or negative test outcome. Predictive values unlike sensitivity and specificity are affected by the disease prevalence in the population and as such cannot be transferred from a population to population. The effect of disease prevalence on PPV and NPV is different; as the prevalence increases, the PPV increases (because the probability of having a false-positive result decreases), while NPV decreases. If the prevalence decreases, the reverse will be true; NPV will increase and PPV will decrease. The effect of prevalence is more significant on PPV than on NPV. For diseases with a low prevalence, a test with high specificity (low false-positive rate) is needed to have an acceptable positive predictive value (Fig. 2.7).

Other accuracy measures can also be extracted. A summary of these measures is shown in the table below (Table 2.2).

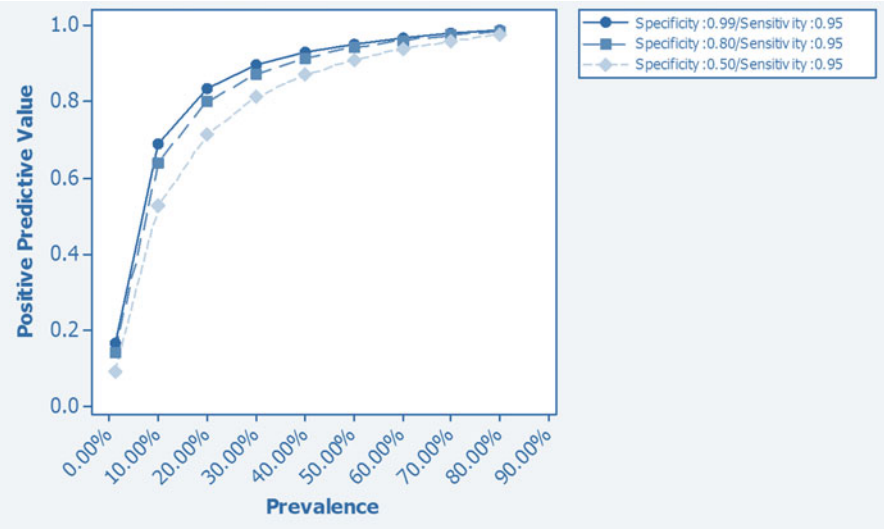


Fig. 2.7 Scatter plot of prevalence versus positive predictive value (PPV) for different test specificities. Note the marked decrease in PPV as prevalence falls below 10%

Table 2.2 Summary of diagnostic accuracy measures

	Test outcome			
	Positive	Positive		
Condition positive	*TP*	*FN*	Sensitivity = TP/(TP + FN)	False-negative rate (FNR) = FN/TP
Condition negative	*FP*	*TN*	False-positive rate (FPR) = FP/TN	Specificity = TN/(TN + FP)
Accuracy = (TP + TN)/(TP + TN + FP + FN)	Positive predictive value (PPV) = TP/(TP + FP)	False omission rate (FOR) = FN/(TN + FN)	Positive likelihood ratio (LR+) = sensitivity/FPR	Diagnostic odds ratio (DOR) = (TP × TN)/(FP × FN) = LR+/LR−
	False discovery rate (FDR) = FP/(TP + FP)	Negative predictive value (NPV) = TN/(TN + FN)	Negative likelihood ratio (LR−) = FNR/specificity	

Receiver Operating Characteristic Curve

The basic assumption for every diagnostic test is that the diseased individuals will have different test outcomes compared to the unaffected population. Many tests return a quantitative range of values instead of a binary "positive" or "negative" value. In these tests, cutoff values must be determined that will set apart the affected from unaffected. Determining cutoff values will depend on the distribution of the values among unaffected and diseased individuals as well as the desired sensitivity and specificity levels. In a perfect test, the outcome values for the affected and unaffected population will have no overlap. There is, however, always a degree of overlap between the two populations making the decision of a cutoff value very important as different cutoff values will lead to different sensitivity and specificity levels (Fig. 2.8).

Receiver operating characteristic curve (ROC) is the graphical illustration of true positive rate (sensitivity) as the *Y*-axis and false-positive rate (1-specificity) as the *X*-axis. ROC curve is generated by plotting the cumulative distribution of sensitivity as a function of cumulative distribution of false-positive rate. Consequently, the ROC curve shows the trade-off between sensitivity and specificity. The basic concept behind the ROC curve is that a reference or test variable (test outcome) is used to classify subjects and classification performance is compared with a classifier variable (gold standard), and at different cutoff values for the test variable, true positive rate and false-positive rate are calculated and plotted as Y-axis and X-axis, respectively.

The ROC curve is usually plotted using a nonparametric generalized linear model. The most common approach was proposed by Tosteson and Begg. In the simplest form of their model, the only classifier is the indicator of the true disease status (χ_1). The other assumption will be that for test outcome values of r_1 through r_j, the subject is classified as negative (T-), and for values greater than r_{j+1}, the subject is classified as positive (T+). Subsequently, for the cumulative probabilities

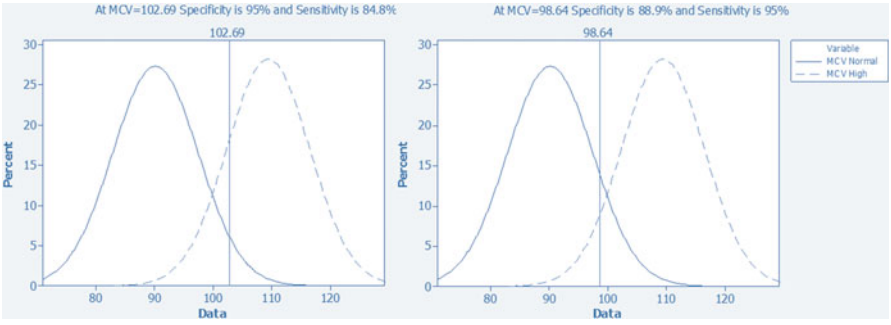


Fig. 2.8 Distribution of mean corpuscular volume (MCV) of normal population and macrocytic anemia (with an assumed 50% prevalence). If the cutoff is set at 102.69 fL, the specificity will be 95% and sensitivity will be 84.8%. Lowering the cutoff will increase the sensitivity and reduce the specificity (at 98.64 fL the sensitivity and specificity will be 95% and 88.9%, respectively)

of response, $\gamma_j(\chi_1)$, determine the response categories (TP, TN, FP, and FN). In this setting, $\gamma_j(0)$ is the probability that an unaffected individual has in a test value outcome of between r_1 and r_j (i.e., test outcome value lower than the cutoff value). This probability represents the "true negative rate" or "specificity" and consequently 1 - $\gamma_j(0)$ represents the "false-positive rate" which forms the *X*-axis of the ROC space. $\gamma_j(1)$ will be the probability that an affected individual has in a test outcome value of between r_1 and r_j (false-negative rate), and thus 1 - $\gamma_j(1)$ will be the "true positive rate" or "sensitivity" which forms the *Y*-axis of the ROC space. The ROC curve will be constructed by plotting all the pairs of 1 - $\gamma_j(0)$ and 1 - $\gamma_j(1)$ for each of the test outcome cutoff points (θ_j). The following generalized linear model will form the ROC curve:

$$g\left[\gamma_j(\chi)\right] = \frac{\theta_j - \alpha'\chi}{\exp(\beta'\chi)} \quad j = 1, \ldots, j-1 \tag{2.24}$$

α' and β' are regression parameters of location and scale, respectively. These two parameters will determine the shape of the curve and in the simplest form are defined as constants that will provide the curve a concave appearance. To make the curve smooth, smoother functions known as link functions are introduced to the generalized linear model. The most common link function used is the probit link which is based on the standard normal cumulative function, Φ. The generalized linear model with the link function applied will be:

$$\Phi^{-1}\left[\gamma_j(\chi)\right] = \frac{\theta_j - \alpha'\chi}{\exp(\beta'\chi)} \quad j = 1, \ldots, j-1 \tag{2.25}$$

ROC curve has been shown to have a nonparametric interpretation like the Mann-Whitney *U* test (see Chap. 6). This means that, while the distribution of test values in the affected and unaffected population usually follows a binormal distribution, other non-normal distributions can also be used in constructing a ROC curve.

ROC space is a square with *X*- and *Y*-axis range of 0–1. A diagonal line connects the top right corner of the space to the bottom left corner. This line is called the line of no discrimination and depicts a complete random association of the test variable with the classifier. The perfect classification point in the ROC space (100% specificity and sensitivity) lies at the top left corner of the space. The further the ROC curve moves away from the diagonal line toward the top left corner, the better the classification properties of the test variables will be (Fig. 2.9).

For determination of the cutoff value, two approaches can be undertaken. In the first approach, a decision must be made on the optimal level of sensitivity and specificity on the ROC curve, and the cutoff value extracted from the table of curve coordinates which shows the corresponding test value for each curve coordinate. This manual search for the cutoff value allows for choices such as choosing a cutoff for screening (high sensitivity) or for confirmation (high specificity) (Fig. 2.10).

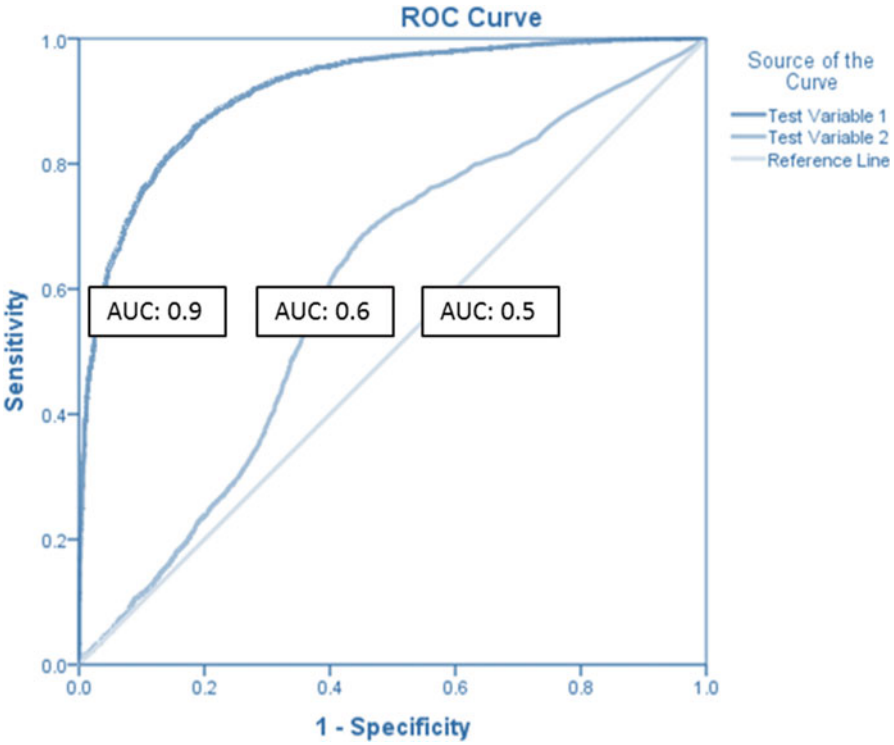


Fig. 2.9 ROC curves for two test variables are depicted. As the curve nears the *top left* corner of the ROC space, the classifying power of the test variable increases. In this figure, note that the test variable 1 is a far better classifier compared to test variable 2

If sensitivity and specificity are given equal weights, then a ROC curve analysis can be employed to determine the optimal cutoff value. Several methods of ROC curve analysis have been established. One of the oldest and simplest methods is called the Youden's index. This index is calculated as the difference of the sum of sensitivity and specificity from 1.

$$\text{Youden}'\text{s index} = (\text{Sensitivity} + \text{Specificity}) - 1 \tag{2.26}$$

Youden's index can assume values between 0 and 1 with 0 showing poor diagnostic accuracy and 1 showing a perfect diagnostic accuracy (sensitivity and specificity of 100%). In ROC curve analysis using the Youden's index, the index is calculated for every coordinate on the ROC curve, and the corresponding test value for the point of maximum Youden's index is then set as cutoff value. The cutoff value set in this approach balances the sensitivity and specificity. Essentially Youden's index determines the cutoff to be at the point where the two distribution of test outcome values of the affected and unaffected population meet. Financial considerations and cost can also be criterions for determining cutoff values and can be incorporated in ROC curve analysis.

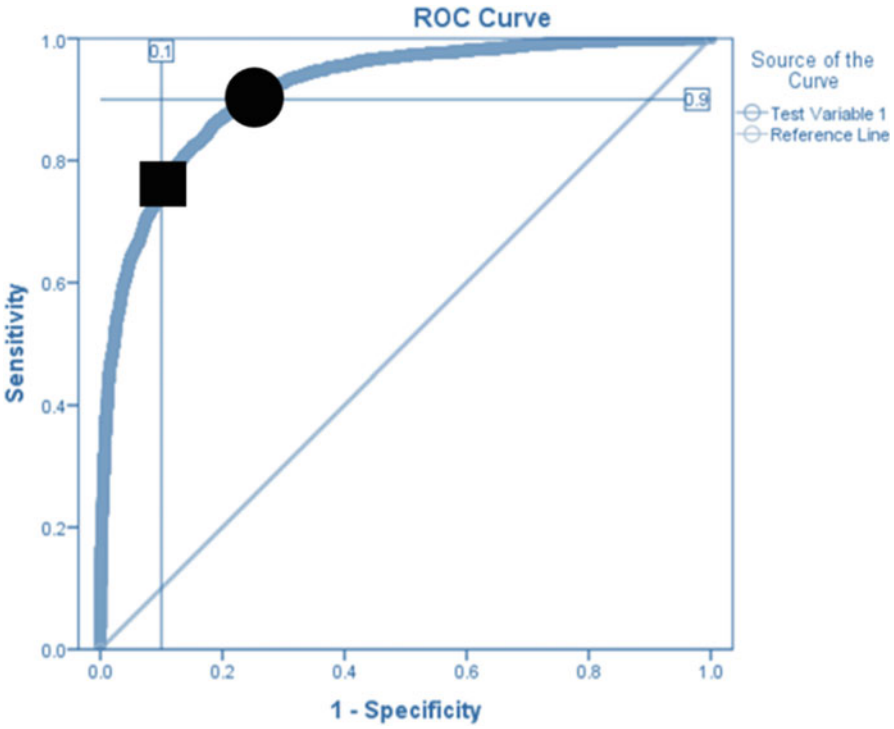


Fig. 2.10 Different cutoff values can be chosen based on the desired level of specificity and sensitivity. The square depicts a point where specificity is 90%. The circle depicts a point where sensitivity is 90%

One of the benefits of ROC curves is the ability to calculate "area under the curve" (AUC). AUC is one of the most useful measures of diagnostic accuracy and discrimination power of a test. A perfect AUC will have a value of 1; this will signify that there exists a cutoff point in test outcome values where the sensitivity and specificity will be 100%. Consequently, as AUC nears 1, the classification (discrimination) power of the test will increase. AUC can be stated in form of the probability that a randomly selected affected individual will have a higher test outcome value than randomly selected unaffected individual.

A rough estimate of levels of discrimination power based on AUC is provided in the Table 2.3 [22, 23].

Calculating AUC

The simplest approach for calculating AUC is using the trapezoidal rule. In this approach, the space under the curve is transformed to a series of rectangles and triangles, and their cumulative area is calculated. If a single cutoff point (j) is used,

Table 2.3 Levels of discrimination power based on AUC

Area under curve (AUC)	Discrimination power
0.9–1	Excellent
0.8–0.9	Very good
0.7–0.8	Good
0.6–0.7	Acceptable
0.5–0.6	Bad
<0.5	Not useful

then the AUC calculated using the trapezoidal method will equal to ½ (Sensitivity$_j$ + Specificity$_j$). This method will always underestimate the true AUC.

Several nonparametric approaches can be used for better estimation of AUC. One of the methods suggested by Hanley and McNeil is called the "Wilcoxon area estimate." In this method, due to inherent similarity between U statistics of a Mann-Whitney test and AUC, the area under curve is calculated using a rank-sum Mann-Whitney U test.

$$\mathrm{AUC} = \frac{n_o n_1 - U}{n_o n_1} \tag{2.27}$$

where n_0 and n_1 represent the sample size of the unaffected and the affected populations. In this calculation, the U statistics is calculated using the rank sum of the unaffected population (R):

$$U = R - \frac{1}{2}\, n_o(n_o + 1) \tag{2.28}$$

The standard error of the AUC estimation by the Hanley method can also be calculated.

$$\mathrm{SE(AUC)} = \sqrt{\frac{\mathrm{AUC}\,(1-\mathrm{AUC}) + (n_1 - 1)\left(Q1 - \mathrm{AUC}^2\right) + (n_o - 1)\left(Q2 - \mathrm{AUC}^2\right)}{n_o n_1}} \tag{2.29}$$

$$Q1 = \frac{\mathrm{AUC}}{(2 - \mathrm{AUC})} \tag{2.30}$$

$$Q2 = \frac{2\mathrm{AUC}^2}{(1 + \mathrm{AUC})} \tag{2.31}$$

While AUC is a relatively simple accuracy measure, recently there have been arguments against using it as a measure of classification power. This is because AUC is a summary measure that includes both relevant and irrelevant parts of a curve; the performance of the test at the extremes of the curve (where specificity will be very high, but sensitivity will be very low or vice versa) is usually not of

interest to the clinicians. Furthermore, AUC gives equal weight to sensitivity and specificity and may not be useful in instances where one of the measures is of greater importance.

Clinical Applicability

Establishing technical accuracy is the first step in appraisal of new tests. The next step will be assessment of diagnostic accuracy. Yet, perhaps even before this step, it is necessary to consider the clinical context in which the test will be used. Most new tests will have a similar test or diagnostic method in the clinical pathways or decision-making algorithms. There are exceptions to this rule, mainly when new screening tests are developed. Thus, an effort must be made to determine the pathway or decision-making algorithm to which the new test belongs. After the pathway is determined, we should identify the possible role of the test in the clinical pathway; a test can be used to screen for or diagnose a disease, it can also be used to guide treatment choices, or it can provide prognostic information. Some tests will also have community or population level indications; for example, they can show the genetic predisposition of the patient's offspring or can determine infectious disease carrier status of a patient.

A new test will either be upstream of a clinical pathway or have a role in "triaging" patients; it may replace an existing test in the clinical pathway or be an add-on to the existing diagnostic pathway. The decision of where the new test will be in regard to the clinical pathway will determine the characteristics and quality metrics of the test. Tests can also have non-diagnostic applications such as monitoring and prognostication.

If a test is to be a substitute for an existing test, it needs to improve upon one or few of the existing test's qualities such as accuracy, cost, harm, ease of performance, etc. Thus, to establish the superiority of the new test (or non-inferiority when factors such as cost or harm are improved in the new test), comparative studies should be conducted to gauge the new test against the existing test. Triage tests or screening tests need to be noninvasive, easy to perform, and cheap; they also need to have high sensitivity. Again, these tests need to be evaluated using clinical trials in order to establish their diagnostic accuracy. Add-on tests will help to further categorize patients in clinical pathways or determine prognosis or treatment options. These tests require higher specificity and are usually time and resource intensive to perform. Currently, a consistent proportion of new tests are focusing on add-on tests as the market for add-on tests is more targeted with less direct competition.

Ideally, the patient outcomes and quality metrics of new diagnostic tests should be measured using well-designed blinded randomized clinical trials. Other trial designs such as controlled trials, before-after studies, and prospective cohorts are also of limited use in estimating the impact of new diagnostic tests. Another option is to determine the effects of the new test on patient outcome via assessment of changes in physicians' intentions for treatment and management. However, given

the current fast pace of innovation, in certain circumstances "modeling" can be used to estimate the impact of the test. We will explore the diagnostic studies in Chap. 12 and data modeling in Chap. 15.

In assessing the clinical benefit of the tests, it is important to identify objective patient outcome measures that are affected by the diagnostic test. In many situations, finding these outcomes is problematic as a direct causative link between diagnosis and patient outcome may be lacking. The effects of a test may not just be physical but also emotional, behavioral, cognitive, or social. Furthermore, a multitude of confounding factors can obscure the true impact of the test. Lack of a targetable outcome with possible treatment options is serious argument against a new test; for example, tests that identify Alzheimer's disease in very early stages are of limited or no clinical use as currently there are no viable treatment options available to the patients.

Secondly, possible trade-offs of utilizing the new test should be identified and balanced. The new test may be associated with direct harm due to invasiveness of the test or administration of possibly toxic or hazardous elements to the patient. Sometimes, the harm can be secondary, for example, a screening test that has good sensitivity yet poor specificity may lead to high false-positive rates which can be subjected to harmful tests or treatments as a follow-up to the screening test. As with benefits, the harm of the test should be identified and measured.

Clinical utility of the test will be determined by comparing the benefit with the harm. While there are objective methods of weighing the benefit versus harm, sometimes a subjective judgment by consensus panels of experts is needed to decide the utility of new tests.

One of the ways to assess the clinical utility of a test is to determine the absolute difference (ΔP) between pretest and posttest probabilities (see Chap. 3). This difference is dependent on test characteristics such as likelihood ratio as well as the pretest probability: as the pretest probability decreases (e.g., prevalence decreases), the likelihood ratio of the test should increase.

$$\text{Absolute difference } (\Delta P) = |\text{pretest probability} - \text{post} - \text{test probability}| \tag{2.32}$$

Absolute probability difference is sometimes difficult to interpret; in reality, the utility of a clinical test is to allow clinicians to alter the care and management of a patient thus providing benefit and avoiding harm. Further calculations are needed to extract the net benefit of a test:

$$\text{Net benefit} = (\Delta P \times r_i \times (b_i - h_i)) - h_t \tag{2.33}$$

where r_i is the rate of changes in the interventions based on probability changes (e.g., to follow curative treatment versus palliative treatment), b_i is the benefit of the changes in the interventions to the patient, h_i is the harm of the changes in the interventions to the patient, and h_t is the harm associated with the test itself. Cost can also be considered in this formula.

The question of clinical utility is usually addressed and assessed by governance and supervisory entities (such as FDA or CLIA) or the manufactures. For the practicing pathologist, it is often more important to be able to evaluate these studies and advisories (see Chap. 13) and decide on the issue of clinical applicability or relevance in their own practice setting. Clinical relevance is determined by answering two questions: transferability and feasibility.

Transferability

Studies for determining the diagnostic and technical accuracy of tests are usually performed in controlled setting with limitation on patient population, biases, and confounding factors. The study setting, as result, can potentially considerably different from clinical setting. Clinical settings can also vary across geographical or community spectra. Nonetheless, the pathologist needs to determine whether test metrics can be transferable to his/her own practice setting.

One of the determinants of transferability is the patient spectrum. It is important that the spectrum of patients seen in the pathologist's practice match or be close to the spectrum of trial subjects. Sometimes, trials upon which diagnostic accuracy is established by manufactures suffer from "spectrum bias" where highly selected patients with controlled parameters are included in the study. Outcome of such studies may not be generalizable, and adopting those tests to a different practice setting with a different population metrics may be problematic. In these setting, before a test is adopted, test validation is required, where the performance of the test is assessed in a representative sample of the population and the results are compared with the test developers' study results. Existence of gold standard tests can help in validating a new diagnostic test (see Chap. 11).

Sensitivity and specificity are thought to be independent of disease prevalence, yet it has been established that in highly controlled study samples with stringent inclusion and exclusion criteria, the calculated specificity and sensitivity may differ from clinical practice, especially in early adaptation settings of a new test due to "indication creep," whereby clinicians order the test for increasingly broad indications, the case-mix, and definition of affected individual changes.

Transferability is also an issue when the test is to be used in a different role than originally anticipated. For example, EGFR status is tested in stage IV lung adenocarcinomas, and the test validation has been performed for that setting. If EGFR status is tested in all lung cancer patients irrespective of stage or type, then a shift in the role has occurred. While, in theory, there can be justification for this transfer, the evidence to support it is lacking.

Another issue in transferability is regarding cross platform and technology transfer. As assays, platforms, technologies, and even test version are changed, there may be a need for revalidation of the test unless there is enough evidence to support the cross validation of these factors.

Feasibility

Determining feasibility of performing a test depends on the practice setting. The needs of the population served and the resources at the disposal of the lab as well as the requirements of the test determine the feasibility of adopting a new test or platform. In smaller setting, answers to these questions can be easier to find, yet in large practice settings, often, a feasibility study is needed.

A feasibility study needs to answer the following questions:

- Will the pathologists, clinicians, patients, and technicians accept the new test? (Acceptability)
- Will there be enough demand for the new test to justify the capital investment, retraining of staff, and the recurring costs? (Demand)
- Will performing the test be possible in the current setting with the available resources? (Implementation)
- Will the test be practical in the practice setting? Is the level of complexity or cost acceptable in the practice setting? (Practicality)
- Can the test be adapted and changed to fit the lab and the population served by the lab? (Adaptability)
- Can the test be integrated into the current lab routines and systems? (Integration)

Feasibility studies need stake holder analysis with participation of the lab director, clinicians, and technicians. Need assessment may sometimes be needed if the demand for a new test is not clear-cut; this assessment is usually in form of practice surveys of the clinicians. Cost analysis and breakdown of cost burden of the new test will also be necessary in evaluating the feasibility of adopting the test. Finally, small-scale runs or pilots can be helpful in better understanding the feasibility of adopting the new test or platform.

It must be noted that clinical utility is a continuous and ongoing issue which needs to be periodically revisited. As the clinical practice and technologies evolve, it may be necessary for the lab to adapt and change, and this requires constant review of the current tests and possible expansions or innovations that can improve, surpass, or replace the current tests [24].

Cost-Effectiveness Analysis

When adopting a new diagnostic test, the final yet perhaps one of the most critical questions will be the cost. The lab directors are eventually responsible for the financial state of the lab and need to decide if adopting the new test will be financially viable. In a purely financially driven setting, outcomes are often ignored, and the entire enterprise is set up to minimize cost and maximize profit. However, in most laboratories, improved patient outcome is the ultimate goal, and thus cost should be considered alongside effectiveness or benefit; if an improved outcome is attainable even at a higher cost, this may justify the cost.

Cost can be broken down to capital and recurrent costs; capital is the financial resources dedicated to procure the equipment, and the recurring costs are the financial resources required for continued operation of the equipment and running of the test (including reagents, controls, and labor). Capital cost should always be weighed against discounting; due to inflation, the true value of an investment made today will be discounted in the future; thus, in calculating investment returns, capital adjustment is needed. The costs should be summarized in form of a per-test cost: the financial resources needed to run a single test. In calculations of cost-effectiveness, we will use the per-test cost as the measure of cost.

Often a test will replace an existing test. In these settings, the per-test costs of the two tests will be directly compared, or, alternatively, a measure called "incremental cost" can be used which is the difference in the per-test cost of the new test versus the old test.

Measuring benefit can be more difficult than measuring cost; crude measures such as life expectancy or changes in mortality can be used, but these measures don't encompass all relevant aspects of patient outcome. Subjective measures such as visual analogue scales can also be used, but they often vary considerably from patient to patient and can be difficult to interpret. Summary measures such as "disability-adjusted life years" (DALY) or "quality-adjusted life years" (QALY) are better suited for measuring effectiveness and are more relevant to patient outcomes.

In measuring both cost and benefit (outcome), the "perspective" should also be considered: whether the cost and outcome are measured at lab level, institution level, or social level. This is a fundamental decision that can make tests that appear cost-ineffective at lab level highly cost-effective at social level (e.g., using advanced nuclear amplification detection methods for screening of tuberculosis instead of smears). Further discussion of the measuring units of costs and effectiveness is beyond the scope of this book.

In comparing a new test (N) versus an existing test (O) and knowing the costs and effectiveness of each test, then cost-effectiveness can simply be stated as

$$\begin{aligned}&\text{Incremental cost} - \text{effectiveness ratio}\\ &= \frac{\text{Cost}_N - \text{Cost}_O\ (\Delta\ \text{Cost})}{\text{Effectiveness}_N - \text{Effectiveness}_O\ (\Delta\ \text{effectiveness})}\end{aligned} \tag{2.34}$$

A perfect test should have a negative "Δ cost" and a positive "Δ effectiveness." A test in which the effectiveness decreases and the cost increases will also be automatically rejected. In cases where the changes in cost and effectiveness are contradictory, the decision will be based on core values of the lab: cost minimization versus patient outcome maximization. One way of measuring effectiveness is to use "posterior odds" (see Chap. 3). Posterior odds are products of prior odds (calculated using disease prevalence) and likelihood ratio.

$$\text{Posterior odds} = \text{Prior odds} \times \text{Likelihood ratio} \tag{2.35}$$

Alternatively, in the formula for incremental cost-effectiveness ratio (ICER), the net benefit of the test (see above) can be used in place of Δ *effectiveness*.

"Decision analytical model" will often provide more insight into cost-effectiveness analysis (Fig. 2.11). This model follows two steps, with the first step involving calculating the "hypothetical performance and cost" of the new test, and if the new test passes this hypothetical step, then a clinical trial is undertaken to calculate the actual cost considerations of the new test. If actual test diagnostic accuracy data and cost estimations are available, then the first step can be skipped, and the cost-effectiveness is determined using the second step.

For example, assume that we are constructing the decision analytical model for a disease with a prevalence of 0.01. The sensitivity and specificity of the current test (*T*0) are 80% and 85%, respectively, and the sensitivity and specificity of the new test (*T*1) are 90% and 90%, respectively. The cost of the current test is 5$ and the cost of the new test is 20$.

In this example, we are employing a health system perspective, and we calculate the costs as the total cost burden of the health system. Based on this, a true positive result will cost 5000$ (early treatment) plus test cost and lead to a DALY of 0.1. A false-negative result will lead to a cost of 10,000$ (late treatment) plus test cost and lead to a DALY of 10. A true negative result will lead to only the test cost and a DALY of 0. The false-positive result will lead to a cost of 5000$ (early treatment) plus test cost and lead to a DALY of 0.1.

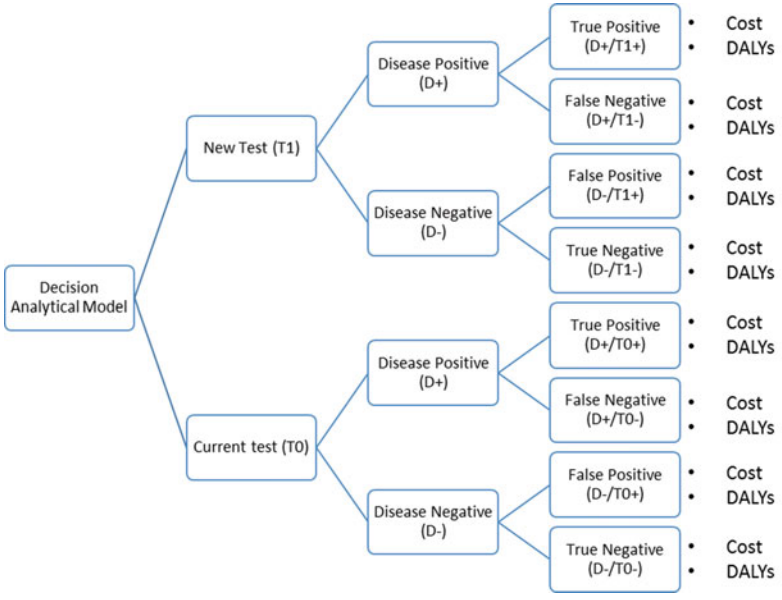
Now we can construct the model (Fig. 2.12). As you can see, despite the higher cost of the new test, at health system level, employing this test will lead to both cost saving and reduction of average DALYs. The results of this cost-effectiveness analysis support the adoption of the new test in place of the current test [25–29].

Summary

We have shown that to assess a diagnostic test, a stepwise approach is needed. The first step of the assessment is technical assessment. In this step, scientific and technical issues related to the test are evaluated, and possible sources of error are identified and addressed. It is important that pathologists know technical test parameters especially precision and accuracy. The pathologist should be able to evaluate the scientific merit of the body of evidence supporting a new test. We will discuss this at length in Chap. 12, where we provide an approach for critical appraisal of literature.

The next important question is to determine the diagnostic accuracy of the test, in other words, to determine if the test can measure what it was designed to measure and whether it has enough discrimination power to be clinically relevant.

This is followed by a closely related step, in which the clinical applicability of the test is assessed: the clinical benefit of the test should be determined and the pathologists needs to assess whether the test is transferable to his/her setting and if so, whether it is feasible to implement the new test or not.



$$Mean\ Cost = \sum(Costs \times Probabilities)$$

$$Mean\ Effectiveness = \sum(DALYs \times Probabilities)$$

$$Incremental\ Cost\ Effectiveness\ Ratio = \frac{Mean\ Cost_{T1} - Mean\ Cost_{T0}}{Mean\ Effectiveness_{T1} - Mean\ Effectiveness_{T0}}$$

Fig. 2.11 Decision analytical model for a new test

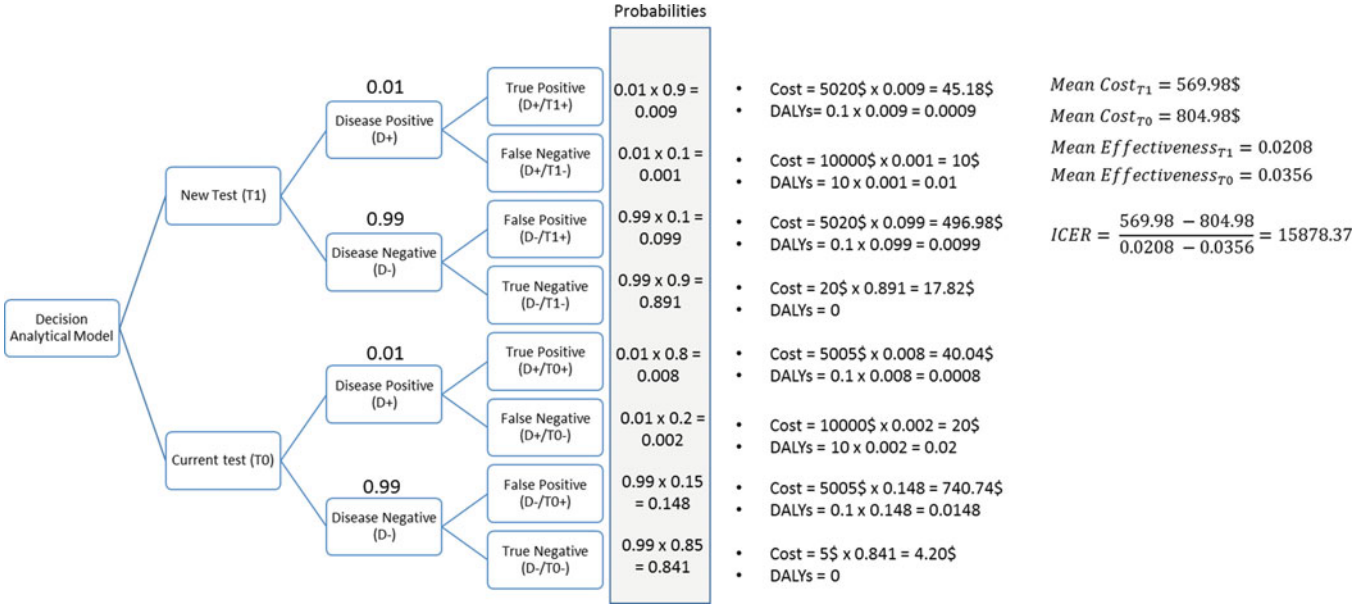


Fig. 2.12 Decision analytical model for the example provided

The final yet critical assessment is to assess the cost of adopting the new test and whether the incremental cost over existing tests is justifiable. Cost-effectiveness analysis is a power tool with which every lab director should be familiar.

We will revisit some of the concept introduced in this chapter further in the book, specifically, in Chap. 10 where we talk about test validation.

References

1. McPherson RA, Pincus MR. Henry's clinical diagnosis and management by laboratory methods. Elsevier Health Sciences: USA; 2016.
2. Crook M. Clinical governance and pathology. J Clin Pathol. 2002;55(3):177–9.
3. Van den Bruel A, Cleemput I, Aertgeerts B, Ramaekers D, Buntinx F. The evaluation of diagnostic tests: evidence on technical and diagnostic accuracy, impact on patient outcome and cost-effectiveness is needed. J Clin Epidemiol. 2007;60(11):1116–22.
4. Garfield S, Polisena J, Spinner DS, Postulka A, Lu CY, Tiwana SK, Faulkner E, Poulios N, Zah V, Longacre M. Health technology assessment for molecular diagnostics: practices, challenges, and recommendations from the medical devices and diagnostics Special Interest Group. Value Health. 2016;19(5):577–87.
5. Koffijberg H, van Zaane B, Moons KG. From accuracy to patient outcome and cost-effectiveness evaluations of diagnostic tests and biomarkers: an exemplary modelling study. BMC Med Res Methodol. 2013;13(1):12.
6. Knottnerus JA, van Weel C, Muris JW. Evaluation of diagnostic procedures. Br Med J. 2002;324(7335):477.
7. Banoo S, Bell D, Bossuyt P, Herring A, Mabey D, Poole F, Smith PG, Sriram N, Wongsrichanalai C, Linke R, O'brien R. Evaluation of diagnostic tests for infectious diseases: general principles. Nat Rev Microbiol. 2008;8:S16–28.
8. Swets JA. Measuring the accuracy of diagnostic systems. Science. 1988;240(4857):1285.
9. Bonini P, Plebani M, Ceriotti F, Rubboli F. Errors in laboratory medicine. Clin Chem. 2002;48(5):691–8.
10. Acken JM, Millman SD. Fault model evolution for diagnosis: accuracy vs. precision. In: Proceedings of the custom integrated circuits conference; 1992. p. 13–4.
11. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D, De Vet HC, Lijmer JG. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem. 2003;49(1):7–18.
12. Jennings L, Van Deerlin VM, Gulley ML. Recommended principles and practices for validating clinical molecular pathology tests. Arch Pathol Lab Med. 2009;133(5):743–55.
13. Šimundić AM. Measures of diagnostic accuracy: basic definitions. EJIFCC. 2009;19(4):203.
14. Yuoh C, Elghetany MT, Petersen JR, Mohammad A, Okorodudu AO. Accuracy and precision of point-of-care testing for glucose and prothrombin time at the critical care units. Clin Chim Acta. 2001;307(1):119–23.
15. Sirota RL. Error and error reduction in pathology. Arch Pathol Lab Med. 2005;129(10):1228–33.
16. Zarbo RJ, Meier FA, Raab SS. Error detection in anatomic pathology. Arch Pathol Lab Med. 2005;129(10):1237–45.
17. Plebani M. The detection and prevention of errors in laboratory medicine. Ann Clin Biochem. 2010;47(2):101–10.
18. Bossuyt PM, Irwig L, Craig J, Glasziou P. Diagnosis: comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ. 2006;6:1089–92.
19. Apple FS, Jesse RL, Newby LK, Wu AH, Christenson RH. National Academy of Clinical Biochemistry and IFCC Committee for Standardization of Markers of Cardiac Damage

Laboratory Medicine Practice Guidelines: analytical issues for biochemical markers of acute coronary syndromes. Circulation. 2007;115(13):e352–5.
20. Katayev A, Balciza C, Seccombe DW. Establishing reference intervals for clinical laboratory test results. Am J Clin Pathol. 2010;133(2):180–6.
21. Bellera CA, Hanley JA. A method is presented to plan the required sample size when estimating regression-based reference limits. J Clin Epidemiol. 2007;60(6):610–5.
22. Tosteson AN, Begg CB. A general regression methodology for ROC curve estimation. Med Decis Making. 1988;8(3):204–15.
23. Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian J Intern Med. 2013;4(2):627.
24. Bowen DJ, Kreuter M, Spring B, Cofta-Woerpel L, Linnan L, Weiner D, Bakken S, Kaplan CP, Squiers L, Fabrizio C, Fernandez M. How we design feasibility studies. Am J Prev Med. 2009;36(5):452–7.
25. Trikalinos TA, Siebert U, Lau J. Decision-analytic modeling to evaluate benefits and harms of medical tests: uses and limitations. Med Decis Making. 2009;29(5):E22–9.
26. Siegel JE, Weinstein MC, Russell LB, Gold MR. Recommendations for reporting cost-effectiveness analyses. JAMA. 1996;276(16):1339–41.
27. Mushlin AI, Ruchlin HS, Callahan MA. Costeffectiveness of diagnostic tests. Lancet. 2001;358(9290):1353–5.
28. Brunstein J. Cost-effectiveness considerations with molecular diagnostics in oncology. MLO Med Lab Obs. 2016;48(5):30.
29. Gray AM, Clarke PM, Wolstenholme JL, Wordsworth S. Applied methods of cost-effectiveness analysis in healthcare. Oxford: OUP; 2010.

Probability and Probability Distribution 3

Introduction

In this chapter, we are going to show that the theory of probability and statistics underlies all quantitative assessments relevant to pathology and laboratory medicine. The probability theory gives rise to different forms of probability function that characterize different physical processes. The probability that an event will occur is a function of a parameter. For example, the probability that the sodium content of a serum sample will be 140 mEq/L will be given by a Gaussian probability curve where the parameter is distribution of normal sodium concentration in the population.

In fact, the overwhelming probability distributions that are used in pathology and laboratory medicine follow the Gaussian (normal) distribution, and thorough understanding of this concept is needed. In this chapter, we have explained normal distribution; however, we have also described the other relevant probability distributions and explained some of the basic concepts of the theory of probability for interested readers. These concepts, however, will be fundamental and require a degree of understanding of mathematical annotation. You may skip this chapter if you are more interested in practical aspects of statistics in pathology.

Probability is a concept that occurs due to randomness; a random event is an event whose outcome cannot be predicted with certainty before the event occurs. Understanding probability also needs a second assumption which is the event or experiment can be repeatable either through time or space; in other words, the event can occur multiple times (indefinitely) under the same conditions (e.g., tossing a die multiple times or tossing multiple dice at the same time). This assumption is the foundation of classical probability theory and allows us to form the probability space.

A random event or experiment has two main components of interest: the first is the "outcome" which is the result of the event that is being recorded. The second is "parameter" which is a constant in the experiment which can affect the outcome.

In biology, most systems are chaotic, in that there is a multitude of parameters and initial conditions affecting each single event which makes determining the

A. Momeni et al., *Introduction to Statistical Methods in Pathology*,
DOI 10.1007/978-3-319-60543-2_3

outcome with certainty difficult. In other words, most events in biology are random and thus are governed by laws of probability, and, consequently, laboratory medicine where our concern is to measure these biological events is also governed by randomness and probability.

Consider the following question:

How many individuals have a hemoglobin level of between 10 and 15 g/dl?

One way to answer this question is to measure hemoglobin levels on every person in the population; this, of course, is practically impossible. The probability theory can help us answer this question using only a fraction of the population which we call a "sample." By understanding probability, we can select a "random sample" from the population and generalize the results to the population with a small degree of uncertainty. This question and similar questions are vital to pathology and laboratory medicine: The entire concept of diagnostic medicine is to be able to differentiate an affected individual from unaffected individuals, and this requires being able to determine normal and abnormal ranges, determining the distribution of a test outcome and making inferences about the results. These, fundamentally, are answered by employing concepts of probability.

Repeatability is used in "compound experiments" where an experiment (E_j) is repeated "j" times from E_1 to E_j with each experiment being independent of the previous experiments. A compound experiment can be running the same test on the same individual for "*j*" times or running the experiment on a random sample of "*j*" size. This sampling, if sufficiently large, can allow us to draw conclusions about the distribution of the experimental outcome in the population. In simplest form of experiments otherwise known as "Bernoulli trials," the outcome can be binary, i.e., only one of two possible outcomes can occur. If multiple outcomes are possible, then the experiment is a "multinomial trial." [1]

To progress further, first we need to introduce some basic concepts:

Sample Set

"Set" or S is a collection of objects also known as elements. For example, the alphabet is a set of letters that is used for writing:

$$S = \{a, b, c, d, \ldots, z\}, \tag{3.1}$$

A "sample space" for a random event includes all the possible outcomes of the event; this set may have elements other than those that are the possible outcomes of the event. For example, the sample set for hemoglobin measurement can be written as $S_{\text{hgb}} = \{0,\infty\}$ even though the range of values of hemoglobin will be only a finite spectrum of the above set. In a compound experiment of n experiments, the sample set will be "Cartesian product" of all the sets of experiments ($S^n = S_1 x\, S_2 x \ldots x\, S_n$).

Event

"Event" is a subset of the sample set. For example, a pregnancy test has two outcomes, either positive or negative. Thus the set for the pregnancy test is

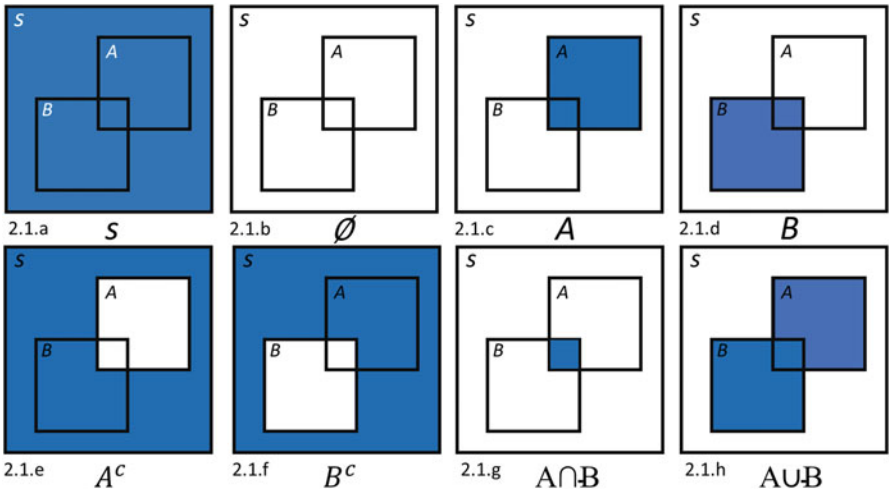


Fig. 3.1 Algebra of events in an experiment with two events: A and B. ∅ is an expression that never occurs as it implies that none of the values in the set are obtained. "∩" or "AND" is an expression that dictates that both events on the two sides of the expression should happen. "∪" or "OR" is an expression that dictates that either of the two sides of the expression should occur

$S = \{Pos, Neg\}$. Positive outcomes are an event or a subset of the test outcomes. In diagnostic test, usually we are interested to determine whether an event occurs or not.

The algebra of events follows a grammar which is shown in Fig. 3.1. Other notable algebraic terms used in probability theory include "A ⊆ B" meaning that occurrence of "A" implies the occurrence of "B" and "A\B" meaning that only "A" occurs and "B" does not occur.

Random Variable

"Random variable" is a variable whose value is determined by a set of random events; it can be expressed as a function that assigns probability to outcome of an experiment. If "s" is the outcome of an experiment with a sample set of "S," then "X" is a random variable which takes the value "$X(s)$":

$$s \in S, \quad X = X(s), \tag{3.2}$$

For example, a urine pregnancy test has two possible outcomes: positive and negative (S). Thus, the result of the urine pregnancy test is a random variable that can only assume values of S, i.e., positive or negative.

Each random variable has a "probability distribution" which is the probability that the value of the variables falls within a certain subset of possible values.

Often, other random variables can be derived from another random variable, i.e., some variables can be a function of another variable. For example, variable "Y" can be derived from random variable "X" using the function "g."

$$s \in S, \quad Y(s) = g[X(s)], \tag{3.3}$$

For example, if troponin levels are a randomly distributed variable, then the probability of a patient suffering from myocardial infarction is another random variable that is derived from the troponin level.

Another form of variables is "indicator variable" which shows if a specific event has occurred. Indicator variables have a value of "1" if the event has occurred and a value of "0" if the event has not occurred. The indicator variable for event "*A*" can be shown as "$1_A(s)$."

$$1_A(s) = \begin{cases} 1, s \in A \\ 0, s \notin A \end{cases} \tag{3.4}$$

Probability Measure and Axioms of Probability

The "probability" of event "*A*" is the set function "*P*" that assigns to each event "*A*" in sample set "*S*" a value "*P*(*A*)." This probability needs to fulfill three "probability axioms" also known as "Kolmogorov axioms":

(a) For every event *A*, $P(A) \geq 0$. That is, event A has a probability between 0 and 1 of occurring.
(b) The probability of sample set is 1, i.e., $P(S) = 1$. That is, if the event A occurs every time the experiment is run, then its probability will be 1.
(c) Given mutually exclusive events (*A*1, *A*2, ...), i.e., [$Ai \cap Aj = \emptyset$, for $i \neq j$], then

$$P(A_1 \cup A_2 \cup \ldots \cup A_i) = P(A_1) + P(A_2) + \cdots + P(A_i) \tag{3.5}$$

$$\text{or} \qquad P\left(\bigcup_{i \in I} A_i\right) = \sum_{i \in I} P(A_i) \tag{3.6}$$

Probability has multiple rules and theorems; some of the basic and fundamental rules are provided below:

1. $P(\varnothing) = 0$, (3.7), i.e., the probability that the result of the experiment is not in the sample set is zero.
2. $P(A^C) = 1 - P(A)$, (3.8), i.e., the probability of event not occurring equals to 1 minus the probability of event A occurring.
3. If A $\subseteq$ B, then P(A) < P(B), (3.9), i.e., if event A only occurs if event B has already occurred, then the probability of event A is smaller than the probability of event B.
4. If $A \subseteq B$, then $P(B \backslash A) = P(B) - P(A)$, (3.10), i.e., if event A only occurs if event B has already occurred, then the probability that event B occurs without event A occurring equals to the difference of the two probabilities.

5. $P(B\backslash A) = P(B) - P(A \cap B)$, (3.11); axiom 4 can be alternatively stated that the probability of event B to occurring without event A occurring equals to the probability of event B minus the probability of the intercepts of event A and B.
6. Product rule – for any two independent events A and B:

$$P(A, B) = P(A) \times \ P(B), \tag{3.12}$$

i.e., the probability of two independent events to occur equals the product of their probabilities (e.g., probability of rolling two dice and getting two sixes equals to probability of rolling one six multiplied by the probability of rolling another six: $1/6 \times 1/6 = 1/36$).

7. Boole's inequality (union bound) provides the upper bound of probability of a union of finite events:

$$\text{If } \{A_i : i \in I\} \text{ is a finite collection of events, then } P\left(\bigcup_i A_i\right) \leq \sum_i P(A_i), \tag{3.13}$$

8. Bonferroni's inequality provides lower bounds of probability of a finite union:

$$\text{If } \{A_i : i \in I\} \text{ is a finite collection of events, then } P\left(\bigcap_i A_i\right) \geq 1 - \sum_i 1 - P(A_i), \tag{3.14}$$

9. Inclusion-exclusion rule – for any events A or B:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B), \tag{3.15}$$

Alternatively, this theorem can be written as

$$P(A \cap B) = P(A) + P(B) - P(A \cup B), \tag{3.16}$$

This rule can be expanded to any number of events:

$$P\left(\bigcup_i A_i\right) = \sum_{k=1}^{n} (-1)^{k-1} \sum_{J \subseteq I, \#(J)=k} P\left(\bigcap_j A_j\right) \tag{3.17}$$

These rules can help answer some of the relevant questions in diagnostic medicine. Here we will provide a couple of genetics examples that use the above theorems:

Example 3.1
Q: Type 1 hemochromatosis is an autosomal recessive hereditary disease that is caused by mutations in the HFE gene. C282Y is one of the common mutations in this gene. If one parent of a child is homozygous for C282Y and the other parent is heterozygous for C282Y, then what is the probability of their child having type 1 hemochromatosis or being heterozygous for the disease?

A: One parent has two alleles with C282Y which means that the probability of passing on this mutation to the offspring is 1. The other parent is heterozygous for the mutation which means that the possibility of passing on this mutation to the offspring is 0.5. Thus, the probability of a homozygous child is 0.5 (product of the two probabilities) and the probability of a heterozygous child is 0.5 as well. The union of these two probabilities is 1.

$$\begin{aligned} P\left(\text{Child}_{\text{Heterozygous or Homozygous}}\right) &= P\left(\text{Child}_{\text{Heterozygous}} \cup \text{Child}_{\text{Homozygous}}\right) \\ &= P\left(\text{Child}_{\text{Heterozygous}}\right) \\ &\quad + P\left(\text{Child}_{\text{Homozygous}}\right) \\ &= 0.5 + 0.5 = 1, \end{aligned} \tag{3.18}$$

Example 3.2
Q: A disease has two causative mutations: A and B. Patients manifest the disease if they have one or both mutations. The probability of a person having mutation A is 0.02 and the probability of a person having mutation B is 0.03 and the probability of having both mutations is 0.01. What is the probability that a person has the disease?

A: This can be answered using the inclusion-exclusion rule:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.02 + 0.03 - 0.01 = 0.04, \tag{3.19}$$

Conditional Probability

Conditional probability relates to the probability of an event occurring if another event has already occurred. For example, instead of asking the probability that a random person develops chronic renal failure, we can ask what is the probability of a random person who has diabetes to develop chronic renal failure. Thus, in conditional probability, a condition (or sets of conditions) needs to be satisfied, before a probability can be assigned to an event. The probability of the event occurring in this setting is proportional to the condition sets; let B be the condition and A the event, we can write conditional probability as $P(A/B)$:

Table 3.1 2×2 confusion matrix of disease A and test B results in 100 individuals

	Test B positive (T+)	Test B negative (T−)	Total
Disease A positive (D+)	40	5	45
Disease A negative (D−)	10	45	55
Total	50	50	100

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \tag{3.20}$$

The condition in conditional probability can itself be a conditional probability measure. For example, for events A, B, and C, A is conditional to B which is conditional to event C:

$$P(A|B \cap C) = \frac{P(A \cap B|C)}{P(B|C)}, \tag{3.21}$$

Conditional probability is very useful in diagnostic medicine: many of the important test measures are conditional probabilities. To better demonstrate this, we will provide an example: Table 3.1 is a 2×2 contingency table that shows the number of individuals who have disease A (D+) in first row and the number of unaffected individuals (D-) in second row. Test B has been developed to test for this disease, and the results of the test are shown in columns with the first column showing the number of individuals with a positive test (T+) and the second column showing the number of individuals who tested negative (T−).

An important question to ask is if a person has the disease A (D+), then what is the probability that he will test positive for test B (T+) or $P(T+|D+)$:

$$P(T+|D+) = \frac{P(T+ \cap D+)}{P(D+)} = \frac{40}{45} \cong 0.89, \tag{3.22}$$

Incidentally, this probability ($P(T+|D+)$) is called "sensitivity" of the test. Other important test conditional probabilities include:

- If a person is unaffected by disease A (D-), then what is the probability that he will test negative for test B (T−) or $P(T-|D-)$? This is known as test "specificity" which is approximately 0.82 in the above example.
- If a person tests positive for test B (T+), then what is the probability that he has disease A (D+) or $P(D+|T+)$? This is known as test "positive predictive value" which is 0.80 in the above example.
- If a person tests negative for test B (T-), then what is the probability that he does not have disease A (D−) or $P(D-|T-)$? This is known as test "negative predictive value" which is 0.90 in the above example.

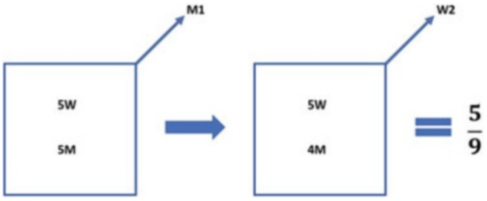
These test measures are explained further in Chap. 2.

Probability axioms also apply to conditional probability. For example, the law of unions can be applied to conditional probabilities; given that $P(B) > 0$ and events $A_1, A_2, \ldots, A_i$*are exclusive*, *then*

$$P(A_1 \cup A_2 \cup \ldots \cup A_i|B) = P(A_1|B) + P(A_2|B) + \cdots + P(A_i|B). \tag{3.23}$$

Example 3.3
Q: If five men and five women are in a group and two people are randomly chosen from the group, then what is the probability of choosing a woman if the first choice has been a man?

A: $P(W_2|M_1)=$



Example 3.4
Q: In 1000 patients, presence of diabetes (*Di*) and proteinuria (*Pr*) is assessed. 150 patients had proteinuria (Pr+), while 200 patients had diabetes (*Di*+). Of the 200 patients that had diabetes, 110 had proteinuria. What is the probability of a person having proteinuria (*Pr*+) without having diabetes (*Di*−)?

$$\text{A} : P(\text{Pr}+|Di-) = \frac{P(\text{Pr}+|Di-)}{P(Di-)} = \frac{{}^{40}/_{1000}}{{}^{800}/_{1000}} = \frac{40}{800} = 0.05, \tag{3.24}$$

Multiplication Rule

Occasionally, conditional probability is the probability that is known; we can use this to calculate other event probabilities.

$$P(A \cap B) = P(B) \times P(A|B) = P(A) \times P(B|A), \tag{3.25}$$

This can be generalized to a sequence of events ($A_1, A_2, \ldots, A_i$) in a random experiment:

$$\begin{aligned} P(A_1 \cap A_2 \cap \ldots \cap A_i) = {} & P(A_1) \times P(A_2|A_1) \times P(A_3|A_1 \cap A_3) \times \ldots \\ & \times P(A_i|A_1 \cap A_2 \cap \ldots \cap A_{i-1}), \end{aligned} \tag{3.26}$$

For example, if we have three events (A, B, C), then

$$P(A \cap B \cap C) = P(A) \times P(B|A) \times P(C|A \cap B), \tag{3.27}$$

Example 3.5
Q: In a standard deck of cards, what is the probability of drawing four consecutive aces from the deck?

$$\text{A} : P(A_1 \cap A_2 \cap A_3 \cap A_4) = {}^{4}/_{52} \times {}^{3}/_{51} \times {}^{2}/_{50} \times {}^{1}/_{49} \approx 0.000003, \tag{3.28}$$

Two events are considered independent, if occurrence of one does not affect the probability of the other. In such cases the multiplication rule can be written as

$$P(A \cap B) = P(A) \times P(B), \tag{3.29}$$

Alternatively, two events are independent if and only if the above formula is correct.

If there are more than two events, then a "pairwise independence" is sought, whereby any pair of events should be independent. For example, for events A, B, and C to be pairwise independent, the following statements must all be true:

$$[P(A \cap B) = P(A) \times P(B)] AND\ [P(A \cap C) = P(A) \times P(C)] \times AND\ [P(B \cap C) = P(B) \times P(C)], \tag{3.30}$$

If in addition to the pairwise independence, the following statement is also true:

$$P(A \cap B \cap C) = P(A) \times P(B) \times P(C), \tag{3.31}$$

then these events are considered as being "mutually exclusive."

Bayesian Probability

Bayes theorem allows us to determine the probability of an event knowing prior conditions or an inverse probability of the event. For example, we can determine $P(B/A)$ if $P(A/B)$ is known. The Bayes theorem borrows from the concepts of conditional probability and can be written as

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}, \tag{3.32}$$

In Bayes theorem, the probability measures that are known and given as input are called "prior probabilities," and the probabilities that are calculated are called "posterior probability." In the next section, we introduce the concepts of pretest and posttest probability which are governed by Bayes theorem.

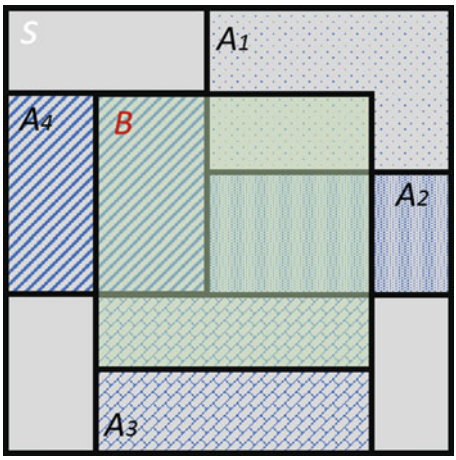


Fig. 3.2 A partitioning of sample space S to finite events A = {Ai: i ∈ I} induces a portioning of event B

The Bayes theorem can be expressed using the law of total probability; in this law probability of event B ($p(B)$) is conditioned on the partition A which is a finite collection of exclusive events with probabilities larger than zero that partition the sample space (S) [$A = \{A_i: i \in I\}$] (Fig. 3.2).

In this setting the law of total probability states that $P(B)$ is the sum of weighted average of the conditional probability of $P(B/A_i)$ with $P(A_i)$ being the weight factors (over $i \in I$):

$$P(B) = \sum_{i \in I} P(A_i)P(B|A_i). \tag{3.33}$$

Thus, the Bayes theorem can be rewritten as

$$P(A_j|B) = \frac{P(A_j)P(B|A_j)}{P(B) = \sum_{i \in I} P(A_i)P(B|A_i)}, \tag{3.34}$$

In this formula, $P(A_j)$ is the prior probability of A_j and $P(A_j/B)$ is the posterior probability of A_j.

We will demonstrate the Bayes theorem further in the following example:

Example 3.6

Q: Going back to the example of proteinuria and diabetes; suppose that the probability of someone having proteinuria (Pr+) is 0.04 and the probability of someone having diabetes (Di+) is 0.3. Now suppose that the probability of a diabetic having proteinuria is 0.1. Now calculate the probability of someone with proteinuria being a diabetic.

A:

$$P(Di+|\mathrm{Pr}+) = \frac{P(\mathrm{Pr}+|Di+)P(Di+)}{P(\mathrm{Pr}+)} = \frac{0.1 \times 0.3}{0.04} = 0.75, \tag{3.35}$$

Bayesian probability is very useful in diagnostic medicine; it allows for calculation of false positive and false negative probabilities if factors such as test sensitivity, specificity, and disease prevalence (prior probability) are known. In the next section, we have elaborated this concept further [2–5].

Pretest and Posttest Probability

When interpreting diagnostic tests, one must exercise caution as there is always an element of error in measurements. As such it rarely occurs that positivity for a test will mean that an individual is affected by a condition with a 100% certainty. Knowing the prior risk (prior probability) of the individual will help interpret the result by allowing us to calculate the posterior probability of the individual.

Prior probability or pretest probability is the probability that the individual being tested has the disease of interest before testing is performed. This probability can be based on disease prevalence or can also consider certain demographic or clinical information about the patient. Posttest probability is the probability of having a disease after a test or set of tests are performed. For example, if a 70-year-old woman tests positive for pregnancy using a highly accurate pregnancy test, it is still highly unlikely that she is pregnant. Measures such as "positive predictive value" and "negative predictive value" are more sensitive to changes in prior probability than sensitivity and specificity.

$$\begin{aligned}\text{Positive predictive value} &= P(D+|T+) = \frac{P(T+|D+)P(D+)}{P(T+)} \\ &= \frac{P(T+|D+)P(D+)}{P(D+)P(T+|D+) + P(D-)(T+|D-)},\end{aligned} \tag{3.36}$$

$$\begin{aligned}\text{Negative predictive value} &= P(D-|T-) \\ &= \frac{P(T-|D-)P(D-)}{P(D+)P(T-|D+) + P(D-)P(T-|D-)},\end{aligned} \tag{3.37}$$

As you can see in the formula above, knowing $P(D+)$ and $P(D-)$ is important in determining whether a positive test means being affected or a negative test means being unaffected. Usually it can be assumed that sensitivity (α) and specificity (β) of tests are stable across populations. Knowing the prior probability of an individual having disease A ($p = P(A)$), then the posterior probability of the individual having disease A based on a positive test ($P = P(A/T+)$) can be calculated:

$$P = \frac{\alpha p}{(\alpha + \beta - 1)p + (1 - \beta)}, \tag{3.38}$$

P is a function of p where if p increases from 0 to 1, then P also continuously increases from 0 to 1. If the sum of sensitivity and specificity is greater than 1, then the function P has a concave downward shape (see ROC curve, Chap. 2). If $P = p$, then $\alpha + \beta = 1$ and the test is only randomly associated with the disease status (independent).

Alternatively, it can be stated that the posttest probabilities can be measured using likelihood ratio; likelihood ratio (LR) is a product of sensitivity and specificity of the test and consequently is less prone to sampling bias or changes in pretest probability. Posttest probability is calculated using the posttest odds which is a product of pretest odds and likelihood ratio. The pretest odds is the ratio of those who have a condition versus those who are unaffected, i.e., the ratio of probability of having a condition (pretest probability (p)) versus not having it.

$$\text{Pretest odds} = \frac{p}{1 - p}, \tag{3.39}$$

Likelihood ratio (LR) can either be positive (LR+) which means the likelihood of a positive test result favors the individual being affected by the condition of interest or it can be negative (LR−) which favors the individual being unaffected. LR+ is the ratio of true positive results to false positive results, and LR− is the ratio of false negative results to true negative results.

$$\text{LR+} = \frac{P(T + |D+)}{P(T + |D-)} = \frac{\text{Sensitivity}}{1 - \text{Specificity}}, \tag{3.40}$$

$$\text{LR-} = \frac{P(T - |D+)}{P(T - |D-)} = \frac{1 - \text{Sensitivity}}{\text{Specificity}}, \tag{3.41}$$

LR+ can have values between 0 and ∞. If LR+ is less than 1, then test positivity decreases the posttest probability. As likelihood ratio nears 1, the test result effect on posttest probability decreases, with a likelihood ratio of 1 denoting that the test has no effect on posttest probability. With LR+ values greater than 1, a positive test outcome has more effect on posterior probability with a LR value of 10 increasing the posttest probability by 45% if the test outcome is positive.

Posterior odds is a product of pretest odds and LR+:

$$\text{Posterior odds} = \text{pretest odds} \times \text{positive likelihood ratio}, \tag{3.42}$$

The posttest probability can be calculated from the posterior odds:

$$\text{Posttest probability} = \frac{\text{Posterior odds}}{\text{Posterior odds} + 1}, \tag{3.43}$$

In tests where a screening test is followed by a confirmatory test, the posttest probability of the screening test will be the pretest probability of the confirmatory test. In these cases, it is important to determine if the two tests are independent and do not have significant overlap. In general, tests of the same modality have considerable overlap and should be avoided as a confirmatory test. In the simplest example, if a test is run for one person twice, the results of the first run should not be interpreted as the posterior probability of the second run.

Example 3.7

Q: Let us assume that prostate-specific antigen (PSA) value of more than 10 ng/ml has a sensitivity of 90% and specificity of 70% for detection of prostatic cancer. If prostate cancer has a prevalence of 0.005 in a 40-year-old man, then what is the posterior probability of a PSA higher than 10 ng/ml in a 40-year-old man?

A: The pretest odds of the patient is approximately 0.005 (0.005/0.995). The LR+ of PSA at this age is 3 (0.9/(1–0.7)). The posterior odds of the patient is 0.015 (0.005×3). The posttest probability of this patient is approximately 0.014 ($0.015/(1 + 0.015)$). Thus, despite the high PSA level in this patient, he is still highly unlikely to have a prostate cancer.

Q: A new array comparative genomic hybridization (CGH) assay has been developed for prostate cancer that has a sensitivity of 99% and specificity of 99%. The test is performed for the above patient after the PSA levels were determined and the test results are positive. What is the posterior probability of the patient having prostate cancer?

A: The pretest odds of the patient is now 0.015. The LR+ of the new test is 99 (0.99/(1–0.99)). The posterior odds of the patient is 1.485 (99×0.015). The posttest probability of patient having prostate cancer is now approximately 0.6.

Fagan nomogram (Fig. 3.3) is a visual estimation of the pretest and posttest probabilities based on the likelihood ratio of the test; using the nomogram a straight line can be drawn from the pretest probability at the left nomogram to the test likelihood ratio. Continuation of this line to the right side of the nomogram will provide the posttest probability at the point where the line crosses the posttest probability scale.

Pathologists should be aware of the concept of pretest and posttest probability and use tests cautiously when approaching cases; tests, however accurate, can lead to misleading diagnoses if pretest probabilities are ignored. Thus, pathologists and especially anatomic pathologists should avoid using shotgun panels and tests (such as ordering a wide panel of immunohistochemical stains) to work up cases. As the example above showed, a test with a specificity and sensitivity of 99% can only change a 1% pretest probability to 40% posttest probability. Therefore, multistep diagnostic approaches or established diagnostic criteria with acceptable statistical power should be employed in diagnosis.

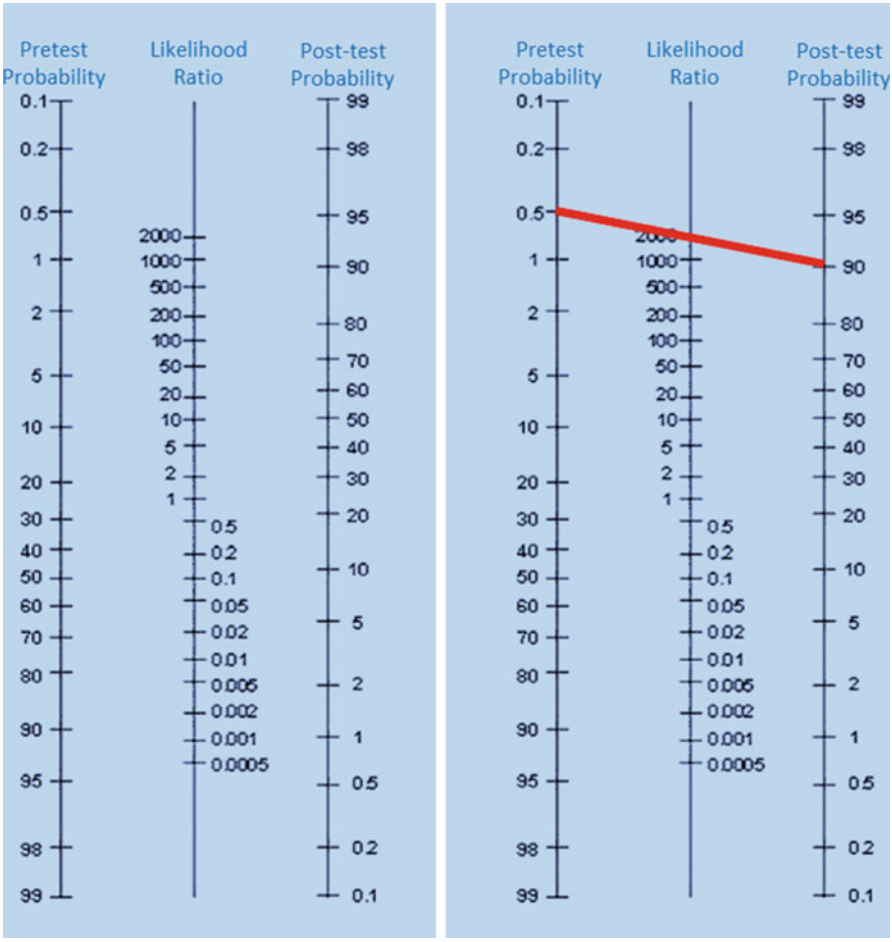


Fig. 3.3 The Fagan nomogram. The panel on the right shows a patient with a pretest probability of 0.5% who had a positive test result with a positive likelihood ratio of 2000 which translates to a posttest probability of 90% [6–10]

Probability Distribution

Probability distribution is the probability of occurrence of different possible outcomes of a random trial or experiment. Probability distribution can be discrete, for example, "binomial distribution" or "Poisson distribution." Probability distribution can also be continuous, for example, "normal distribution." Understanding probability distribution is fundamental to understanding statistics. Many of the applications of statistics in pathology and laboratory medicine require an understanding of probability distribution and its different forms. Concepts such as

"confidence interval," "mean," "reference range," "error," etc. are all determined using probability distribution.

While what follows may appear exhaustive, we encourage the readers to at least familiarize themselves with concepts of "probability mass function," "cumulative distribution function," "probability density function," "normal distribution," "log-normal distribution," "mean," and "variance." At the end of the chapter, we have provided an introduction of the different plots used to depict probability distributions.

Discrete Distribution

A discrete random variable is a countable finite or infinite random variable; let us assume that we call a discrete random variable X, and then the possible ranges of X is a countable set ($\{R_x = x_1, x_2, x_3, \ldots\}$). Now assume that event A is the set of outcomes (s) in sample space S for which the value of X is equal to x_j.

$$A = \{s \in S | X(s) = x_j\}, \tag{3.44}$$

We can show the probabilities of the events where variable X assumes the value x_j ($\{X = x_j\}$) as the "probability mass function" (PMF) of X. PMF is also known as "probability distribution" for discrete random variables.

$$P_X(x_j) = \begin{cases} P(X = x_j), \text{for } j = 1, 2, 3, \ldots \\ 0 \qquad\qquad\qquad\qquad \textit{if } x_j \notin R_x \end{cases}, \tag{3.45}$$

PMF is a function that provides the probabilities of different possible values of a discrete random variable. Thus, the theorems and laws of probability also apply to PMF. For example, the sum of all probabilities of X will be 1.

$$\sum_{x \in A} P_x(x) = 1. \tag{3.46}$$

Also:

$$0 \leq P_x(x) \leq 1 \text{ for all } x \text{ and,} \tag{3.47}$$

$$\text{for any set } A \subset R_x, P(X \in A) = \sum_{x \in A} P_x(x). \tag{3.48}$$

Example 3.8
Q: Two consecutive qualitative HIV tests are run; if X is the number of positive results, define the PMF for X.

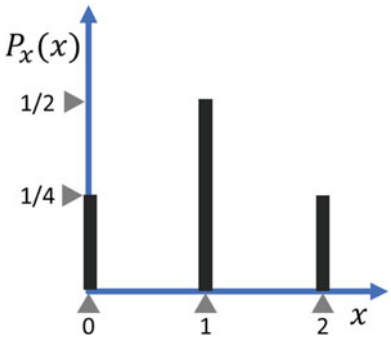


Fig. 3.4 PMF plot of an experiment consisting of a binary event repeated twice (see Example 3.8)

A: Here the sample space will be:

$$S = \{++, +-, -+, --\}, \tag{3.49}$$

The possible ranges of values for X will be:

$$R_x = \{0, 1, 2\}, \tag{3.50}$$

The PMF for X will be:

$$P_x(0) = \frac{1}{4}, \quad P_x(1) = \frac{1}{2}, \quad P_x(2) = \frac{1}{4} \quad , \tag{3.51}$$

The values of the PMF can be plotted (Fig. 3.4). The figure shows that if the experiment is repeated infinitely, then half of the times we will observe the value of X to be 1 ($P_x(1) = 1/2$). This peak can also be observed in the plot.

The random variable X also has a "cumulative distribution function" (CDF) otherwise known as "distribution function of X." CDF is the probability that the random variable X, evaluated at x, takes a value equal or smaller than x.

$$F_X(x) = P(X \leq x) = \sum_{m=0}^{x} f(m) = f(0) + f(1) + f(2) + \cdots + f(x). \tag{3.52}$$

CDF has certain properties:

1. The values of CDF range from 0 to 1.
2. CDF is a nondecreasing function of x, for $-\infty < x < \infty$.
3. The probability of X taking a value between x_1 and x_2 is a cumulative function as well:

$$P(x_1 < X < x_2) = F_x(b) - F_x(a), \tag{3.53}$$

CDF is more useful for evaluating the distribution of random continuous variables. The cumulative distribution function of a continuous random variable is equal to the area under the curve of the "probability density function":

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt. \tag{3.54}$$

Example 3.9
Q: Considering the PMF of Example 3.8, what is the CDF of $X \leq 1$?

A: For $X \leq 1$, it means that X can either have a value of 0 or 1. Then the CDF can be calculated as

$$F_X(x \leq 1) = P_x(0) + P_x(1) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}, \tag{3.55}$$

Let us assume that we have a finite population with a size of N that contains a number of a certain type (m) and the remainder of another type ($N - m$). "Hypergeometric distribution" then is the probability mass function of drawing a number of items of type m from the population *without replacement* (i.e., with each successive draw, the population decreases) (x).

$$P(X = x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}, \tag{3.56}$$

The term $\binom{m}{x}$ is called a binomial coefficient which states the number of ways for picking x unordered outcomes from m possibilities and is read as "m choose x." The value of the binomial coefficient can be calculated with the following formula:

$$\binom{m}{x} = \frac{m!}{(m-x)!x!}\ , \tag{3.57}$$

Example 3.10
Q: There are 10 patients in a ward and 3 of them have contracted a hospital-acquired infection. Your staff randomly picks 5 patients for a blood culture. Let X be the number of infected patients selected. What is the PMF of the X variable?

A: In this example, the size of population (N) is 10, the number of draws is 5 (n), and the number of successful draws is (x) which can have values of {0, 1, 2, 3}. Thus, the PMF is

$$P(X = x) = \begin{cases} \dfrac{\binom{3}{x}\binom{10-3}{5-x}}{\binom{10}{5}} & x = 0, 1, 2, 3 \\ 0 & \text{Otherwise} \end{cases}, \tag{3.58}$$

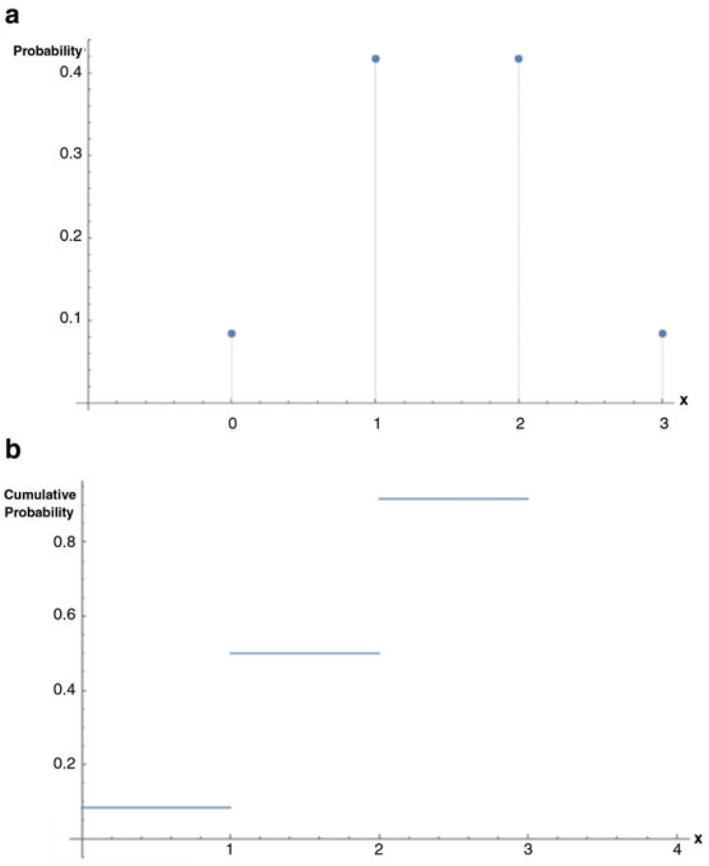


Fig. 3.5 PDF (**a**) and CDF (**b**) plots of Example 3.10

Q: What is the probability of drawing exactly three infected patients?
A:

$$P(X=3)=\frac{\binom{3}{3}\binom{7}{2}}{\binom{10}{5}}=\frac{1}{12}, \tag{3.59}$$

The PMF and CDF functionsof the above example can be plotted (Fig. 3.5) [11].

Binomial Distribution

"Binomial random variables" are discrete random variables that count the number of successes in a fixed number of trials. In each trial, the choices are binary: there will either be success (the event of choice occurs) or failure; these trials are otherwise known as "Bernoulli trials." These trials are independent, and there will be replacement meaning that the probability of success is the same for each trial. Each binomial random variable (X) has a probability mass function:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \tag{3.60}$$

where p is the probability of success in each trial, n is the number of trials, and x is the number of successes.

Each binomial variable has a "binomial distribution":

$$X \tilde{} b(n, p), \tag{3.61}$$

n and p are the parameters of the distribution of X. If p is 0.5 (i.e., half of the trials are successes), then the distribution of the binomial variable will be "symmetrical" (Fig. 3.6a). If p is small, the successes are less likely to occur, and the distribution will be "skewed right" with the smaller numbers constituting the bulk of the distribution and the distribution tailing off toward larger numbers (this holds true

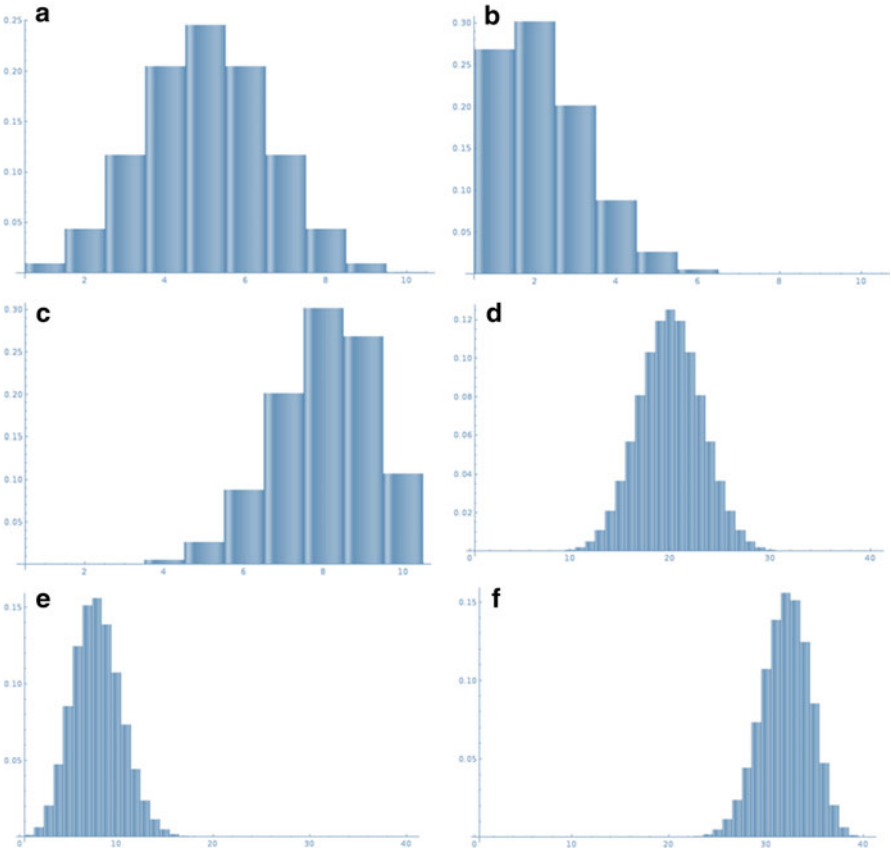


Fig. 3.6 Symmetrical distribution with $p = 0.5$ and $n = 10$ (**a**). Right skewed distribution with $p = 0.2$ and $n = 10$ (**b**). Left skewed distribution with $p = 0.8$ and $n = 10$ (**c**). As the number of trials increases ($n = 40$), the distribution approaches symmetry irrespective of p (frames **d**–**f**)

if the number of trials (n) is small) (Fig. 3.6b). Conversely, if the p is large with small trial size, then the distribution is said to be "skewed left" (Fig. 3.6c). If the number of trials is sufficiently large, then the distribution "approaches symmetry" irrespective of the probability of success (Fig. 3.6d–f).

Example 3.11
Q: You have noticed that in each run of 10 samples in your blood gas machine, one sample result is incorrect ($p = 0.1$). You choose a random sample from each run for 5 runs ($n = 5$), what is the probability of choosing the sample with the wrongly reported result, 3 times out of 5 ($x = 3$)?

A:

$$f(3) = \binom{5}{3} 0.1^3 (0.9)^2 = 0.0081, \tag{3.62}$$

Q: What is the probability of finding the sample at least two times?

A: Here you can calculate the probability by calculating the CDF of $2 \leq x \leq 5$.

$$\begin{aligned} F(2 \leq x \leq 5) &= f(2) + f(3) + f(4) + f(5) = 1 - (f(0) + f(1)) \\ &= 0.08146, \end{aligned} \tag{3.63}$$

In reality, the binomial distribution occurs only rarely in the practice of pathology. However, it is of interest that this distribution can be used to generate all isozyme forms of multi-subunit enzymes such as creatine kinase (CK) and lactate dehydrogenase (LDH). Here, we know that CK has two isozymes termed M and B that exist as dimers. The question is how many distinct dimers can exist? The answer is $(M + B)^2$ or $(M^2 + 2\,MB + B^2)$. Thus, there are three forms. For LDH, there are two isozymes, H and M, that form tetramers. To generate the combinations, we compute $(H + M)^4$ or $H^4 + 4H^3M + 6H^2M^2 + 4HM^3 + M^4$ or a total of five different tetramers. In general, the number of distinct polymer forms with two subunits is $(A + B)^N$ where A and B are the distinct isozymes and N is the number of units per polymer. It is easy to see that the number of distinct forms, each form containing N subunits, is $N + 1$ [12, 13].

Geometric Distribution
Geometric distribution is a discrete probability distribution of the number of "Bernoulli trials" needed to obtain one success (X) or alternatively the number of failures before a success is achieved ($Y = X-1$). Geometric distributions have two parameters: the probability of success (p) and the number of trials until a successful trial (x).

The probability mass function of a geometric distribution of number of trials before a success is attained is written as

$$f(x) = P(X = x) = (1 - p)^{x-1} p, \tag{3.64}$$

Alternatively, the distribution of number of failures before a successful trial can be written as

$$f(x) = P(X = x) = (1-p)^x p, \tag{3.65}$$

The cumulative distribution function of a geometric distribution of number of trials before a success is attained is

$$F(x) = P(X \leq x) = 1 - (1-p)^x, \tag{3.66}$$

And the CDF for the distribution of number of failures before a successful trial is

$$F(x) = P(X \leq x) = 1 - (1-p)^{x+1}, \tag{3.67}$$

Example 3.12
Q: Going back to Example 3.11, what is the probability of going through three runs before finding one sample with an incorrectly reported result?
A:

$$f(x) = P(X = 3) = 0.9^2 \times 0.1 = 0.081, \tag{3.68}$$

In geometric distributions, the number of trials needed before a successful trial is the inverse of the probability of success ($^{1}/_{p}$). Incidentally, $^{1}/_{p}$ is also the mean of a geometrically distributed variable.

This concept is used in two useful epidemiologic measures: "number needed to treat" and "number needed to harm." Simply stated, these measures mean the number of individuals who must receive a treatment or be exposed to a hazard to have one person be cured or to develop an adverse outcome in case of number needed to harm. This follows a geometric distribution and they can be written as

$$\text{Number needed to treat (NNT)} = \frac{1}{\text{Absolute risk reduction}}, \tag{3.69}$$

$$\text{Number needed to harm (NNH)} = \frac{1}{\text{Absolute risk increase}}, \tag{3.70}$$

The absolute risk difference (either reduction or increase) is the difference between the probability of cure (or harm) in the experimental group and the control group.

Negative Binomial Distribution
A "negative binomial" random variable is a binomial variable that denotes "the number of Bernoulli trials choosing the r^{th}- *1* success" or $\binom{x+r-1}{r-1}$ where x is the number of failures and $r-1$ is the number of successful trials with success on the $(x+r)^{\text{th}}$ trial. The negative binomial distribution (also known as "Pascal distribution") can be written as

$$X \tilde{} nb(r,p), \tag{3.71}$$

The probability mass function of a negative binomial distribution is given by

$$f(x) = P_{r,p} = \binom{x+r-1}{r-1} p^r (1-p)^x, \tag{3.72}$$

And the CDF is given by

$$F(x) = \sum_{n=0}^{x} \binom{x+r-1}{r-1} p^r (1-p)^x, \tag{3.73}$$

Example 3.13
Q: Going back to Example 3.11, what is the probability of finding the third wrongfully reported sample (r) in the seventh run ($x + r$)?
A:

$$P_{3,0.1} = \binom{6}{2} 0.1^3 (0.9)^4 = 0.0172187, \tag{3.74}$$

Q: What is the probability of finding the third wrongfully reported sample on or before the seventh run?
A:

$$F(4) = \sum_{n=0}^{4} \binom{x+r-1}{r-1} p^r (1-p)^x = 0.0701908, \tag{3.75}$$

Mean and Variance

Before we can introduce the concept of "Poisson distribution," we need to explain "mean," "variance," and "moment-generating functions" (MGF).

Each random variable (X) has an "expected value" ($E[X]$) which is the weighted average of values that X can take on. For discrete variables, the weights of values are determined by their respective probability.

$$E[x] = \sum_{x} f(x) P(x). \tag{3.76}$$

In other words, the expected value of X can be thought of as the "mean of X." For continuous variables, the expected value is the integral of its probability density function.

$$E[x] = \int f(x)P(x)dx, \tag{3.77}$$

If a random experiment is repeated many times, the average value of outcomes converges on the expected value of the experiment. Thus, the mean is the "central tendency" or "location" of a probability distribution.

The expected values of powers of X are called the "moments of X." The "moments about the mean of X" are expected values of powers of $[X-E[X]]$.

In continuous variables, the zeroth moment (E^0) equals 1 and is the total probability. The first moment (E^1) is the mean of the variable, the second central moment ($(X-E[X])^2$) is the variance, the third central moment ($(X-E[X])^3$) is the skewness, and finally the fourth central moment ($(X-E[X])^4$) is the kurtosis.

Thus, variance (σ^2) is the expected value of the squared difference of the random variable from its mean. Variance shows the dispersion of the probability distribution.

$$\sigma^2 = E\left[(X - E[X])^2\right], \tag{3.78}$$

Moment-generating function is a real function, the derivatives of which at zero are equal to moment of the random variable. The MGF can characterize the distribution of the random variable. The MGF can be given by

$$M_x(t) = E\left[e^{tX}\right], \tag{3.79}$$

where t is all real numbers belonging to the closed interval $[-h,h]$ $[-h, h] \subseteq R$, for which the expected value $E[e^{tX}]$ exists and is finite.

A summary of formulas for mean, variance, and moment-generating function of different distributions is provided in Table 3.2.

Poisson Distribution

Poisson distribution is a probability distribution that gives the probability of a given number of events occurring in a fixed interval (of time or space) if these events are independent and the average rate of the events is known. For example, if X is the

Table 3.2 Characteristics of different random distributions

Distribution	Mean	Variance	Moment-generating function
Bernoulli	$E(x) = p$	$\sigma^2 = p(1-p)$	$1 - p + pe^t$
Binomial	$E(x) = np$	$\sigma^2 = np(1-p)$	$(1 - p + pe^t)^n$
Geometric	$E(x) = \frac{1}{p}$	$\sigma^2 = \frac{1-p}{p^2}$	$\frac{pe^t}{1-(1-p)e^t}$
Negative binomial	$E(x) = \frac{pr}{1-p}$	$\sigma^2 = \frac{pr}{(1-p)^2}$	$\frac{(1-p)^r}{(1-pe^t)^r}$
Poisson	$E(x) = \lambda$	$\sigma^2 = \lambda$	$e^{\lambda(e^t-1)}$
Normal	$E(x) = \mu$	σ^2	$e^{t\mu+\frac{1}{2}\sigma^2t^2}$
Chi-squared	$E(x) = k$	$\sigma^2 = 2k$	$(1-2t)^{-\frac{k}{2}}$

number of tests ordered in a day with the number of tests ordered each day being independent of the number of tests ordered in prior days and if the average number of tests ordered is known, then X is a Poisson random variable.

Other conditions that must be met for a variable to follow Poisson distribution are:

- Two events must be separated (i.e., they cannot occur at the same time).
- The rate of occurrence of events is constant with the probability of events occurring in an interval being proportional to length of the interval.

The probability mass function of a Poisson variable is given by

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \tag{3.80}$$

where λ is the mean and variance of X and e is the mathematical constant with a value of 2.7182 to four decimal points.

The distribution of a Poisson variable is symmetrical and is centered around its mean.

The cumulative distribution function of a Poisson variable can be written as

$$F(x) = e^{-\lambda}\sum_{i=0}^{k}\frac{\lambda^i}{i!}, \tag{3.81}$$

where k is the floor function of x which is the largest integer less than or equal to x.

Example 3.14

Q: If X is the number of tests ordered in a day with the number of tests ordered each day being independent of the number of tests ordered in prior days and if the average number of tests ordered is 10 per day, then what is the probability of having a day in which 8 tests are ordered?

A:

$$f(8) = \frac{e^{-10}10^8}{8!} \cong 0.112599, \tag{3.82}$$

Q: What is the probability of having 8 or less tests per day?

A: (Fig. 3.7)

$$F(8) = e^{-10}\sum_{i=0}^{8}\frac{10^i}{i!} = 0.33282 \tag{3.83}$$

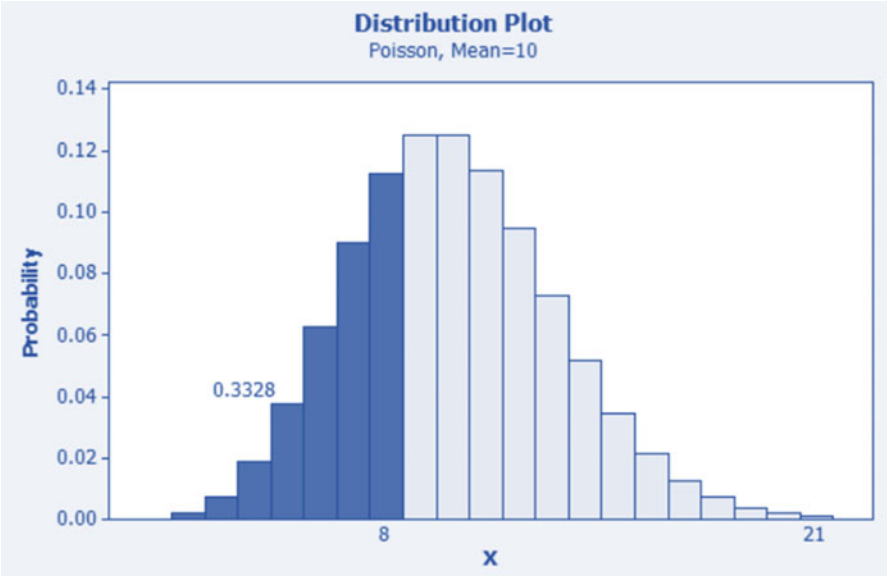


Fig. 3.7 Probability distribution plot of Example 3.14 with the darker area showing the cumulative probability of X assuming values of 8 or less

Continuous Distributions

Unlike discrete random variable where the sample set consists of a range of finite values, the continuous variables have a range that is infinite and uncountable. In these variables, single points among the range have a probability of 0, and only ranges of values can have a non-zero probability (Fig. 3.8).

Probabilities of continuous random variables are derived from the area under curve of its probability distribution plot. Hence, instead of probability mass function, we use "probability distribution function" to characterize continuous variables. Probability density function (PDF) is the integrable function $f(x)$ for continuous variable x which for all x in the sample set, $f(x) > 0$ and the area under curve for the entire range of the sample set are equal to 1. The integral of $f(x)$ for any interval $(a \leq x \leq b)$ in the sample set is the probability of x falling within that interval. PDF can be given as

$$f(x) = P(a \leq x \leq b) = \int_a^b f_X(x)dx. \tag{3.84}$$

The cumulative distribution function of continuous variables is a nondecreasing continuous function (unlike discrete variables where CDF is a nondecreasing step

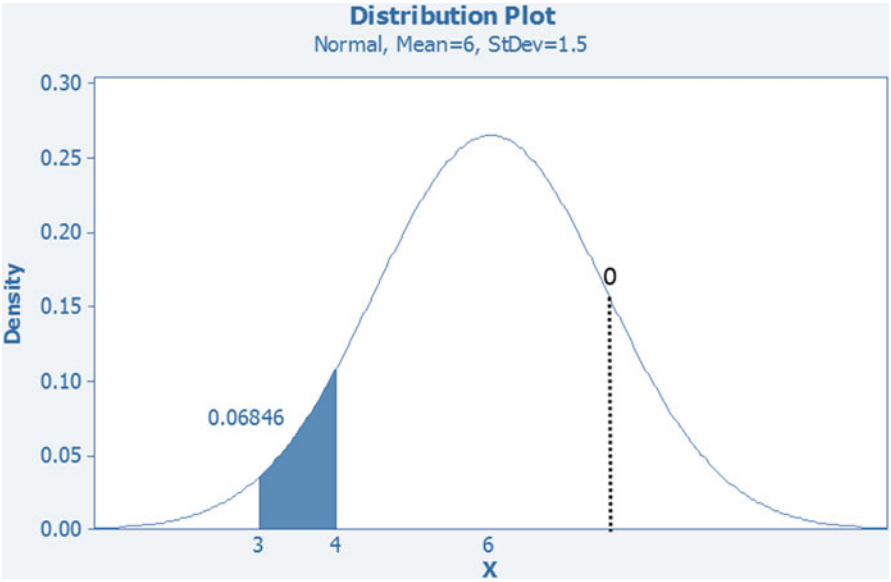


Fig. 3.8 In this normal distribution, the *shaded area* shows a range of values of X and has a non-zero probability (0.06846); the *dotted line*, however, represents a single point in the range of X values and has a probability of 0

function). We introduced the cumulative distribution function of a continuous variable before (Formula (3.54)).

Continuous variables are very important in clinical pathology as most tests are quantitative and return an outcome from a continuous range. One of the important properties of continuous variables is that they can have "percentiles." The concept of percentiles is very important since factors such as "confidence interval" and "reference range" are determined using percentiles. A percentile is a value (π_p) below which a given percentile of observations or outcomes falls. For each π_p, the area under the curve to the left of the value has a probability p which is equal to the percentile (Fig. 3.9). In other words, each percentile signifies a CDF from $-\infty$ to π_p.

The mean (expected value) of a continuous variable is called μ and is calculated as

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx. \tag{3.85}$$

Consequently, the variance (σ^2) of a continuous variable is given by:

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx. \tag{3.86}$$

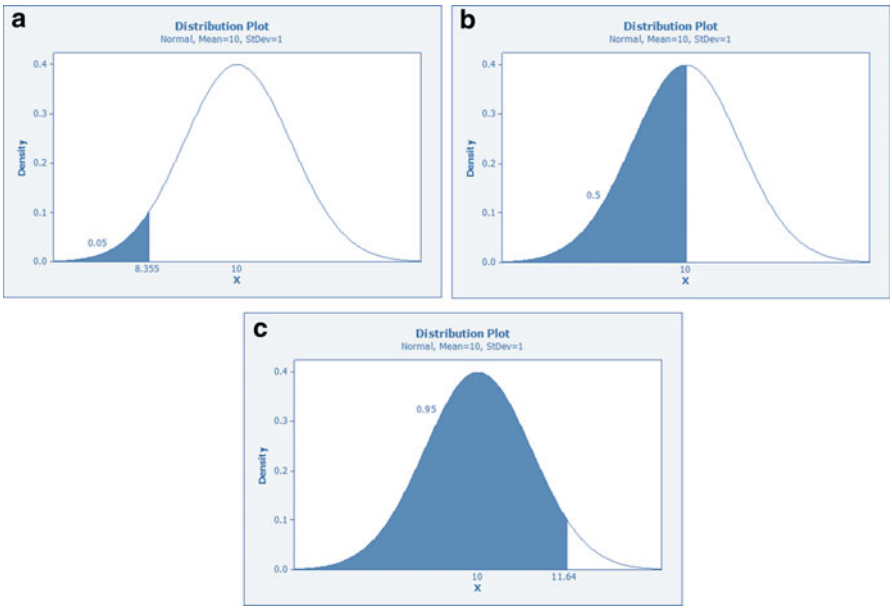


Fig. 3.9 Examples of percentile: 5th (**a**) 50th (**b**) and 95th (**c**) percentiles

Normal Distribution

"Normal distribution" otherwise known as "Gaussian distribution" is perhaps one of the most important probability distributions in laboratory medicine. This crucial role of the Gaussian distribution is due to the "central limit theorem." Simply stated, this theorem postulates that if a measurement (e.g., a diagnostic test result) is influenced by infinite uncertainty sources, then the distribution function of the measurement will approach a normal distribution, irrespective of the probabilities and distributions of the uncertainty sources. In reality, even a finite but sufficiently large number of uncertainty sources will shift the probability distribution toward a normal distribution. In biology and by extension laboratory medicine, each measurement or test is influenced by many sources of uncertainty like age, gender, individual traits, nutritional status, etc. Thus, many of the quantitative test results in laboratory medicine have either a normal distribution or log-normal distribution.

The normal distribution curve has the famous bell-shaped (Fig. 3.9) appearance with probability at μ being the maximum height of curve. It means that normal distribution is perfectly symmetrical around its mean and its moments beyond mean and variance are zero (i.e., there is skewness or kurtosis and so on). In normal distribution, the bulk of the values is within a few standard deviations of the mean, and for practical purposes the probabilities of values falling more than four standard deviations from the mean are practically considered to be zero ($\lim_{x \to \pm\infty} f(x) = 0$) (see Westgard rules, Chap. 10).

The two main parameters of a normal curve are mean and variance: the mean determines the location of the curve and the variance determines the spread of the curve. As the variance increases, the curve becomes flatter or vice versa, as the variance decreases, the curves become taller.

The probability density function of a normal (Gaussian) distribution is given by:

$$f\left(x|\mu,\sigma^2\right) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{3.87}$$

The CDF function of normal distribution is denoted with Φ and is written as

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt, \tag{3.88}$$

Direct calculations of the integral are technically difficult; to facilitate these calculations, each variable with normal distribution ($N(\mu,\sigma^2)$) can be transformed into a standard normal distribution ($N(0,1)$).

If X is the random normal variable with $N(\mu,\sigma^2)$, then the standardized distribution of X is given by:

$$Z = \frac{x-\mu}{\sigma}, \tag{3.89}$$

This standardized distribution now has the distribution $N(0,1)$. Thus, any interval ($a \leq x \leq b$) can be transformed into a Z-value and the corresponding probability calculated using the values of the Z-table (Appendix A). Z-table contains the mathematical values of standardized Φ [14, 15].

Example 3.15

Q: The mean corpuscular volume of red blood cells follows a normal distribution with mean of 92 fL and variance of 16. What is the probability of a normal individual (i.e., no anemia) having a MCV of less than 80 fL?

A: The standardized value of 80 fL can be calculated as

$$Z = \frac{80-92}{4} = -3, \tag{3.90}$$

Going to the Z-table, you can see that there are no negative Z-values. Thus, for negative values, you use the corresponding complementary cumulative Z-value which shows that the probability of a value falling below Z-score of 3 is 0.00135. Thus, a healthy individual has a 0.00135 probability of having a MCV of less than 80 fL.

Alternatively, if a probability is provided, then the corresponding value of x can be calculated by transforming the probability to its corresponding Z-value and then calculating the corresponding x value using the following formula:

$$x = \mu + z\sigma, \tag{3.91}$$

One of the important questions in statistics is whether a continuous variable has a normal distribution or not. We will talk about this is detail in Chap. 6 where we will introduce the concepts of parametric and non-parametric measures as well as testing for normality.

Log-Normal Distribution

In physiologic measurements, many of the measurement values converge to a limiting distribution. In a simplified manner, these measurements are usually bound by a lower limit and exhibit additive percentage changes that are usually unidirectional. For example, for every gender/age group, there is a limit to how low a person's weight can be; on the other hand, the weight of individuals can be very high (Fig. 3.10). In other words, many physiologic processes are right skewed. These continuous measurements will not fit normal distribution, and "log-normal distribution" should be used.

Log-normal distribution is a continuous probability distribution whose logarithm follows a normal distribution (if $X \sim \ln N(\mu, \sigma^2)$ then $\ln(X) \sim N(\mu', \sigma'^2)$). If X is the log-normal random variable with μ and σ being the mean and standard deviation of the variable natural logarithm, then we can express X as

$$X = e^{\mu + \sigma Z}, \tag{3.92}$$

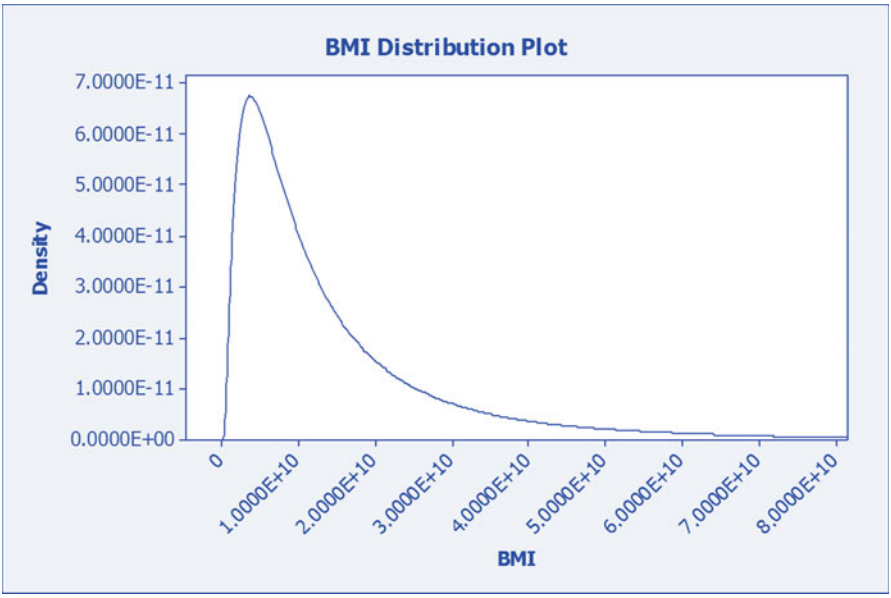


Fig. 3.10 Body mass index (BMI) distribution plot follows a log-normal distribution with a lower threshold of 15, scale of 1, and a location of 23

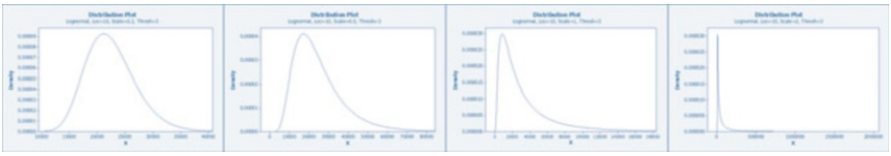


Fig. 3.11 Changes in shape of log-normal curves with changes in σ (scale)

The probability density function of a log-normal variable with a known mean and standard deviation of the variable natural logarithm (μ and σ) is given by:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, \tag{3.93}$$

σ is the main determinant of the shape of the log-normal distribution. As σ increases, the curve shifts toward the left increasing the skewness, and as σ decreases, the curve shifts to the right increasing the symmetry (approaching normal distribution) (Fig. 3.11). μ on the other hand determines the location of the curve, with the peak of the curve corresponding to μ.

The cumulative distribution function of a variable with log-normal distribution is written as

$$F(x) = \Phi\left(\frac{\ln x}{\sigma}\right), \tag{3.94}$$

where Φ is the cumulative distribution function of standard normal distribution [16].

Introduction to Distribution Plots

Distribution plots are graphs that visualize the distribution of a random variable. Examining distribution plots allows for a subjective and rapid assessment of data and can be a powerful tool in data analysis and reporting results. We will introduce some of the distribution plots here. Most of these plots mainly concern visualizing continuous random variables especially normally distributed variables.

Normal Distribution Plot

"Normal distribution plots" or "normal probability plots" are used to assess whether the data follows a normal distribution. This allows for a fast and visual inspection of normality before statistical procedures that are designed for normal distributions can be used. In normal distribution plots, the observed values of the variable are plotted against a theoretical normal distribution; if the variable has a normal distribution, then the points in the plot should approximate a straight line (Fig. 3.12).

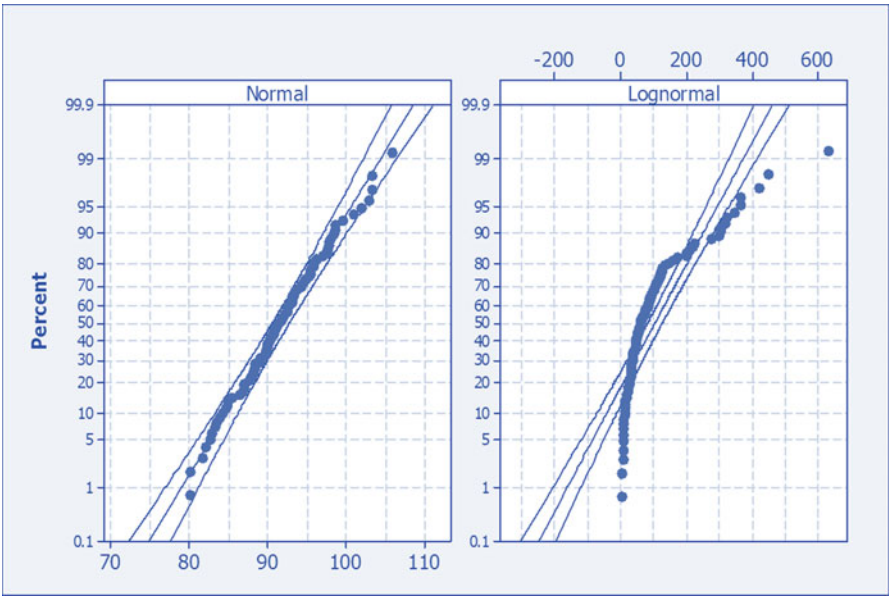


Fig. 3.12 The panel on the left shows a normal distribution; note that the points in the plot approximate a straight line. The panel on the right shows a log-normal distribution; note the departure of the points from the straight line

Quantile-Quantile Plots

"Quantile-quantile plots" also known as Q-Q plots provide a visual estimation of a comparison of two variable distributions. This graph plots the quantile distributions of one variable against quantile distributions of another variable (note quantile distributions and not the actual values). If the points in the Q-Q plot approximate a straight line, then the two variables plotted are likely to be from the same distribution family (e.g., both are normally distributed). Significant departures from a straight line show that the two variables have different distribution families (Fig. 3.13).

Cumulative Distribution Plots

CDF plot plots the cumulative distribution of probabilities of values of a variable less than a given x. The Y-axis of the plot is the cumulative probability and ranges from 0 to 1. The CDF plot for discrete variables consists of a series of steps and for continuous variables is a curve (Fig. 3.14).

Histogram

"Histogram" is a visual representation of the distribution of a variable. The Y-axis of a histogram consists of either frequency or relative frequency in which the frequencies are normalized to a scale (e.g., percents). The X-axis of the plot consists of values of the variable put in "bins." Bins are range of values of the variable that

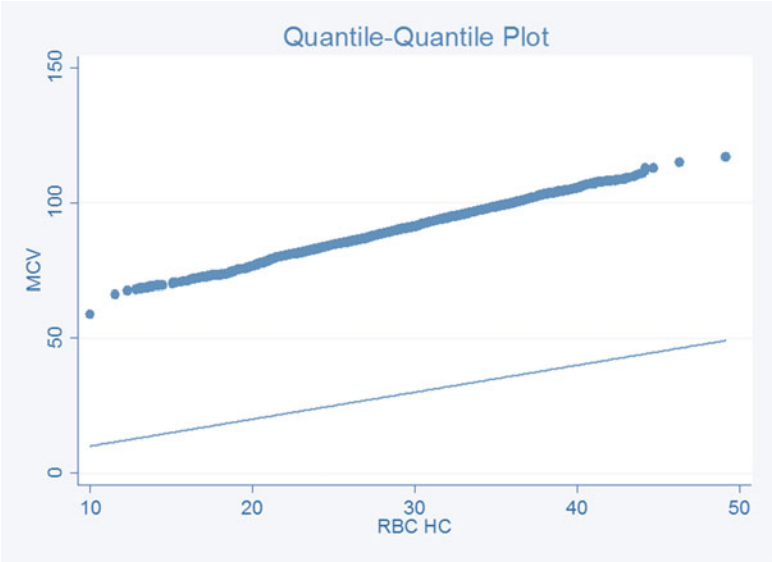


Fig. 3.13 Q-Q plot of MCV versus RBC hemoglobin concentration. Note that the points form a straight line suggesting that both variables have the same distribution family

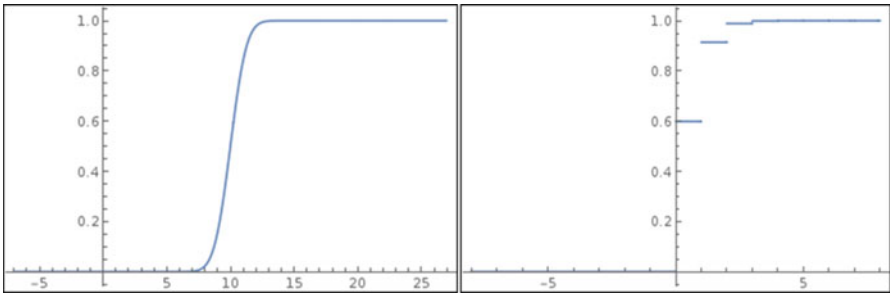


Fig. 3.14 The panel on the *left* shows the CDF plot of a continuous random variable, and the panel on the *right* shows the CDF plot of a discrete random variable

are clustered together; in other words, the X-axis is divided into equal length intervals (bins) (Fig. 3.15).

The choice of bin size is very important in drawing a histogram: too wide intervals will hide crucial variations in data, while too narrow intervals can show too much noise.

Often, you can fit distributions to a histogram. In these instances, a curve best fitting the distribution of the data is added to the histogram. Interpretation of the fitted distribution curve is often easier than interpreting the histogram itself.

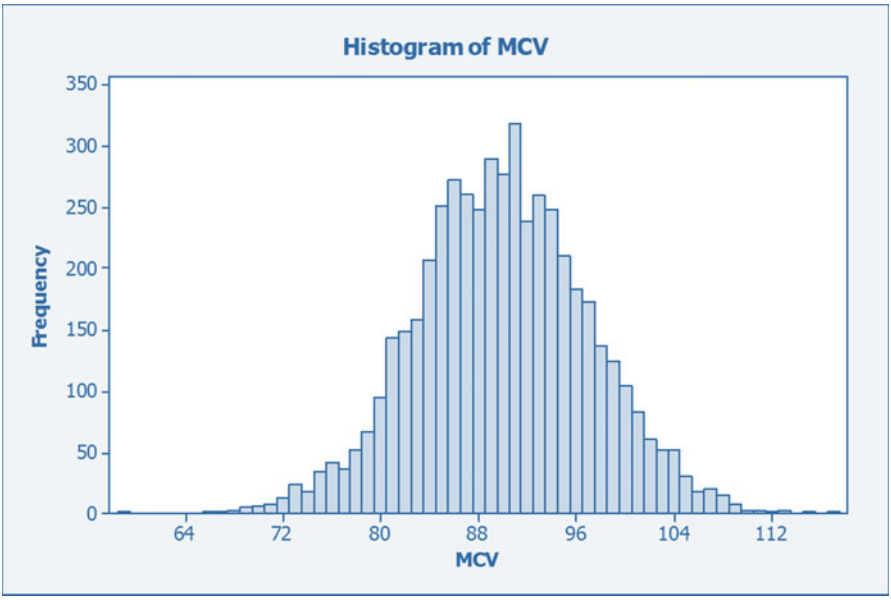


Fig. 3.15 Histogram of mean corpuscular volume

Boxplot

Each randomly distributed variable can be summarized in a five-number summary which is composed of the median, first quartile, third quartile, and minimum and maximum of the data range. Boxplot graphs allow visualization of these five-number summaries and are powerful tools in comparative exploratory analysis. Boxplots consist of a rectangle, the upper and lower bounds of which mark the third and first quartile, respectively. The median is marked by a horizontal line through the rectangle. The minimum and maximum are shown as linear extensions from the rectangle. Outliers can be marked as points or circles along the axis of the plot (Fig. 3.16) [17].

Summary

In this chapter, we reviewed two fundamental concepts of statistics and statistical inference: probability and distribution. Many of the concepts that we will introduce in later chapters require an understanding of these fundamentals.

A set of rules or theorems govern probability. Probability has its own alphabet; different elements can define each random variable such as sample set, event rate, probability mass function, and cumulative distribution function.

It is imperative that the readers understand the nature of the experiment or measurement which gives rise to the random variable. Calculations and

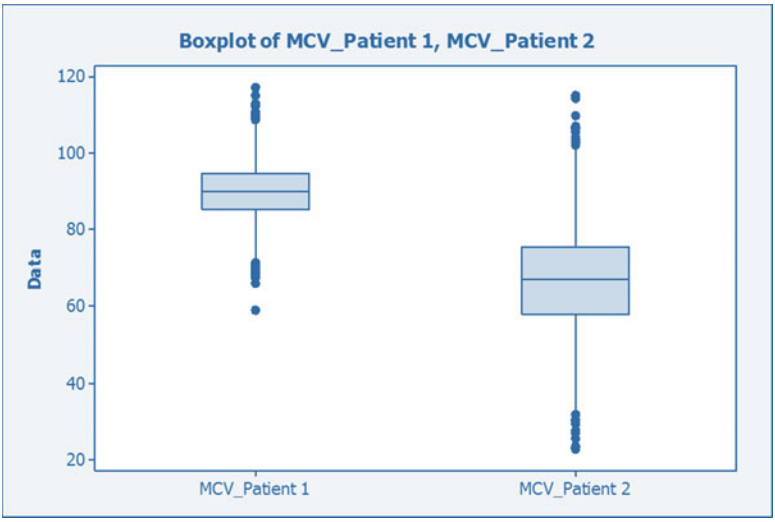


Fig. 3.16 Boxplot of mean corpuscular volume in a normal individual (patient 1) and in patient 2 who has iron deficiency anemia (decreased MCV and increased RDW)

measurements of probability differ based on the type of experiment. Conditional probabilities and Bayesian probabilities are two broad categories of probability.

Finally, one of the most important features of a random variable is its distribution. The nature of the distribution, whether discrete or continuous, and their subclassifications will determine the statistical inference tests that can be used to analyze and compare random variables.

References

1. Strike PW. Statistical methods in laboratory medicine. Butterworth-Heinemann; UK. 2014.
2. Kolmogorov AN. Foundations of the theory of probability. Chelsea publishing company. New York; 1956.
3. Durrett R. Probability: theory and examples. Cambridge: Cambridge university press; 2010.
4. Chung KL. A course in probability theory. San Diego: Academic press; 2001.
5. Fishburn PC. The axioms of subjective probability. Stat Sci. 1986;1:335–45.
6. Jaeschke R, Guyatt GH, Sackett DL, Guyatt G, Bass E, Brill-Edwards P, Browman G, Cook D, Farkouh M, Gerstein H, Haynes B. Users' guides to the medical literature: III. How to use an article about a diagnostic test B. What are the results and will they help me in caring for my patients? JAMA. 1994;271(9):703–7.
7. Akobeng AK. Understanding diagnostic tests 2: likelihood ratios, pre-and post-test probabilities and their use in clinical practice. Acta Paediatr. 2007;96(4):487–91.
8. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. BMJ. 2004;329(7458):168–9.
9. Fagan TJ. Nomogram for Bayes theorem [letter]. N Engl J Med. 1975;293(5):257.
10. Caraguel CG, Vanderstichel R. The two-step Fagan's nomogram: ad hoc interpretation of a diagnostic test result without calculation. Evid Based Med. 2013;18(4):125–8.
11. Doob JL, Doob JL. Stochastic processes. New York: Wiley; 1953.

12. Kaplan NO. I. Multiple forms of enzymes. Bacteriol Rev. 1963;27(2):155.
13. Basu MK, Selengut JD, Haft DH. ProPhylo: partial phylogenetic profiling to guide protein family construction and assignment of biological process. BMC bioinformatics. 2011;12(1):434.
14. Galen RS, Gambino SR. Beyond normality: the predictive value and efficiency of medical diagnoses. New York: Wiley; 1975.
15. Elveback LR, Guillier CL, Keating FR. Health, normality, and the ghost of Gauss. JAMA. 1970;211(1):69–75.
16. Gaddum JH. Lognormal distributions. Nature. 1945;156(3964):463–6.
17. Ryan TA, Joiner BL, Ryan BF. Minitab™. Hoboken: Wiley; 2004.

Linear Correlations

4

Linear Correlations in the Medical Laboratory

Central to all statistical evaluations of the reliability of results in quantitative laboratory medicine is the correlation of results of testing for analytes on more than one analyzer. Almost always, at least two analyzers are available in clinical laboratories for performing routine analyses, either functioning concurrently to handle the testing volume or with one analyzer serving as a "backup" analyzer for the other analyzer which is handling the testing volume. Since both analyzers report patient values, the question arises as to how close are the values that the two (or more) analyzers are reporting. This question is answered by performing testing on given patient samples on each analyzer and evaluating whether the values on each sample are the same or different. CLIA and its surrogate regulatory agency, the College of American Pathologists (CAP), require correlation studies at least twice per year, preferably at 6-month intervals. There are at least two ways in which the results of such a study can be evaluated.

Two-Tailed T-Test

First, each sample can be run multiple (say, five) times on each of two analyzers so that the mean and standard deviation can be computed for the sample on each analyzer. These data can then be used in a two-tailed student t-test, as described in Chap. 6. If the p value is >0.05, the two sets of data from the two analyzers are not statistically significantly different from one another, and it can be concluded that the two analyzers give results that are the same.

Of course, this approach assumes that the results that are generated are based on the same method and that the two analyzers are the same. However, there are instances in which, as we discuss below, the results may be generated by two different methods where the units of measurement may not even be the same. In these cases, if the results are proportionate, then they may be expressed in common

A. Momeni et al., *Introduction to Statistical Methods in Pathology*,
DOI 10.1007/978-3-319-60543-2_4

standardized units. For example, the results from each method can be expressed as a fraction of the highest value found. These fractions can then be compared using the T-test.

Irrespective of the results being compared, this method would require around five determinations per sample on each analyzer. Generally, at least ten different samples should be tested covering low, normal, and elevated values for the analytes. Since, in most clinical chemistry laboratories, analyzers now perform upwards of 40 tests, around 2000 determinations must be performed. Since at least two analyzers are involved, the number of tests performed must be doubled to 4000. Irrespective of the speed of the analyzers, this is a time-consuming process. There is an alternate method that requires fewer determinations [1].

Correlation Plots

In this alternate approach, single determinations of each analyte are performed on each of the ten samples referred to in the preceding paragraph. The value for an analyte for each sample on one analyzer (called analyzer 1) is then plotted on the *Y*-axis (ordinate) against the value determined on the other analyzer (called analyzer 2) on the *X*-axis (abscissa). This process is repeated for all ten samples. The resulting plot is called a correlation plot because it shows how good the correlation is between the values determined on each analyzer.

The general equation of a straight line is

$$Y = mX + b \tag{4.1}$$

where m is the slope and b is the *Y*-intercept.

Ideally, both analyzers should give the same value for a given sample. If this were to occur, then the plot of the data would show a straight line with a slope of 1 and a *Y*-intercept of 0. This would indicate that the two analyzers correlate perfectly for testing for a particular analyte.

An example of the "perfect" straight line is shown in Fig. 4.1 (the data for which are given in Table 4.1). This figure is a plot of the points resulting from analyzing eleven samples for BUN over a range of values from about 5–50 mg/dL. To the nearest tenth of a mg/dL, the results for each sample on the two analyzers were the same. However, most of the time, there are differences between the values, and the points do not all fall exactly on the best straight line.

Then, what happens when the points do not fall perfectly on a straight line? We know that there is always statistical error behind each point making the perfect plot unlikely for every analyte. Given that two analyzers are the same, i.e., are made by the same manufacturer and have been previously tested and found to correlate well for all analytes which are tested on the analyzers, we make the assumption that the points generated lie on a straight line. The question is what is the "best" straight line that the points "should" lie on. They do not lie exactly on this line because of the error involved in their determinations. This assumption therefore excludes the

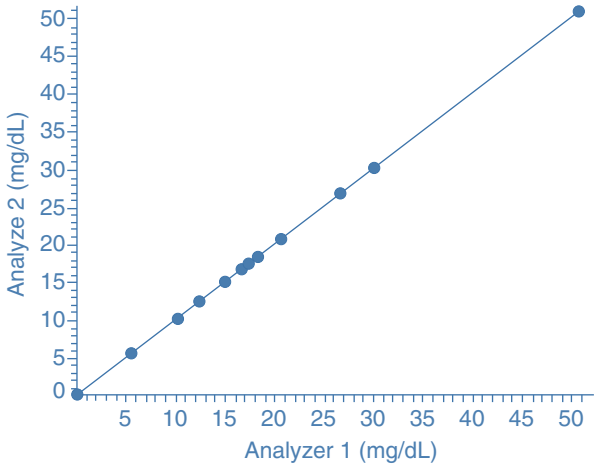


Fig. 4.1 The least squares best fit line to the points from Table 4.1, represented by filled circles, generated from running 11 serum BUN samples on two identical analyzers. All the points lie on or very close to the best fit line. Only one point, the lowest value, point 1 in Table 4.1, deviated slightly from the best fit line. The best fit equation for this line is Y (analyzer 2) = 0.999 X (analyzer 1) + 0.0374. As can be seen from the slope and the intercept, this line is very close to the "perfect" 1:1 line with a slope of 1 and a Y-intercept of 0.0

Table 4.1 Values of serum BUN plotted for 11 samples run on two identical analyzers

Number	X	Y	Y,est.
1	5.5	5.9	5.5
2	10.2	10.2	10.2
3	12.4	12.4	12.4
4	15.0	15.0	15.0
5	16.8	16.8	16.8
6	17.3	17.3	17.3
7	18.3	18.3	18.3
8	20.6	20.6	20.6
9	26.7	26.7	26.7
10	30.2	30.2	30.2
11	50.6	50.6	50.6

The X values (analyzer 1) are given in the first column, the corresponding Y values (analyzer 2) are given in column 2, and the best fit Y values are given in the third column. These values are seen to be virtually identical to those in column 2 except for the first (lowest) value. The plot of column 2 values against the column 1 values is shown in Fig. 4.1. Y,est. is the "best fit" Y value to the points in column 2; these values are seen to be identical to the X values. The straight line in Fig. 4.1 is the plot of Y,est. vs. X

possibility that a nonlinear law better fits the data, which, for this case, appears to be a physically reasonable one.

In fact, in general, correlations between two parameters are not necessarily linear. They may follow any number of other laws such as a quadratic law wherein the best parabola that fits the data would be of the form $AX^2 + BX + C$ or some higher order form as $A_nX^n + A_{n-1}X^{n-1} + A_{n-2}X^{n-2} + \ldots + A_1$. The basic objective for fitting a curve or a straight line to a given set of points is to determine the values of the coefficients for each term in X that gives the lowest sum of squares of the deviations of the Y values from the X values.

Here, we present how the best values for the slope and intercept are determined for determining the *straight line* that gives the *lowest sum of squares of deviations* of the determined Y values from their corresponding predicted values on the "best" straight line [2, 3].

Determination of the "Best" Straight Line Through Experimentally Determined Points

The simplest approach to the determination of the "best" line is to define it as the straight line that results in the lowest possible deviation of the individual points from this line. Thus, for each X value on the correlation plot, we want the Y value to deviate as little as possible, i.e., we want the error, S_i, for the Y value, Y_i, when $X = X_i$, to be as small as possible, and we define Si as:

$$S_i = Y_i - (mX_i + b) \tag{4.2}$$

where m and b are the slope and intercept, respectively, of the "ideal" line Y value. Notice that in this treatment we consider that there is no error in the X_i value, i.e., X_i*is assumed to be absolutely accurate*.

Now, there are the other experimentally determined points for which we wish the same condition to hold. Suppose one point gives an error of, say, -2 and the other gives an error of $+2$. Even though both points deviate from the ideal line, one error cancels the other giving a net error of 0. To avoid this occurrence, we define the square of the error for each point as

$$S_i^2 = [Y_i - (mX_i + b)]^2 \tag{4.3}$$

and

$$\sum S_i^2 = \sum [Y_i - (mX_i + b)]^2 \tag{4.4}$$

where the sum, $\sum$, is taken over all points, in the above example shown in Fig. 4.1, points 1–11. In general, the sum is written as:

$$\sum_{i=1}^{N} S_i^2 = S_1^2 + S_2^2 + S_3^2 + \ldots + S_N^2 +$$

which reads "sum of Si squared i = 1 to N = S12 + S22 + S32 + ... + SN2". For brevity, we represent the sum simply as $\sum$.

Overall, the object is to have the sum of the squares of the individual errors add up to as small a value as possible; that is, we wish to determine the value for m and the value for b that will give the lowest possible value for $\sum S_i^2$.

Derivation of the Least Square Best Fit Line Through the Experimentally Determined Points

Using the methods of the calculus of functions of more than one variable (here, two variables), we have to minimize $\sum S_i^2$, i.e., we must take its first derivative with respect to the slope, m, and the intercept, b, and set each resulting equation equal to 0. This process gives the values for m and b for which $\sum S_i^2$ is a minimum.

Therefore, for the slope,

$$\partial/\partial m \sum S_i^2 = \sum 2S_i \cdot \partial S_i/\partial m = 0, \tag{4.5}$$

and

$$\sum S_i \cdot \partial S_i/\partial m = 0 \tag{4.6}$$

Now, $\partial S_i/\partial m$, from Eq. 4.2, is $-X_i$. So Eq. 4.6 becomes

$$\sum S_i \cdot \partial S_i/\partial m = \sum -Y_i X_i + m \sum X_i^2 + b \sum X_i = 0 \tag{4.7}$$

Similarly, for the intercept,

$$\partial/\partial b \sum S_i^2 = \sum 2S_i \cdot \partial S_i/\partial b = 0 \tag{4.8}$$

and

$$\sum S_i \cdot \partial S_i/\partial b = 0 \tag{4.9}$$

Now, $\partial S_i/\partial b$, from Eq. 4.2, is -1. Therefore,

$$\sum S_i \cdot \partial S_i/\partial b = 0 = -\sum -Y_i + m \sum X_i + \sum b. \tag{4.10}$$

We now have two Eqs. 4.7 and 4.10, with two unknowns, i.e., m and b. We can write these, respectively as:

$$m\sum X_i^2 + b\sum X_i = \sum X_i Y_i \tag{4.11}$$

and

$$m\sum X_i + \sum b = \sum Y_i, \text{and} \tag{4.12}$$

$$m\sum \mathrm{Xi} + nb = \sum \mathrm{Yi} \tag{4.13}$$

where n is the number of points.

Solving these simultaneous equations for m and b, we get:

$$m = \left(n\sum Y_i X_i - \sum X_i \sum Y_i\right) / \left(n\sum X_i^2 - \left[\sum X_i\right]^2\right) \tag{4.14}$$

and

$$b = \left(\sum Y_i \sum X_i^2 - \sum X_i \sum X_i Y_i\right) / \left(n\sum X_i^2 - \left[\sum X_i\right]^2\right). \tag{4.15}$$

For the data plotted in Fig. 4.1, the slope, m, of the best fit line is 0.999, and the Y intercept, b, is 0.0374.

Equations 4.14 and 4.15 give values for m and b such that the deviations of the determined Y points will have the lowest possible deviation from the best straight line that runs through these points. This least squares best fit line to the experimentally determined points is referred to as the *regression line*, and the process described using Eqs. 4.5, 4.6, 4.7, 4.8, 4.9, 4.10, 4.11, 4.12, 4.13, 4.14 and 4.15 is referred to as regression analysis.

We now rewrite Eq. 4.4, once the optimal values of m and b are computed, as

$$\sum S_i^2 = \sum (Y_i - Y_i, \text{est})^2 \tag{4.16}$$

where Y_i,est. is the computed value of Y from X_i using the optimized values for m and b from Eqs. 4.14 and 4.15 and further defines the average deviation, S^2, as

$$S^2 = \sum (Y_i - Y_i, \text{est})^2 / N \tag{4.17}$$

and

$$S = \left[\sum (Y_i - Y_i, \text{est})^2 / N\right]^{1/2} \tag{4.18}$$

In this formulation, S^2 is the average of the sum of the squares of the deviations of Y_i points from the corresponding computed values which we call Y_i,est. (or Y_i, estimated or computed) using the optimized values for m and b. This expression is of the identical form to the variance described in Chap. 3, and S in Eq. 4.18 is

identical in form to the equation for computing the standard deviation of points from the mean. In fact, we can construct lines in which 1S and integral multiples of S are added to b, the Y-intercept for the best fit straight line, and obtain 68% (1S), 95% (approximately 2S), and 99.7% (approximately 3S) of all the points that would be included between the resulting lines.

The value of S for the BUN correlation plot in Fig. 4.1 is 0.22. From Chap. 3, the coefficient of variation of CV is the standard deviation, which for the correlation plot is S in Eq. 4.18, divided by the mean. For the data plotted in Fig. 4.1 (black line), the mean is 20.281. Thus, the CV = 1.08%, an indication of low error. Normally, the two lines representing + 2 standard deviations and −2 standard deviations from the best line shown in Fig. 4.1 in which 95% of the points should fall are drawn parallel to the best fit line. For the data plotted in Fig. 4.1, the standard deviation is so small that the two lines lie very close to the best fit line and have therefore not been drawn in this figure [4, 5].

Correlation Coefficient

Once the best slope and intercept for the points from a given set of determinations is completed, the question arises as to how good the correlation is between the corresponding points. The quantitative value for the closeness of the correlation is given by Eqs. 4.17 and 4.18. The problem in evaluating the closeness of fit from S^2 is that it is difficult to evaluate the meaning of specific values for S^2 without having some standard for comparison. Obviously, if S^2 is 0 or close to 0, the fit can be judged to be excellent. For higher values, the degree of closeness of fit becomes less apparent.

The currently accepted method for evaluating closeness of fit is to define the *total variation of Y values* from the *mean value*, i.e., $\sum(Y_i - Y_{av})^2$, where Y_{av} is the mean of the Y values. It can be shown that

$$\sum (Y_i - Y_{av})^2 = \sum (Y_i - Y_i, \text{est})^2 + \sum (Y_i, \text{est} - Y_{av})^2 \tag{4.19}$$

There are two components to this expression, $\sum(Y_i - Y_i, \text{est})^2$, which is the sum of the squares of the minimum deviations of the points from the corresponding points on the best straight line through them, which contains the uncertainty of Y_i since there is error in its determination, and $\sum(Y_i, \text{est} - Y_{av})^2$, which is a computed set of values, Y_i,est. derived from the best fit line to the points and Y_{av} which is the average of all of the Y_i points. Due to these considerations, the first term in Eq. 4.19 is called the *unexplained variation* (due to error in determination of the Y values), and the second term is referred to as the *explained variation.* The *correlation coefficient, r, is then defined as the square root of the ratio of the explained variation to the total variation,* i.e.,

$$r = \left[\sum (Y_i, \text{est} - Y_{\text{av}})^2 / \sum (Y_i - Y_{\text{av}})^2\right]^{1/2}$$
$$= \left[\sum (Y_i, \text{est} - Y_{\text{av}})^2 / \left(\sum (Y_i - Y_i, \text{est})^2 + \sum (Y_i, \text{est} - Y_{\text{av}})^2\right)\right]^{1/2} \quad (4.20)$$

Note from the right side of Eq. 4.20, if all of the points fall out exactly on the best fit straight line, then the first term in the denominator, i.e., $\sum(Y_i - Y_i, \text{est})^2$, the unexplained variation, is 0, and r therefore must equal 1. On the other hand, if the unexplained variation becomes much larger than the explained variation due to large deviations of the Y_i from the best fit straight line values, then the value of the ratio in Eq. 4.20 approaches 0. Thus, r ranges between 0 and 1, 0 being totally uncorrelated and 1 being a perfect correlation.

Using Eq. 4.20, the *r value for the regression plot in Fig. 4.1 is 0.999* which indicates a strong correlation between the X_i and Y_i values [6, 7].

Problems with This Approach

While Eq. 4.20 gives a "standard" to evaluate the closeness of fit of the computed best straight line through the points, the division of terms into "explained" and "unexplained" deviations is artificial. Errors in Y_i will affect Y_{av}, which is the average of the Y_i values and will also affect their closeness to the best straight line through them and hence will also affect Y_i,est.

Correlation Versus Closeness of Fit to a Straight Line

There is also the question as to the best straight line through points that occur on a perfect horizontal line, i.e., where the slope is 0. Using the example, at the beginning of this chapter, of two analyzers that are determining the concentration of an analyte in a series of serum samples, if one analyzer is not sensitive to changes in analyte concentration and gives the same result for all samples, then there is no correlation between the two analyzers.

From Eq. 4.20, if the Y_i values are identical to the corresponding Y_i,est. values, then r in equa. $20 = [\sum(Y_i, \text{est} - Y_{\text{av}})^2 / \sum (Y_i, \text{est} - Y_{\text{av}})^2]^{1/2}$. Normally, this would be 1.0, but Y_{av} is the same as Y_i,est. The argument is made that, since this makes the numerator $r = 0$, then the value of the fraction $=0$. Thus $r = 0$, i.e., there is no correlation. Omitted from this argument is that the denominator also $=0$. This gives rise to an indeterminate form, 0/0. The actual limit of the fraction $\sum(Y_i, \text{est} - Y_{\text{av}})^2 / \sum (Y_i, \text{est} - Y_{\text{av}})^2$ as Y_i,est approaches Y_{av} is 1. This means that r is 1, indicating the "perfect" correlation when, in fact, there is none even though there is a perfect fit of the points to the horizontal line.

Slopes and Intercepts

Equations 4.14 and 4.15 give the least squares best fit values for the slope, m, and the intercept, b. These values are important for further evaluation of the correlation between the results of the two analyzers.

For the case we presented earlier in this chapter, there are two identical analyzers that perform analyses for analyte levels on the same serum samples. Here, we would expect not only a high value of r, i.e., ≥ 0.9, but also a slope close to 1 and an intercept close to 0. If the r value is high but m is <0.9, and b is significantly high, there is evidence of differences in performance of the two analyzers, and these findings suggest that an investigation of both analyzers is warranted.

In most commercial programs, such as EP-Evaluator (1), that evaluate correlation data, in addition to providing the plot of the data, i.e., the actual points and the best line through these points and the value of r, the correlation coefficient, another plot is presented in which the so-called one-to-one line is also plotted. This is a plot of the line with a slope of 1.0 and an intercept 0.0. This line is termed one-to-one because the angle that this "perfect" line makes with the x-axis is 45°, giving a slope of 1 (the slope is the tangent of this angle and equals 1 for 45°), and the change in value for any two successive X_i values is the same. If this line significantly differs from the best fit line actually obtained, this finding would prompt an investigation.

However, not all correlation studies are constrained by the above considerations. When a correlation study is performed on two different methods that analyze for levels of the same analyte, slopes of <0.9 and intercepts that differ significantly from 0.0 can be accepted. If the r value is >0.9, the correlation equation may be used to "convert" the result obtained using one method to the result that would have been obtained using the other method. This situation is often encountered in immunoassays. One company uses one monoclonal antibody to detect the antigen of interest, while another company may use a polyclonal antibody or a different monoclonal antibody. These antibodies may react with different determinants with different affinities, and the different determinants may be subjected to different rates of proteolytic degradation. In such cases, while the r values are >0.9, the slopes may be considerably lower (or higher) than 1, and the Y-intercepts may differ significantly from 0.0.

Errors in the Slopes and Intercepts

Since, most often, most Y_i values in a correlation study lie off the least squares best fit line, giving rise to deviations or errors, there will be, in general, errors in the slopes and intercepts.

Determination of the errors in the slopes and intercepts of correlation studies is important because these errors determine the range of values that can be assumed by each of these two parameters. In the case of the two identical analyzers discussed above, for a correlation coefficient of >0.9, suppose we find that $m = 0.85$ and

$b = 5.0$. The question is will the error in m include 1 as a possible value and the error in b allow inclusion of 0?

Determination of the error for each of these two parameters allows us to compute what is termed the confidence interval or CI. Knowledge of the CI allows direct determination of whether the slope and intercept can assume values of 1.0 and 0.0, respectively.

Error in the Slope

The accepted definition of the error in the slope for the regression line is

$$\text{Error in Slope} = \sigma = \left[\sum (Y_i - Y_i, \text{est})^2 / (N-2) / \sum (X_i - X_{\text{av}})^2\right]^{1/2} \quad (4.21)$$

where N is the number of points and X_{av} is the average of the X values. The numerator is S in Eq. 4.18 except instead of N, we are using $N - 2$. The difference between these becomes small as N increases. we can therefore write

$$\sigma = S / \left[\sum (X_i - X_{\mathbf{av}})^2\right]^{1/2} \quad (4.22)$$

This equation that results from a theoretical analysis of the variance of a function of random variables states that the error in the slope is the error in the experimentally determined Y values, as defined in Eq. 4.18, i.e., the square root of the average of the squares of the errors or differences between the individual Y_i points from the corresponding regression least squares best fit line values, divided by the sum of the squares of the differences between the corresponding X values and the average X value. The $(N - 2)$ value in Eq. 4.21 (we used N in Eq. 4.18) is used rather than N because designation of a slope involves two points so there are $N - 2$ degrees of freedom.

Overall, this equation computes the square of the errors in Y divided by the sum of the squares of the change in X. This ratio is divided by $N - 2$ to obtain an average square of the error in Y as X changes. The square root of this quantity gives the average error in Y as X changes, and this is equated to the error in the slope.

Ideally, computation of the error in the slope should involve differences between successive Y values for given X changes. Since it is assumed that all variations in values that occur are due to variations in Y (remember all X values are assumed to be completely accurate), the error in the slope should be computed as the error of the difference between two successive Y values divided by the difference in the X values corresponding to these two Y values. However, it is impossible to compute the error of the difference between two points, each of which has its own respective error.

An Alternate Method

Another approach to computing the error of the slope (and, also, the error of the intercept) might be to consider that, for N points, there are $N(N-1)/2$ pairs of points. Since two points determine a straight line, there are $N(N-1)/2$ lines that can be drawn for the N points. Each of these lines has a slope and an intercept.

The error can be assessed in two ways. The first is to compute the mean slope and the mean intercept from the $N(N-1)/2$ lines and then compute the respective standard deviations for these. The second would be to compute the difference between the slope and the intercept of each line and the slope and intercept of the regression line. The square root of the sum of the squares of these differences divided by the number of lines minus 2 (to take into account the degrees of freedom), i.e., $[N(N-1)/2]-2$, would give the error. Note that, if all points lie on the best fit line, then the slopes and intercepts of all lines will be the same as the slope and intercept of the best fit line, and there will be zero error for both slope and intercept.

There is a formulation based on this overall approach, called the Passing-Bablok method, which is often referenced when the results of regression analysis of data points are presented. In this method, point pairs that have slopes around 0 are discarded as outliers.

Confidence Interval for the Slope

Regardless of what method is used to compute the error in the slope and intercept, it is necessary to compute the confidence interval for the slope (and intercept) values. Confidence intervals are discussed in Chap. 2.

By way of review, the basic concept of the confidence interval for regression lines is to *calculate the error in the mean or, in our present case, the error of the slope (and the intercept)* if we repeat the exact same determinations of the samples run on the two analyzers. We would expect to generate points that are similar but not, for the most part, identical to the ones we obtained in the first experiment. We can further expect that the slope and intercept of the regression line will also be similar but not identical to those determined in the first experiment. The question then arises as to what variation, or error, in the slope and intercept can we expect to occur if we repeat the same experiment a large number of times? This error can be computed using a simple equation

$$\text{Error of the mean} = \sigma/\sqrt{N} \tag{4.23}$$

where σ is one standard deviation from the mean.

Since the distributions of values for the slopes and intercepts that would be determined in successive experiments are assumed to follow a Gaussian distribution, this error is the same as one standard deviation from the mean (or, here, one standard deviation from the regression value). Notice, in this equation, the error decreases when N increases because the effect of random fluctuations is diminished when the number of determinations becomes large.

It may be recalled from Chap. 3 that approximately 80% of values occur within one standard deviation, about 95% of values occur within 2 (really 1.96) standard deviations, and about 99% values occur within 3 (2.57) standard deviations. The actual percent of inclusion of values in a Gaussian distribution depends on the number of degrees of freedom.

Thus, for a given mean and number of degrees of freedom, one can compute the number of standard deviations that will include a desired percent of all values. This percent of all values is called the level of confidence, which is mostly the 95% level. The parameter that determines the number of standard deviations that will give this desired level is the so-called Z (t for small values of N) parameter.

We can now define the confidence interval (CI) as

$$\mathrm{CI} = m \pm Z \cdot \sigma / \sqrt{N} \tag{4.24}$$

where m is the regression line slope.

Note, in this equation, Z is a multiple of one standard deviation. σ for the slope of the regression line in Fig. 4.1 is 0.0056, using Eq. 4.21. The CI for 95% confidence level, using Eq. 4.24, is 0.986–1.012. This interval includes 1.0, indicating good correlation between the two analyzers. Thus, the slope for this line may be written as 0.999 ± 0.013.

Error in the Intercept (S_{int})

The accepted definition of this error, again based on the theoretical analysis of the variance of a function of random variables, is

$$S_{\mathrm{int}} = \left[\sum (Y_i - Y_i, \mathrm{est})^2/(N-2)\right]^{1/2} \cdot \left[\left(\sum X_i{}^2\right)/[N-2]\left(\sum X_i - X_{\mathrm{av}}\right)\right]^{2}\Bigg]^{1/2} \tag{4.25}$$

Using Eqs. 4.18 and 4.21, this equation can be written as

$$S_{\mathrm{int}} = \sigma \cdot \left[\sum X_i{}^2\right)/[N-2]\Big]^{1/2} \tag{4.26}$$

where σ is the error in the slope as defined in Eq. 4.21. That the error in the slope influences the error in the intercept may be seen by changing the angle between a straight line and the X-axis, which determines the slope. Even small changes in this angle will cause significant changes in the Y-intercept.

The confidence interval for the intercept is, using Eq. 4.24,

$$\mathrm{CI} = b \pm Z \cdot S_{\mathrm{int}}/\sqrt{N} \tag{4.27}$$

where b is the computed intercept.

For the regression line in Fig. 4.1, σ for the intercept, using Eq. 4.24, is 0.1332, and the CI using 95% confidence level is −0.2640 to 0.3389. This interval contains the value 0.0. We can write the *Y*-intercept for the data in Fig. 4.1 as 0.0374 ± 0.302. Thus, the line with a slope of 1 and an intercept of 0 is within the range of possible lines that fit the data.

Bias

This refers to the extent to which (X_i,Y_i) points may be skewed to higher or lower values in correlations. If, say, most of the Y_i points are higher in value than the corresponding X_i points, respectively, then there is a bias toward higher *Y* values. If this finding is consistent, it may indicate the presence of a systematic error as defined in Chap. 2. There is no generally accepted method for measuring bias, but the extent of bias can be measured at least semiquantitatively in at least two ways. First, one can compare the means for the X_i and Y_i values to detect any significant differences. The second is to plot the differences between Y_i and X_i on the *Y*-axis and the X_i values on the *X*-axis to detect possibly consistent bias trends. There are basically two types of trends: random and nonrandom. Random implies that about as many values lie above the *X*-axis as below it and the points are distributed randomly above and below or on the *X*-axis. Nonrandom means that there is a definite pattern. Figure 4.2 shows the bias plot for the data used to construct the line in Fig. 4.1. As can be seen in this figure, there is no appreciable trend in either direction, indicating overall no significant bias. In contrast, Fig. 4.3 shows a nonrandom "U-shaped" distribution of differences. Nonrandom patterns suggest possible systematic errors involved in one or the other analyzer [8, 9].

Linearity and Calibration

Linearity is the term used for the procedure to establish the limits of sensitivity of a given quantitative assay. For clarity, we use an example of linearity analysis assays based on measuring the absorbance of a compound or complex of compounds to determine its concentration. The concentration of a compound is proportional to its absorbance at an appropriate wavelength of light. This relationship may be written as Beer's law, i.e.,

$$A = \epsilon \times C \times I \tag{4.28}$$

where *A* is absorbance, *C* is the concentration of the compound in solution, ϵ is the proportionality constant, called the molar extinction coefficient, and l is the path

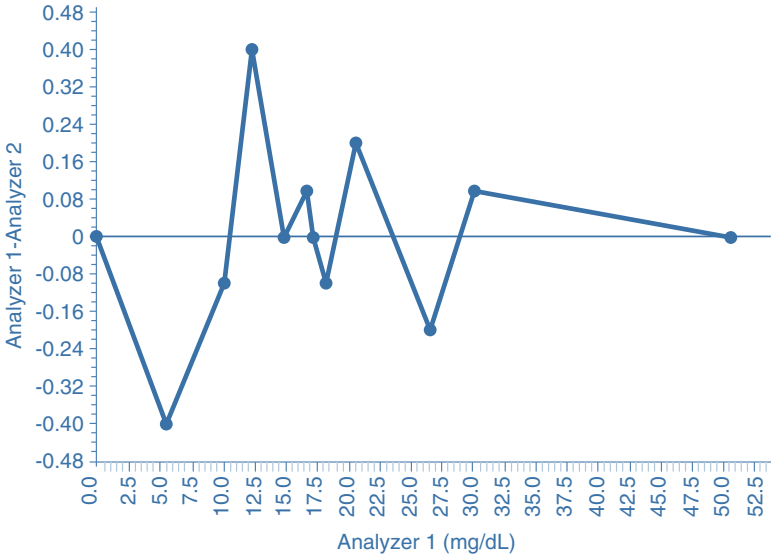


Fig. 4.2 Plot of the bias of the points from Table I. This plot is the difference in values between analyzer 1 and analyzer 2 (analyzer 1 value-analyzer 2 value) plotted against the *X* value (analyzer 1) for the analyte level. Note that there is no constant bias in one direction or the other: as many points lie below the *X*-axis as lie above it, and 4 of the 11 points lie on the *X*-axis, i.e., a bias of 0

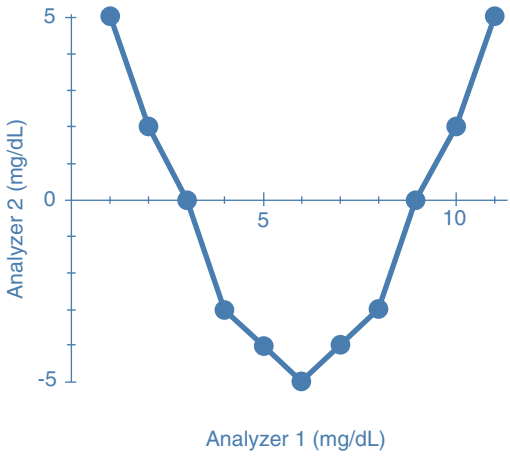


Fig. 4.3 A plot showing an example of bias wherein intermediate values show a distinct negative bias

length in cm. Since in overwhelmingly most cases, $l = 1$ cm, Eq. 4.27 can be written as

$$A = € \times C \tag{4.29}$$

At low or high concentrations, the linear relationship between A and C may break down. Thus, the purpose of linearity analysis is to determine the range of concentrations over which the relationship of A to C is linear, i.e., Eq. 4.28 holds.

One manner in which this task can be accomplished is to make a solution of the compound at the highest level desired at very high accuracy and then to dilute very accurately the solution with an appropriate diluent serially to low levels. These solutions are termed standard solutions or just standards.

The absorbance of each of these solutions is then determined, and a least squares best fit line is constructed for these absorbances using Eqs. 4.14 and 4.15. If this line has an r value (correlation coefficient) >0.9, a slope of 1 that is included in the confidence interval for the slope, and an intercept that has a value of 0 that is included in its confidence interval, the levels of the lowest and the highest concentrations in the study "define" the linearity range as defined in Eqs. 4.24, 4.25 and 4.26 above. This range is referred to as the "analytical measured range" or AMR.

Further procedures that can be performed include comparison of the computed value of €, which is the slope of the best fit line from Eq. 4.28, with its known value. If these are similar or identical, then the linear relationship is further confirmed. In addition, determination of the value of the absorbance of the diluent itself (with no compound present), used in making up the standard solutions, can be performed. This value should be 0 or close to it and should be included as a point for the construction of the best fit line [10, 11].

Practical Considerations

CLIA and CAP require that linearity studies be performed in the same manner as correlation studies, i.e., at least twice per year, preferably every 6 months as described above. Most laboratories do not have the facility for composing standard solutions and, even if they did, do not have the appropriate staff to make these solutions.

To fill in this gap, commercial companies have established facilities for manufacturing standard solutions for most analytes assayed for in the clinical chemistry laboratory. These standard solutions, which are prepared by manufacturers with laboratories dedicated to making these standard solutions, contain, or should contain, very accurately determined concentrations of the analyte over a wide range of concentrations. With all other factors being equal, the results on the medical laboratory analyzer should be close to the value of the accurately determined analyte level at the manufacturer's laboratory.

Calibration

There is a problem that arises with this procedure. There are many different analyzers in medical laboratories. Each analyzer has different calibrators for their

instruments. Calibration is a process in which, still using absorbance measurement as an example, the absorbance of a standard solution, whose concentration is known to great accuracy, is set to have an absorbance that will give this concentration for that solution.

This process is repeated for another standard solution. Depending on the analyzer and the analyte, this process may be performed several times or may just be performed for two standard solutions, a so-called two-point calibration. Once the channel used for measuring the analyte has been calibrated in this manner, assays of samples can then be performed.

The problem that arises in performing linearity studies is that the standard solutions provided by the manufacturers may not have similar compositions to the calibrator solutions used on the medical laboratory analyzer. This can lead to significant differences between the laboratory value and the one determined by the manufacturer for any given analyte.

Therefore, the manufacturer performs assessments of the results for all medical laboratories and divides the results into peer groups, each group using the same manufacturer's analyzer. For each peer group, the mean value and the standard deviation for the level of the analyte in question for each solution are determined for the peer group. If a medical laboratory is found to have a value for a given solution that lies outside ± 2 standard deviations from the peer group mean, the point is said to fail linearity. If two or more points are found to lie outside the 2 standard deviation range, the entire assay for the analyte in question is judged to have failed linearity. Thus, the criterion for linearity is based on peer-group statistics.

However, there is a flaw in this procedure. A regression analysis of the points obtained for the standard solutions for a given analyte is performed. The least squares best fit line is the determined for these points, and the slope and the intercept and the confidence interval for each are likewise determined. It is possible that, even for several points that differ by >2 standard deviations from the group mean, the correlation coefficient can be >0.9, and the confidence interval for the slope and the intercept can include 1.0 and 0.0, respectively. In such cases, the values can be said to lie on a straight line with a slope of 1.0 and an intercept of 0. With the correlation coefficient >0.9, the points pass linearity.

Summary

We have shown the methods that are to evaluate how well two analyzers correlate when they perform quantitative assays for specific analytes. If the two analyzers are identical and use the same methods, the correlation between the values obtained on the two analyzers should be linear. To test whether the values determined do correlate linearly, regression analysis is applied to the experimentally determined points. This analysis gives the slope and the intercept for the straight line that can be drawn through these points that give the minimum or lowest deviation of the sum of the squares of these points from the corresponding values computed from the slope and intercept of this line.

To judge whether the points do conform to a straight line, the correlation coefficient, which is computed as the ratio of the explained error to the sum of the explained + the unexplained error, should be greater than 0.9 although this is not a statistically determined criterion. In addition, the confidence interval for the slope should include the value of 1.0, and the confidence interval for the intercept should include the value of 0. Although many correlations are linear, especially the ones between values of analytes assayed by two identical analyzers, correlations in general may not be linear but can follow other functional forms. The general method is the same as that used for linear correlations, i.e., determination of the values of the coefficients for the mathematical function that minimize the sum of the squares of the deviations of the experimental values from the corresponding values predicted by the mathematical function.

The same methodology can be applied to linearity analysis where standards solutions whose concentrations have been accurately determined are assayed by a medical laboratory. Ideally, the results should be close to the predetermined value.

References

1. Strike PW. Statistical methods in laboratory medicine. Butterworth-Heinemann: UK; 2014.
2. Almeida AM, Castel-Branco MM, Falcao AC. Linear regression for calibration lines revisited: weighting schemes for bioanalytical methods. J Chromatogr B. 2002;774(2):215–22.
3. Williams EJ. A note on regression methods in calibration. Technometrics. 1969;11(1):189–92.
4. Coresh J, Astor BC, McQuillan G, Kusek J, Greene T, Van Lente F, Levey AS. Calibration and random variation of the serum creatinine assay as critical elements of using equations to estimate glomerular filtration rate. Am J Kidney Dis. 2002;39(5):920–9.
5. Craven BD, Islam SM. Ordinary least-squares regression. Sage Publications: CA, USA; 2011.
6. Sedgwick P. Pearson's correlation coefficient. BMJ. 2012;345(7):e4483.
7. Daniel WW, Cross CL. Biostatistics: a foundation for analysis in the health sciences. Wiley: USA; 2013.
8. Plebani M. The detection and prevention of errors in laboratory medicine. Ann Clin Biochem. 2010;47(2):101–10.
9. Howanitz PJ. Errors in laboratory medicine: practical lessons to improve patient safety. Arch Pathol Lab Med. 2005;129(10):1252–61.
10. Panteghini M, Forest JC. Standardization in laboratory medicine: new challenges. Clin Chim Acta. 2005;355(1):1–2.
11. Jhang JS, Chang CC, Fink DJ, Kroll MH. Evaluation of linearity in the clinical laboratory. Arch Pathol Lab Med. 2004;128(1):44–8.

Cross Tabulation and Categorical Data Analysis

5

Introduction

Statistical inference is a powerful tool that allows us to make sense out of data. Using statistical tests, we can draw conclusions about the distribution of data, associations of events with each other, or their correlation with each other. In this chapter, we will introduce the concept of hypothesis testing and explain statistical tests used for hypothesis testing for categorical variables. These tests generally benefit from cross tabulation of data. Cross tabulation is the summarization of categorical data into a table with each cell in the table containing the frequency (either raw or proportional) of the observations that fit the categories represented by that cell. The summary data presented in cross-tabulated form can then be used for many statistical tests most of which follow a distribution called chi-squared distribution. These relatively simple tests are very powerful tools that can help a pathologist in many aspects from result verification to test validation. For example, if we have a new test with a yes and no answer and we want to see if this diagnostic test can diagnose a condition, then we need to use chi-squared tests to determine the usefulness of the test [1].

Most statistical software programs are equipped to run these tests; however, users need to understand the context in which each test can be applied, learn how to interpret the test results, and know the possible limitations of each test.

Before we can delve into these concepts, however, we need to define categorical variables and explain the notion of contingency tables.

Categorical Variables

"Categorical variables" (also known as "nominal variables") represent qualitative properties that allow categorization of observation units (or test subjects) into nominal categories. These variables usually take on discrete and sometimes fixed

A. Momeni et al., *Introduction to Statistical Methods in Pathology*,
DOI 10.1007/978-3-319-60543-2_5

unordered values, with each possible value designated as a "level." Categorical values follow discrete distributions (see Chaps. 2 and 3).

Categorical variables in the simplest form have binary values, i.e., they can only assume one of the two values. For example, disease status can be a categorical variable that lists individuals as either "affected" or "healthy." Categorical variables can also be polytomous, meaning that they can assume more than two values. For example, in Bethesda guidelines for reporting of thyroid cytology specimens, the values can be one of the six categories: nondiagnostic, benign, atypia of undetermined significance, follicular neoplasm, suspicious for malignancy, and malignancy.

Statistical tests used for categorical variable are different from those used for continuous variables. In categorical variables, usually the statistical questions are whether group allocations are similar or dissimilar between individuals and variables. Alternatively, it can be stated that statistical tests for categorical variables in general either test for independence (association) or homogeneity.

Categorical variables commonly assume values that are determined by qualitative properties of the unit of observation. For example, if the categorical variable is disease status, then the values that the variable can assume are inherently qualitative like "yes/no" or "affected/unaffected." Quantitative properties and data, however, can be discretized into categorical variables or dichotomized into binary variables. For example, hemoglobin values can be dichotomized into anemic and non-anemic. In Chap. 2, we showed that using receiver operating characteristic (ROC) curve a continuous variable can be dichotomized into a binary categorical variable based on desired specificity and sensitivity. While continuous variables contain more information, discretization into categorical variables can allow easier interpretation and analysis of the data.

Contingency Table

Contingency tables are used to analyze the associations of two categorical variables. We saw an example of contingency table in Chap. 2 which is often used in pathology and laboratory medicine: the 2×2 contingency table of test status versus disease status. Using tests to categorize individuals into disease status groups is one of the ultimate goals of pathology. Contingency tables are two dimensional matrices composed of rows and columns. The rows (r) represent the possible values of one of the variables, and the columns represent the possible values of the other variable (c). T*hus, a contingency table is a* $r \times c$ *matrix. In each cell of the table, the numbers or proportions of the units* of observation that fit the categories represented by that cell are provided. Table 5.1 represents a contingency table of eye color versus hair color in 100 individuals.

Overall the most common contingency table used is the 2×2 table which sets two binary variables against each other. Some of the most important statistical tests for categorical variables use the 2×2 contingency table. As such, sometimes

Table 5.1 A 4 × 4 contingency table of hair color versus eye color. Each cell represents the number of individuals that fit the categories represented by the cell. Row and column totals are usually included in a contingency table

Eye / hair	Black	Brown	Green	Blue	Total
Black	9	10	3	1	23
Brown	5	5	5	3	18
Blonde	1	2	6	8	17
Ginger	0	1	6	5	12
Total	15	18	20	17	70

categorical variables with more than two values are dichotomized (combined) into a binary variable for inclusion in a 2 × 2 contingency table.

Some of the information that can be from a contingency table are descriptive measures. The simplest of these measures is "count" which is the raw observed frequency of a cell. Another measure is "relative frequency" stated in proportion or percentage which shows the proportion of total data, row total, or column total that is represented in the cell. If relative frequencies are used, it is imperative that the table states which relative frequency (share of total, row total, or column total) is represented.

As stated earlier, values in each cell can either represent actual counts (raw observed frequencies) or proportions. However, if the purpose of drawing up a contingency table is to perform statistical tests then one should avoid using relative frequencies and proportional data in the cells.

Bar charts are usually used to visualize the information contained in categorical variables. In bar charts, each bar represents a value of the category, and the length of the bars is represented by the frequency of the observed category. Information from a contingency table can also be shown by bar charts. Clustered bar charts are ideal for showing the distribution of categories of a variable within another variable (subcategories), where for each category of the first variable bars representing the frequencies of the second variable are plotted. However, if the goal is to show the proportions of both categories, stacked bar charts should be used, where the bars represent the frequencies of one variable and the fill patterns or colors represent the second variable. Figure 5.1 represents the bar chart of Table 5.1.

Pie charts can also be used to visualize categorical data. Pie charts are especially ideal when the goal is to show the relative frequencies of the categorical variable. If categorical variables have too many levels, then pie charts are less suitable for plotting the variable unless some of levels (or categories) with small relative frequencies are combined. Figure 5.2 represents the pie chart of hair color from Table 5.1.

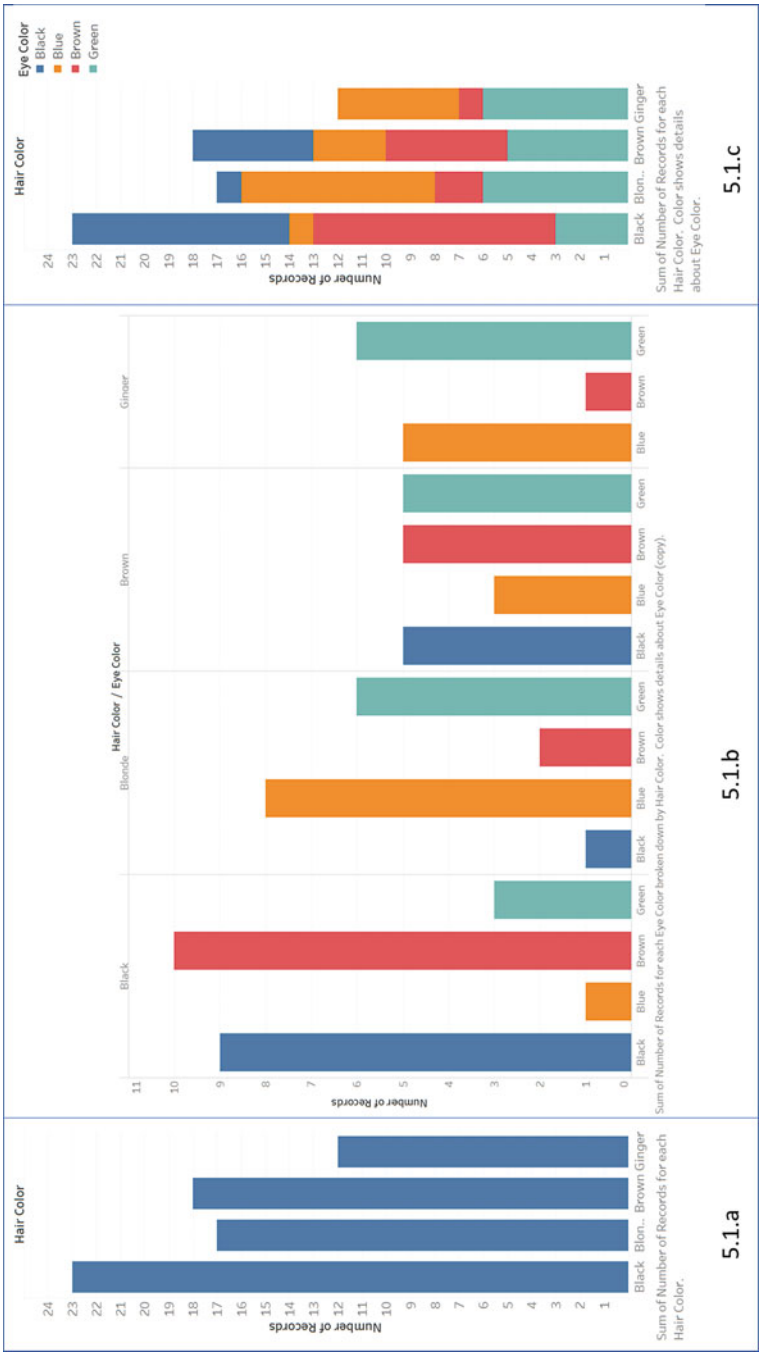


Fig. 5.1 The bar chart of hair color from Table 5.1 (**a**). The clustered bar chart (**b**) and stacked bar chart from Table 5.1 [2, 3]

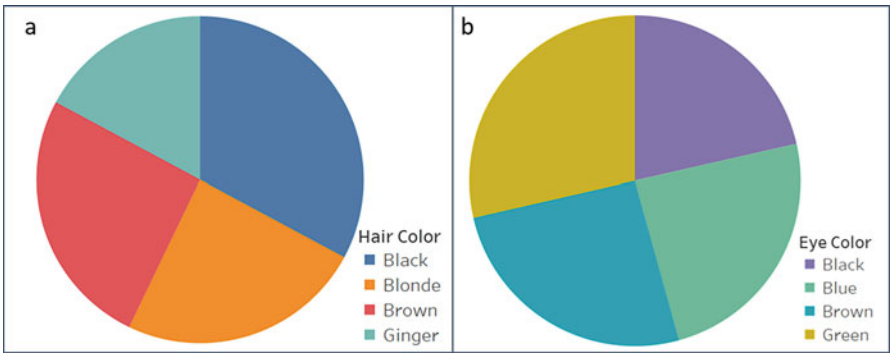


Fig. 5.2 Pie chart of hair color (**a**) and eye color (**b**) from Table 5.1

Hypothesis Testing

Hypothesis is a testable assumption about parameters or phenomenons. The purpose of testing is to determine whether the assumption is true (accepted) or false (rejected). In statistics, "hypothesis testing" is part of "statistical inference" and aims to assess the proposed or assumed statistical relationship between two or more datasets and parameters.

The first step in hypothesis testing is to state the "null hypothesis" (H_0) and the "alternative hypothesis" (H_1 *or* H_a). In simplified terms, the null hypothesis states that there is no statistical relationship between the compared datasets and any observed relationship is either due to error or to chance. Essentially, the null hypothesis is true (or considered likely) if the observed data can be justified using chance and randomness. Alternatively, the null hypothesis is rejected (or considered unlikely) if the observed data cannot be justified using chance alone which means that the alternative hypothesis is likely to be true. Simply stated, alternative hypothesis is the postulation that the observations are due to a real effect. In statistical inference, we usually assume that the null hypothesis is true and tests are designed to reject the null hypothesis.

Stating the null and alternative hypotheses is very important. If they are not clearly defined, then perhaps inappropriate statistical tests are used to test them, or the results of the test can be interpreted incorrectly. For example, in a 2×2 contingency table, the null hypothesis can be that there is no association between two variables (H_0 of testing of independence) with the alternative hypothesis being that the two variables are associated. Another possibility is that the null hypothesis states that the distribution of the categorical variable is the same in two populations (H_0 of testing for homogeneity) with the alternative hypothesis being that the distribution of the categorical variable differs across the populations. While in 2×2 tables the analysis is the same for both these null/alternative hypothesis pairs, the conclusions drawn from the tests are different.

Table 5.2 This table shows the two types of error in statistical hypothesis testing

		H_0	
		False	True
Statistical test for H_0	Rejects	True positive	False positive (type I error)
	Fails to reject	False negative (type II error)	True negative

The second step in hypothesis testing is to choose appropriate statistical tests to assess the hypothesis. This choice depends on the properties of the variables used for testing including their values and distribution as well as the hypothesis being tested. Furthermore, each statistical test has relevant test statistics such as mean, variance, etc.

Table 5.13 provides a summary on choosing the appropriate tests for categorical variables.

The third step in hypothesis testing is to set "power" (β) and "significance level" (α). To understand power and significance level, first you need to understand the concepts of "statistical error." Generally, there are two types of statistical error:

- Type I error: When the null hypothesis (H_0) is true, but it is rejected, then a "type I error" has occurred. The rate of type I error is known as significance level (α) and is the probability of rejecting the null hypothesis while it is true. The significance level is commonly set at 0.05 or 0.01. An alpha level of 0.05 means that there is a 5% probability that the null hypothesis is rejected while it is true.
- Type II error: If the null hypothesis (H_0) is false, but the test fails to reject it, then a "type II error" has occurred. The probability of type II error is called the beta rate (β). The "power of the test" is determined as $1 - \beta$.

Type I and type II errors are akin to false-positive and false-negative concepts introduced in Chap. 2. In fact, a 2×2 contingency table can be drawn to better explain these errors (Table 5.2).

Statistical Power

Statistical power is the complement of β. Power is a concept like sensitivity and is the probability of correctly rejecting the null hypothesis. In other words, power is the ability of the test to detect a statistical effect, if it truly exists.

Power is determined by statistical significance level, magnitude of effect, and sample size. As more stringent significance levels are employed, the power of the study decreases, i.e., assuming all the condition remains constant, a study has more power for a significance level of 0.05 versus a significance level of 0.01. As the significance level is mostly constant, then the statistical power is mostly determined by magnitude of the effect and sample size in real-world situations.

The magnitude of the effect is the difference of the test statistic between the groups being compared. Ideally, the magnitude of the effect should be a

standardized test statistic to have information about both the location and spread of the data. As effect magnitude increases, the test power increases and vice versa.

One of the most important determinants of power is the sample size. As sample size increases, the test power increases. Statistical power analysis is used in trial designs for sample size calculations. Researchers usually set β and α as constants when designing trials. They then choose the desired magnitude of effect and use these to calculate the sample size (choosing the magnitude of effect in clinical setting usually means choosing clinically significant effect). We will explore the concept of sample size calculation further in Chap. 12.

P-Value

Alpha is often mistaken with p-value. Alpha is the level of significance set by the investigators before a test is run. Alpha is usually arbitrarily chosen, but the consensus is that alpha levels of 0.05 or 0.01 are adequate for most practical purposes. After the test is run, the likelihood that the observed statistic occurred due to chance (based on the distribution of the sample) is called the p-value. If p-value is less than or equal to the significance level (p-value $<\alpha$), then the null hypothesis is rejected, and the test result is statistically significant. Alternatively, if p-value $>\alpha$ then the test has failed to reject the null hypothesis, and we can state that the test is statistically nonsignificant. In simple terms, if the observed phenomenon was likely to happen due to chance only, then its p-value will be greater than the chosen alpha level, and we cannot rule out the null hypothesis. The alpha is the cutoff that we choose to say whether something has occurred due to randomness or a true effect and, thus, if the p-value is smaller than alpha, then we can say that what we have observed is probably because of a true effect, and we can reject the null hypothesis.

Unlike α, p-value is dependent on sample size as the p-value calculation requires computing the sampling distribution under the null hypothesis which in turn is dependent on the statement of the null hypothesis, the test statistic used (and its distribution), and finally the data (including the size of the sample).

Essentially, to calculate the p-value, the cumulative distribution function (CDF) of the sampling distribution under the null hypothesis is calculated. After the test is run, the observed values are compared to the CDF, and the p-value then is the probability (assuming H_0 is true) that a value equal or more extreme than what was observed is obtained (Fig. 5.3).

There is a common mistake in interpreting p-values: if the p-value is greater than the significance level, it implies that the test has failed to reject the null hypothesis at the stated significance level; this, however, **does not mean that the null hypothesis is true**.

Let us use a simple example for the concept of alpha and p-value. We have measured the sodium concentration of a serum sample a hundred times, and we have a normally distributed value with mean of 140 mEq/L and standard deviation of 2.5 mEq/L. We measure the sodium concentration for the 101st time. Is the value obtained due to random distribution of the results or is it because of a measurement error?

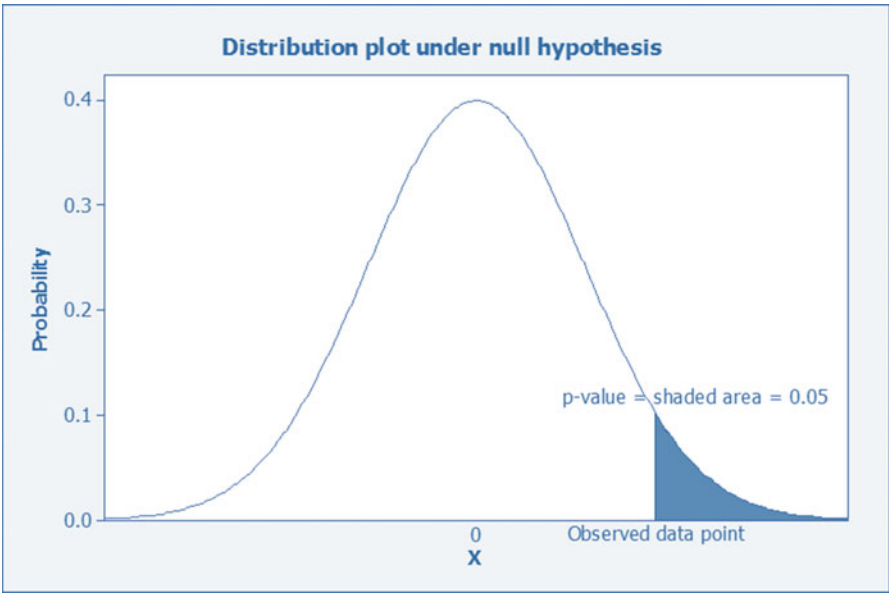


Fig. 5.3 Visual presentation of p-value using the probability distribution plot under null hypothesis

- H_0: The value of 101st measurement is part of the random distribution of the true value.
- H_1: The value of 101st measurement has occurred nonrandomly, most probably because of a measurement error.

For this example, we are going to set the alpha level at 0.05 meaning that if the measured values fall within 95% of the probability distribution around the mean, then we are dealing with a random variation. This 95% incidentally is the 95% confidence interval in normally distributed variables. The 95% CI of our measurements is the mean plus and minus two standard deviations (i.e., 140 ± 5). Thus, if the sodium value of the 101st measurement falls within the range of 135–145 mEq/L, then we can say that we cannot reject the null hypothesis. If we get a value of, for example, 147.5 mEq/L because it is beyond our cutoff of 0.05 probability, then we can say that it occurred nonrandomly and most likely because of a measurement error. In fact, the probability of observing 147.5 mEq/L in the 101st time is exactly 0.00135 which is our p-value. This is shown in Fig. 5.4 [1].

Bonferroni Correction

Under certain conditions the probability of a type I error increases. For example, when multiple comparisons are made, a type I error becomes more likely. In these situation adjustments are needed to the significance level. One of the methods used for adjusting the significance level is called the "Bonferroni method."

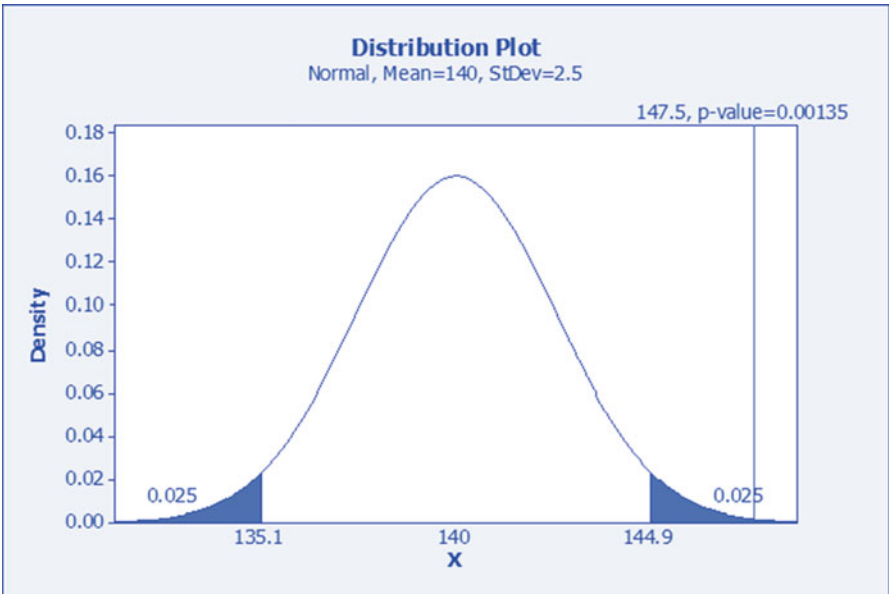


Fig. 5.4 Normal distribution curve of sodium concentration with mean of 140 mEq/L and standard deviation of 2.5 mEq/L. The shaded area shows a two-tailed alpha of 0.05. The reference line shows the value of 147.5 mEq/L. The probability of its random occurrence is 0.00135 which is also its p-value. Since p-value is smaller than the alpha, then we can reject the null hypothesis

If an α level is decided for an entire experiment and the experiment has m tests (T_i, $1 \leq i \leq m$) for the alternative hypothesis (H_i) under the null hypothesis assumption that all the alternative hypotheses are false, then the α level for individual tests (T_i) should be set in a way that:

$$\sum_{1}^{m} \alpha_i \leq \alpha, \tag{5.1}$$

The simplest way of satisfying the above equation is to set the individual test significance levels at α/m. For example, if there are five tests in an experiment with an overall significance level of 0.05, then the significance level for each test can be set at 0.01. Bonferroni correction does not require that all the tests have equal significance levels. This is helpful in studies with interim analysis where the significance level can be set at lower thresholds for earlier phases of the study with higher significance level for the final analysis. For example, if a study has one interim and one final analysis and the overall α is set at 0.05, then the significance level for the interim level can be set at 0.01, and the significance level for the final analysis can be set at 0.04.

Analysis of Risk Ratios

One of the simplest comparisons in a 2 × 2 table is comparing the risks in two groups known as "risk ratio" (also "relative risk") or RR. The calculation of risk ratio is simple with direct comparison of cumulative rate of event in the two groups.

$$\text{Relative risk} = \frac{\text{Probability when exposed}}{\text{Probability when non} - \text{exposed}}, \tag{5.2}$$

Based on a 2 × 2 (Table 5.3) the relative risk can be stated as

$$\text{Relative risk} = {}^{a}\Big/ {}_{c}\Big/ {}_{(c+d)}^{(a+b)}, \tag{5.3}$$

Example 5.1
Q: We are evaluating a new test for diagnosis of pancreatic cancer. The results of the study are summarized in Table 5.4. What is the relative risk of having pancreatic cancer if the test is positive?

A:

$$\text{Relative risk} = \frac{{}^{80}\!/_{95}}{{}^{10}\!/_{85}} \cong 7.15, \tag{5.4}$$

which means that, if the test outcome is positive in an individual, he/she is seven times more likely to have pancreatic cancer than a person with negative result.

Risk ratio can assume values between 0 to ∞. If risk ratio is close to 1, it implies that there is little or no risk difference between the groups, with risk ratios of greater than 1 suggesting increased risk and risk ratios smaller than 1 suggesting decreased risk.

A confidence interval (CI) can be calculated for risk ratio; as the logarithm of relative risk has a sampling distribution that is approximately normal, we can calculate a confidence interval that is located around the logarithm of relative risk.

Table 5.3 A 2 × 2 table of an event status versus disease status

	Disease +	Disease −
Event +	a	b
Event −	c	d

Table 5.4 2 × 2 table of results for Example 5.1

	Pancreatic cancer +	Pancreatic cancer −
Test +	80	15
Test −	10	75

$$\mathrm{CI}_{RR} = \log RR \pm SE \times z_{\alpha}, \tag{5.5}$$

z_{α} is the standard score for the level of significance, and SE is the standard error which can be calculated as

$$\mathrm{SE} = \sqrt{\left(\frac{1}{a} + \frac{1}{c}\right) - \left(\frac{1}{(a+b)} + \frac{1}{(c+d)}\right)}, \tag{5.6}$$

If the confidence interval of relative risk excludes 1, then the relative risk is said to be statistically significant. For example, a relative risk of 1.2 with CI of (1.1–1.3) is statistically significant while a relative risk of 1.5 with confidence interval of 0.5–2.5 is not statistically significant.

Chi-Squared Tests

Comparison of distributions of two categorical variables can be done using the "chi-squared test" (χ^2). Chi-squared statistics compares the counts of categorical variables between two independent groups, or it determines if two categorical variables from a single population are independent or not. χ^2 tests contain a broad set of statistical hypothesis tests where the sampling distribution of the test statistic follows a chi-squared distribution if H_0 is true. In this chapter, we will introduce some of the chi-squared tests including Pearson chi-squared test, McNemar test, and Cochran-Mantel-Haenszel test. Before we introduce these tests, however, we need to explain the χ^2 distribution.

The underlying concept for a chi-squared test is that if we have two nominal variables that occur randomly then we expect the probability of the events to be random. For example, coin toss is a random nominal variable, and we expect that the probability of heads and tails are equal. This is called the expected value. In terms of a 2×2 contingency table, we can say that the expected value, means that the true positive count and false-positive count are equal as well as the true negative count and false-negative counts being equal. In Table 5.4, this means that of the total 95 patients who tested positive, half (47.5) would have the disease and half would not (47.5). Similarly, for the patients who tested negative for the disease (85), half of them have the disease (42.5), and half do not have the disease (42.5). The chi-squared test compares the observed counts to the expected counts. The further the observed counts are from the expected counts, the more likely they are to have occurred nonrandomly.

In fact, the ratio of the squared deviation of the observed values of each cell from expected values of that cell to the expected value is the chi-squared test. The chi-squared statistic that is produced from this equation follows a continuous probability distribution called the chi-squared distribution (which is a kind of gamma distribution). Just like a normal distribution, we can set levels of significance for this distribution and then look if the calculated chi-square statistic is

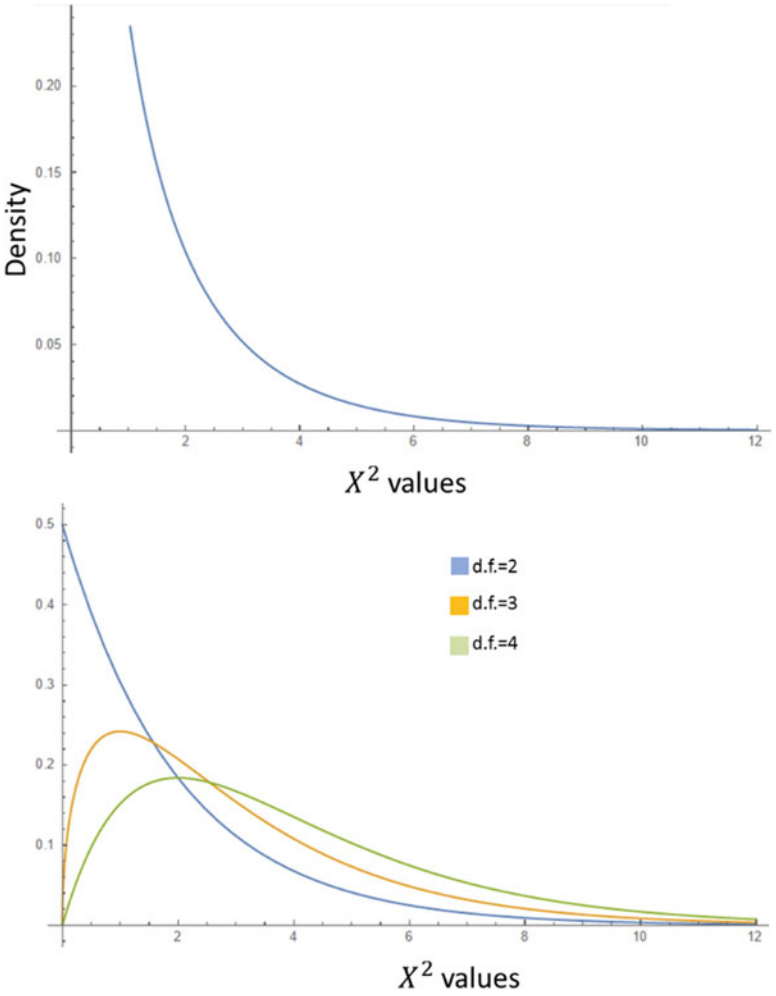


Fig. 5.5 Chi-squared distribution plot with one degree of freedom (upper panel). Chi-squared distribution plots with two, three, and four degrees of freedom (lower panel).

larger than the cutoff value (thus failing to reject the null hypothesis) or smaller than the cutoff value (thus rejecting the null hypothesis). The shape of the chi-squared distribution curve is determined by the degrees of freedom (Fig. 5.5). We discuss the chi-squared distribution in detail below for those interested in the mathematics that results in this distribution. For those who prefer not to peruse this section, please proceed to the section below entitled "The Chi-Squared Probability Distribution Function." [4–6]

Degrees of Freedom

"Degrees of freedom" is the number of values that are free to vary in the calculation of a statistic. The degrees of freedom in contingency tables are the number of cells in the two-way table of the categorical variables that can vary, given the constraints of the row and column totals. Thus, the degree of freedom in contingency tables is determined by the number of columns and rows and can be given by

$$d.f. = (r-1)(c-1), \tag{5.7}$$

where r is the number of rows and c is the number of columns [7].

Chi-Squared Distribution

χ^2 distribution ($Q \sim \chi^2(v)$) is the distribution of sum of the squares of v number of independent standard normal random variables with v degrees of freedom. In other words, it is the distribution of the sum of squared normal deviates (Z_i).

$$\chi^2 \equiv \sum_{i=1}^{v} Z_i^2, \tag{5.8}$$

The χ^2 distribution is a gamma distribution with θ of 2 and α of $v/2$.

"Gamma distribution" is a continuous probability distribution with a scale parameter (θ) and a shape parameter (α). These parameters are positive real numbers. A gamma distributed variable (X) can be stated as

$$X \tilde{} \Gamma(\alpha, \theta), \tag{5.9}$$

To calculate the probability density function and cumulative distribution function of a gamma distributed variable, we need to use "gamma function" and "lower incomplete gamma function." Gamma function (Γ) of n (n being a positive integer) is a sort of factorial function represented as

$$\Gamma(n) = (n-1)!, \tag{5.10}$$

The gamma function can also be stated in integral form (Euler integral form):

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt, \tag{5.11}$$

One of the most commonly used gamma functions is the $\Gamma\left(\frac{1}{2}\right)$ which equals $\sqrt{\pi}$. The proof for this value is beyond the scope of this book. The complete gamma function can be generalized into upper and lower incomplete gamma functions. In the gamma distribution, the lower incomplete gamma function ($\gamma(a, x)$) is of interest to us and is given by

$$\gamma(a,x) = \int_0^x t^{a-1}e^{-t}dt, \tag{5.12}$$

The probability distribution function (PDF) of a gamma distributed variable is given by

$$f(x;\alpha,\theta) = x^{\alpha-1}\frac{e^{-\frac{x}{\theta}}}{\theta^{\alpha}\Gamma(\alpha)} \quad \text{for } x > 0, \tag{5.13}$$

The cumulative distribution function (CDF) is then stated as

$$F(x;\alpha,\theta) = \gamma\left(\alpha,\frac{x/\theta)}{\Gamma(\alpha),}\right. \tag{5.14}$$

The Chi-Squared Probability Distribution Function

The PDF for variables with chi-squared distribution of 1 d.f. can be given as

$$f\left(x;\frac{1}{2},2\right) = \begin{cases} \dfrac{0.797885e^{-2x}}{x^{0.5}} & for\ x > 0, \\ 0 \text{ for } x = 0 \end{cases} \tag{5.15}$$

This is plotted in Fig. 5.4a. The probability distribution plots of a chi-squared distribution with two, three, and four degrees of freedom are plotted in Fig. 5.4b.

In fact, a chi-squared distribution with one degree of freedom is the square of a standard normal distribution ($\chi^2(v) = Z^2$). This close approximation to standard normal distribution makes the chi-squared test ideal for hypothesis testing as it can simplify many statistical calculations. Thus, we can assume that extreme values of a chi-squared distribution, just like in a normal distribution, have low probabilities and consequently will have lower p-values. In fact, in a 2×2 contingency table, if 0.05 is designated as the level of significance, χ^2 values of more than 3.84 are considered statistically significant (i.e., p value <0.05). If 0.01 is chosen as α, then any χ^2 value more than 6.63 will be statistically significant.

In summary, if the observed values vary greatly from the expected values, the chi-squared statistic will increase; if it increases beyond 3.84, then with a significance level of 0.05, we can state that it is highly unlikely that the counts we are observing occurred due to chance, and, in fact, there is a true effect (either independence or homogeneity) present, and we can reject the null hypothesis.

It must be remembered that a chi-squared distribution only approaches a normal distribution if the sample size is sufficiently large (see central limit theorem, Chap. 3), and with small sample sizes, alternative approaches (e.g., Fisher's exact test) should be employed.

The chi-squared distribution table for different degrees of freedom is given in Appendix B.

Pearson Chi-Squared Test

One of the most commonly used chi-squared tests is called the "Pearson chi-squared test"; this statistical test evaluates whether there is a statistically significant difference between distributions of two categorical variables. In other words, the general null hypothesis for this test is that any differences observed between the two categorical variables are due to chance.

Pearson's chi-squared test can test three different sets of null/alternative hypothesis pairs. It is important to remember that while conducting the test is similar for all these pairs, the interpretation of the results is different.

The most common hypothesis tested with Pearson's chi-squared test is a test of independence (also known as test of association). In this case, we are interested in knowing whether there is association between two variables. For example, is there association between having a positive PPD test and having tuberculosis? The null and alternative hypothesis in testing for independence state:

- H_0: There is no association between the two variables, i.e., the variables are independent (e.g., there is no association between PPD result and tuberculosis status).
- H_1: There is association between the two variables, i.e., the variables are associated (e.g., there is association between PPD result and tuberculosis status).

Alternatively, Pearson's chi-squared test can be used to test for homogeneity. In this test, we are comparing the distribution of a categorical variable in two populations. For example, is the distribution of diabetes similar between men and women? The null and alternative hypotheses in this setting are:

- H_0: The distribution of the categorical variable is similar between the two populations (e.g., prevalence of diabetes is the same in men and women).
- H_1: There is a difference in the distribution of the categorical variable between the two populations (e.g., the prevalence of diabetes is different between men and women).

Finally, Pearson's chi-squared test can be used to test for "goodness of fit." In this setting, the observed frequency distribution of a categorical variable is compared to an expected or predicted distribution. For example, we wish to determine if men are 1.5 times more likely than women to have Hodgkin's lymphoma, in a random sample of Hodgkin lymphoma cases drawn from a given population, as has been found to occur in other populations. We inquire if the gender distribution in the population under study is similar to our expectation. The null/alternative hypothesis pair in this case states:

- H_0: The distribution of the categorical variable is similar to the expected distribution (e.g., in our randomly drawn sample from patients with Hodgkin's lymphoma, there are 1.5 times more men than women).
- H_1: The distribution of the categorical variable is different from the expected distribution (e.g., in our randomly drawn sample from patients with Hodgkin's lymphoma, the men and women are equally represented).

After formulating the hypothesis, the next step is to test our data under the null hypothesis. If we are testing for independence or homogeneity, this requires us first to calculate the expected value of the two categorical variables for each cell of the contingency table. The calculation is based on the contingency table with r rows and c columns (in case of a 2×2 table, two rows and two columns):

$$E_{i,j} = \frac{\sum_{k=1}^{c} O_{i,k} \sum_{k=1}^{r} O_{k,j}}{N} \quad \text{for } i(1,2) \text{and } j(1,2), \tag{5.16}$$

where $E_{i,j}$ is the expected value for the cell in column i and row j, $O_{i,j}$ is the observed value for the cell in column i and row j, $O_{k,j}$ is the observed value for the cell in column k and row j, and N is the total count of the contingency table.

The next step is to calculate the value of the chi-squared test (χ^2):

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}}, \tag{5.17}$$

If the degree of freedom is 1, i.e., the contingency table is 2×2, then the formula can simply be written as

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}} = \frac{(O_{1,1} - E_{1,1})^2}{E_{1,1}} + \frac{(O_{1,2} - E_{1,2})^2}{E_{1,2}} + \frac{(O_{2,1} - E_{2,1})^2}{E_{2,1}} + \frac{(O_{2,2} - E_{2,2})^2}{E_{2,2}}, \tag{5.18}$$

Notice that, as we stated earlier, the chi-squared value is the sum of the squares of deviates and with an α of 0.05; if the calculated χ^2 score is greater than 3.84, then we can state that the null hypothesis is rejected.

For goodness of fit studies, we don't need to calculate the expected value using the abovementioned formula: we can derive the expected value for each cell from the expected distribution in the null hypothesis.

Example 5.2

Q: Going back to Table 5.4 (repeated here for convenience), we want to test for association of the new test with pancreatic cancer. State the appropriate null/alternative hypothesis, and determine if the two variables are associated at 0.05 significance level.

A:

- H_0: There is no association between the test and pancreatic cancer.
- H_1: There is an association between the test and pancreatic cancer.

The expected cell counts are:

$$E_{1,1} = \frac{\sum_{k=1}^{2} O_{i,j} \sum_{1}^{2} O_{k,1}}{180} = \frac{(O_{1,1} + O_{1,2}) \times (O_{1,1} + O_{2,1})}{180} = \frac{90 \times 85}{180} = 42.5, \tag{5.19}$$

$$E_{1,2} = \frac{(O_{1,1} + O_{1,2}) \times (O_{1,1} + O_{2,1})}{180} = \frac{90 \times 85}{180} = 42.5, \tag{5.20}$$

$$E_{2,1} = \frac{(O_{2,1} + O_{2,2}) \times (O_{2,2} + O_{1,2})}{180} = \frac{95 \times 90}{180} = 47.5, \tag{5.21}$$

$$E_{2,2} = \frac{(O_{2,1} + O_{2,2}) \times (O_{2,2} + O_{1,2})}{180} = \frac{95 \times 90}{180} = 47.5, \tag{5.22}$$

In other words, if the data were completely random, then the 2 × 2 contingency table would look like below:

	Pancreatic cancer +	Pancreatic cancer −
Test +	47.5	47.5
Test -	42.5	42.5

The expected values for cell 1,1 and cell 1,2 are simply calculated as the sum of observed values for those cells divided by two. The expected values for cell 2,1 and 2,2 are simply the sum of observed values for those cells divided by two. The observed values are provided in Table 5.4. The chi-squared test is the sum of the ratios of squared differences between the observed values of each cell with its expected value divided by the expected (random) value. For example, for patients who have pancreatic cancer and test positive for it, the ratio is 80 minus 47.5 squared divided by 47.5. We do this operation for all four cells and add the results. The final result is the chi-squared score.

In other words, the χ^2 score can be calculated as

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}} = \frac{(O_{1,1} - E_{1,1})^2}{E_{1,1}} + \frac{(O_{1,2} - E_{1,2})^2}{E_{1,2}} \\ &+ \frac{(O_{2,1} - E_{2,1})^2}{E_{2,1}} + \frac{(O_{2,2} - E_{2,2})^2}{E_{2,2}} = \frac{(75 - 42.5)^2}{42.5} \\ &+ \frac{(10 - 42.5)^2}{42.5} + \frac{(15 - 47.5)^2}{47.5} + \frac{(80 - 47.5)^2}{47.5} \\ &= 94.18,\end{aligned} \tag{5.23}$$

The χ^2 is 94.18 which is much greater than the 3.84 significance threshold; in fact, the p-value for this chi-square value is less than 0.00001. Thus, the null hypothesis is rejected, and the two variables are associated, i.e., the outcome of the new test is

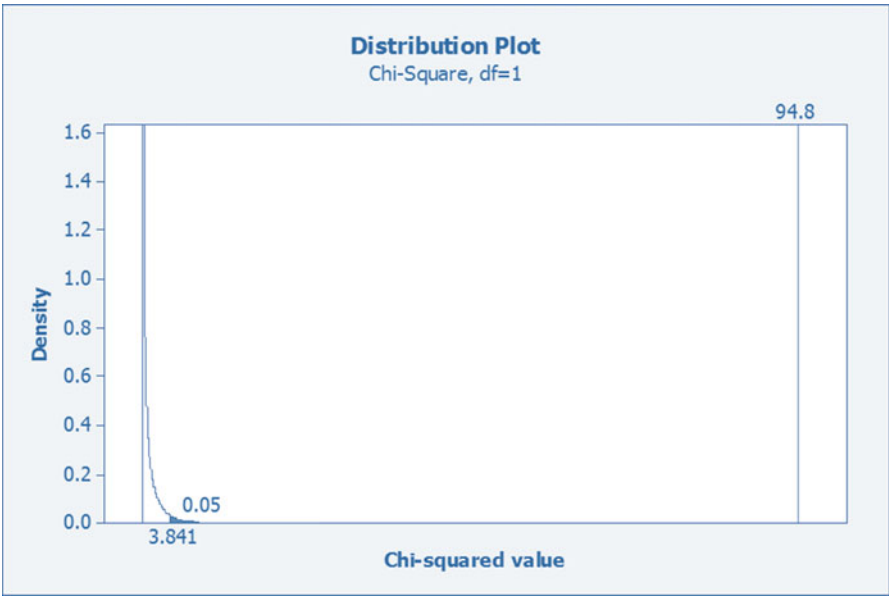


Fig. 5.6 Chi-squared distribution plot for Example 5.2. The shaded area is the 0.05 significance level. You can see that the calculated test statistic is at the very extreme end of the distribution curve

associated with the pancreatic cancer status. In fact, the association is so strong that if we plot the chi-squared distribution, the p-value will be at the extreme end of the distribution tail (Fig. 5.6) [4, 5, 8].

Measures of Association

As you saw earlier, one of the hypothesis tests performed using the chi-squared test is the test for association. The chi-squared score can be used to construct measures known as "measures of association" which show the degree to which two variables are associated.

Phi coefficient (ϕ) is one of such measures that can be calculated for 2×2 contingency tables and is calculated by the following formula:

$$\phi = \sqrt{\frac{X^2}{N}}, \tag{5.24}$$

where χ^2 is the chi-squared score from the Pearson test and N is the total number of observations. ϕ has a range of [0–1]. If ϕ is 0, it shows that the two variables are independent. If the number of rows and columns is more than two, then alternative measures of association should be used. One of these is called the "Cramér V coefficient" and it can be calculated as

$$\text{Cramé r's } V = \sqrt{\frac{X^2}{N(k-1)}}, \tag{5.25}$$

where k is the number of columns or rows (whichever is smaller). As with phi, the Cramér V coefficient can assume values between 0 and 1 [9, 10].

McNemar's Test

In Pearson's chi-squared test, the assumption is that the two sets of variables are different, i.e., they quantify different things; however, if we are using a repeated measure in a single population, then we cannot use the Pearson correlation test, and instead we must use the McNemar test. Essentially, the McNemar test is a test that measures consistency in responses in two categorical binary variables. For example, if we want to compare the number of positive blood samples for influenza virus in a community before and after a vaccination campaign, then we should use the McNemar test. In McNemar's test the assumption is that the two variables are paired because they are quantifying the same parameter (e.g., detection of influenza virus).

The reason for this different approach is that in paired variables there will be a degree of association and thus testing for association using a Pearson chi-squared test is not appropriate. In fact, you need to measure for disagreement between the two variables. The null/alternative hypothesis pair for McNemar's test can be stated as:

- H_0: The distribution of the responses after the intervention is the same as the distribution of the responses before the intervention (e.g., the probability of a person having a positive or negative influenza blood test is the same before and after vaccination).
- H_1: The distribution of the responses after the intervention is different from the distribution of the responses before the intervention (e.g., the probability of a person having a positive or negative influenza blood test changes after vaccination).

In the above pair, the hypotheses are two sided, i.e., any change (whether positive or negative) is considered. However, in many before-after or paired comparisons, we are only interested in one-sided change (e.g., how much the prevalence of positive influenza blood test decreases after vaccination). Thus, a one-sided hypothesis pair can be states as:

- H_0: The probability of a person having a positive or negative influenza blood test is the same before and after vaccination.
- H_1: The probability of a person having a positive influenza blood test decreases after vaccination, or the probability of a person having a negative influenza blood test increases after vaccination.

Note that McNemar's test is only applicable in 2×2 contingency tables (Table 5.5). For instances where there are more than two response categories for

Table 5.5 2×2 contingency table for McNemar's test

	Test 2 positive	Test 2 negative
Test 1 positive	Positive agreement (a)	Disagreement (b)
Test 1 negative	Disagreement (c)	Negative agreement (d)

each variable, the Cochran Q test should be used (the explanation of the Cochran Q test is beyond the scope of this book). Based on the contingency table above, the McNemar chi-squared statistics can be given by

$$X^2 = \frac{(b-c)^2}{b+c}, \tag{5.26}$$

The test statistics has a chi-squared distribution with one degree of freedom, and thus the critical values can be looked up in a chi-squared table. In simple terms, the McNemar test shows if the difference between the two tests is statistically significant, i.e., it shows whether the difference is a true effect or it occurred because of randomness.

Cochran–Mantel–Haenszel Test

"Cochran-Mantel-Haenszel test" or the CMH test for short allows us to compare two binary variables across multiple strata or matched categorical data. In McNemar's test we can only use paired data, for example, a before-after result, but CMH test is suitable if there have been multiple repeats of measurement. The reason for using the CMH test instead of multiple McNemar tests or Pearson chi-squared tests is to avoid the "Simpson paradox." Simply stated, Simpson's paradox occurs when a trend is observed in different groups or strata of data, but it either dissipates or reverses by combining these groups. In other words, by running the tests multiple times, we may wrongfully reject or fail to reject the null hypothesis.

Data stratification is the partitioning of units of observation or results into subgroups based on a factor. For example, we can stratify colon adenocarcinomas into well-differentiated, moderately differentiated, and poorly differentiated subgroups. Usually, stratification is done to control confounding variables (e.g., in case of colon adenocarcinomas, we can control for histologic grade when assessing treatment effect or survival). However, over-stratification can lead to small subgroups and consequently loss of statistical power.

For example, we want to evaluate the effectiveness of imatinib in patients with gastrointestinal stromal tumor (GIST). We have stratified the GISTs into three molecular subgroups: KIT mutants, PDGFRA mutants, and double negative group (without KIT or PDGFRA mutations). Our goal is to evaluate the response to imatinib treatment in the case group in comparison with placebo in the control group. We measure the response to treatment as a binary outcome (e.g., cured

Table 5.6 The i-th 2 × 2 contingency table showing the raw counts for the i-th stratum

$i-th$ 2 × 2 contingency table	Cured	Not cured	Total
Case (imatinib)	a_i	b_i	R_{1i}
Control (placebo)	c_i	d_i	R_{2i}
Total	C_{1i}	C_{2i}	T_i

versus not cured). In this example, the appropriate test to be used to show the treatment effect is the CMH test.

As always, the first step is to determine the null/alternative hypothesis pair:

- H_0: There is no association between the two categorical variables, i.e., they are independent (e.g., imatinib has no effect on treatment outcome for GIST patients).
- H_1: There is association between the two categorical variables (e.g., imatinib has an effect on treatment outcome for GIST patients).

The second step in the CMH test is to summarize the data for each stratum in 2 × 2 contingency tables. The $i-th$ table is shown as Table 5.6.

The CMH tests whether the combined odds ratio (R) is equal to 1 (or near 1). The further the odds ratio gets from 1 the more likely the test becomes to reject the null hypothesis. Thus, the null/alternative hypothesis pair can be restated as

- H_0: $R = 1$
- H_1: $R \neq 1$

The combined odds ratio is given by

$$R = \frac{\sum_{i=1}^{N} \frac{a_i d_i}{T_i}}{\sum_{i=1}^{N} \frac{b_i c_i}{T_i}}, \tag{5.27}$$

where N is the number of strata. To test the hypothesis, the following test formula is used:

$$\chi^2_{\text{CMH}} = \frac{\sum_{i=1}^{N} \left(a_i - \frac{R_{1i}C_{1i}}{T_i}\right)^2}{\sum_{i=1}^{N} \frac{R_{1i}R_{2i}C_{1i}C_{2i}}{T_i^2(T_i-1)}}, \tag{5.28}$$

χ^2_{CMH} follows a chi-squared distribution with one degree of freedom, and consequently the same inferences about the p-value as the chi-squared test can be made for the CMH test [11, 12].

Table 5.7 The stratified results for Example 5.3

				Response	
				Cured	Not cured
				Count	Count
Group	Case	Subgroup	Double neg	5	3
			KIT	20	7
			PDGFRA	10	5
	Control	Subgroup	Double neg	2	3
			KIT	12	21
			PDGFRA	7	5

Example 5.3

Q: Table 5.7 shows the stratified results of the study of imatinib for treatment of GIST. Calculate the χ^2_{CMH} score and determine whether imatinib is effective for treatment of GIST at significance level of 0.05.

A:

$$\chi^2_{\text{CMH}} = \frac{\sum_{i=1}^{N}\left(a_i - \frac{R_{1i}C_{1i}}{T_i}\right)^2}{\sum_{i=1}^{N}\frac{R_{1i}R_{2i}C_{1i}C_{2i}}{T_i^2(T_i-1)}} = \frac{\left(5-\frac{8\times7}{13}\right)+\left(20-\frac{27\times32}{60}\right)+\left(10-\frac{15\times17}{27}\right)}{\frac{8\times5\times7\times6}{13^2(12)}+\frac{27\times33\times32\times28}{60^2(59)}+\frac{15\times12\times17\times10}{27^2(26)}}$$

$$= 6.498, \tag{5.29}$$

Going to the chi-squared distribution table (Appendix B), we can see that the test's p-value is 0.011; thus, we can reject the null hypothesis and conclude that imatinib is effective in treating GIST. The results of stratified association tests can be shown using a clustered bar chart (Fig. 5.7).

Fisher's Exact Test

"Fisher's exact test" is a powerful tool for analysis of categorical variables and contingency tables. Most statistical tests calculate statistical significance by approximating the distribution of values to a normal distribution and thus providing an approximation of the p-value; this requires that the sample size is large enough for this approximation to be valid. Fisher's exact test, however, calculates the exact value of the p-value. The most important characteristic of this test is that it is not dependent on sample size and thus it can be used in cases where the sample size is small (making chi-squared distributed tests invalid).

Fisher's exact test is a test for independence (or association). While the test can be used for a table of any size, it is most commonly used for 2×2 contingency tables as the calculations for larger tables will be computationally taxing. The null/alternative hypothesis pair can be stated as:

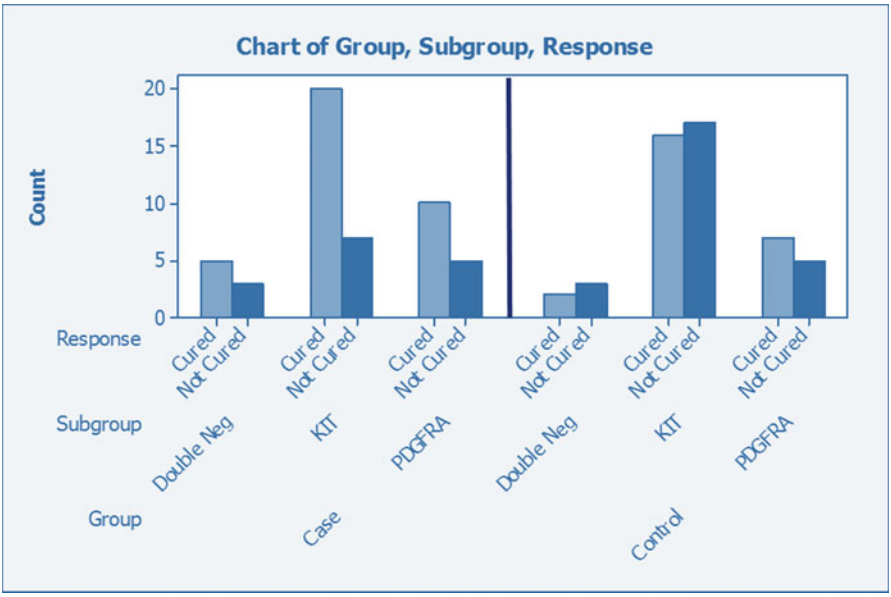


Fig. 5.7 Clustered bar chart for Example 5.3

– H_0: There is no association between the two variables (i.e., they are independent).
– H_1: There is a nonrandom association between the two variables.

Table 5.8 is a 2 × 2 contingency table which we will use for calculating the Fisher exact test.

The probability of obtaining the observed values in a contingency table follows a hypergeometric distribution (see Chap. 3). The first step in calculation is to determine the cutoff p-value (p_{cutoff}) which is the conditional probability of the observed values:

$$p_{\text{cutoff}} = \frac{\binom{r_1}{a}\binom{r_2}{c}}{\binom{N}{c_1}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!N!}, \tag{5.30}$$

The next step is to find all possible tables where the row totals (r_1, r_2) and column totals (c_1, c_2) remain constant and calculate the associated conditional probability for each table using Equation 5.30. After this, all the sum of all the p-values that are smaller than the cutoff p-value is calculated. The sum of these p-values then is the overall p-value of the test. If the overall p-value is smaller than the significance level, then we can say that the null hypothesis is rejected [13].

Table 5.8 A 2 × 2 contingency table

	Disease +	Disease −	Total
Event +	a	b	r_1 (a + b)
Event −	c	d	r_2 (c + d)
Total	C_1 (a + c)	C_2 (b + d)	N (a + b + c + d)

Table 5.9 2 × 2 contingency table for Example 5.4

	Disease +	Disease −	Total
Event +	4	0	4
Event −	1	3	4
Total	5	3	8

Example 5.4

Q: For Table 5.8 calculate the exact p-value, and determine whether the Fisher exact test is statistically significant at α of 0.05 Table 5.9.

A:First let us calculate the cutoff p-value:

$$p_{\text{cutoff}} = \frac{(4)!(4)!(5)!(3)!}{4!0!1!3!8!} = \frac{24 \times 24 \times 120 \times 6}{24 \times 1 \times 1 \times 40320} = 0.428571, \tag{5.31}$$

Now let us find all the tables that have the same row and column totals.

$$\begin{pmatrix} 1 & 3 \\ 4 & 0 \end{pmatrix}, \begin{pmatrix} 2 & 2 \\ 3 & 1 \end{pmatrix}, \begin{pmatrix} 3 & 1 \\ 2 & 2 \end{pmatrix}$$

The conditional probability for each of these tables is 0.071428, 0.428571, and 0.428571, respectively. Only the first conditional probability is smaller than the cutoff p-value; thus, the sum of p-values will be 0.071428. Because this p-value is larger than the significance level, then we have failed to reject the null hypothesis, i.e., there is no association between the event status and disease status. Remember that this is a one-tailed p-value. If we are interested in two-tailed p-value, then we must double the sum of p-values (e.g., in this case the two-tailed p-value is 0.1429).

Generally, if the expected count for a cell is below 5, it is advisable to avoid chi-squared tests and use Fisher's exact test.

Measures of Agreement

Sometimes the reason for analysis of a contingency table is to determine the degree of agreement between the sets. In pathology, measuring agreement is used in test validation as well as for determining inter- and intraobserver reliability and agreement. For example, when two pathologists look at the same histopathology slide, do they come to the same conclusion? One of the commonly used statistics for measuring agreement is the "Kappa coefficient."

Table 5.10 2×2 contingency table for measuring agreement

	Yes	No	Total
Yes	a	b	r_1 (a + b)
No	c	d	r_2 (c + d)
Total	C_1 (a + c)	C_2 (b + d)	N (a + b + c + d)

Table 5.11 Levels of agreement based on Kappa coefficient

Kappa coefficient	Level of agreement
0–0.20	Poor
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Good
0.81–1	Excellent

Cohen's Kappa

"Cohen's Kappa" is a relatively simple statistic to calculate, yet it is powerful metric since it considers the possibility that the agreement has occurred by chance. Cohen's Kappa is calculated by comparing the observed proportionate agreement (p_0) with the overall probability of random agreement (p_e). For these calculations, we are going to use a 2×2 contingency table (Table 5.10).

The observed proportionate agreement (p_0) is given by

$$p_0 = \frac{a+d}{N}, \tag{5.32}$$

The overall probability of random agreement (p_e) is given by

$$p_e = \frac{(r_1 c_1) + (r_2 c_2)}{N^2}, \tag{5.33}$$

The Cohen Kappa coefficient can then be calculated using:

$$\kappa = \frac{p_0 - p_e}{1 - p_e}, \tag{5.34}$$

If the observed agreement is more than chance alone, then the Kappa coefficient will have values of between 0 and +1. The closer the Kappa coefficient gets to +1, the stronger the agreement becomes. Table 5.11 is a rule of thumb index for degrees of agreement based on Kappa coefficient.

Example 5.5

Q: Two pathologists are reviewing voided urine cytopathology slides for the presence of high-grade urothelial carcinoma. Table 5.12 summarizes their results. Calculate the Cohen Kappa coefficient for agreement between the two pathologists.

A: Let us calculate the p_0 and p_e first.

Table 5.12 2 × 2 contingency table of agreement for Example 5.5

		Pathologist 1		
		Positive for high-grade urothelial carcinoma	Negative for high-grade urothelial carcinoma	Total
Pathologist 2	**Positive for high-grade urothelial carcinoma**	23	4	27
	Negative for high-grade urothelial carcinoma	5	33	38
	Total	28	37	65

$$p_0 = \frac{a+d}{N} = \frac{23+33}{65} = 0.8615, \tag{5.35}$$

$$p_e = \frac{(r_1c_1)+(r_2c_2)}{N^2} = \frac{(27\times 28)+(38\times 37)}{4225} = 0.5117, \tag{5.36}$$

Now we can calculate the Cohen Kappa coefficient:

$$\kappa = \frac{p_0 - p_e}{1-p_e} = \frac{0.8615 - 0.5117}{0.4883} = 0.7163, \tag{5.37}$$

The results show that there is good agreement between the two pathologists.

If the categorical variables are paired, then the Kappa coefficient will overestimate the agreement; in these situations, McNemar's test with a null/alternative hypothesis pair of homogeneity should be used.

Fleiss's Kappa

Cohen's Kappa coefficient is only used in situations when there are two raters. If there are more than two raters or more than two categories for the binary variable being compared, then another measure of agreement called the "Fleiss Kappa coefficient" should be calculated. As before, the coefficient is a measure of overall consistency in rating units of observation. The overall Kappa formula is the same as Cohen's Kappa, but p_0 and p_e are calculated differently (here $\bar{P}_0$ *and* $\bar{P}_e$ since they are the means of probabilities).

$\bar{P}_0$ is given by

$$\bar{P}_0 = \frac{1}{Nn}(n-1)\left(\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij}^{\ 2} - Nn\right), \tag{5.38}$$

where N is the total number of subjects, n is the number of rating per subject, and k is the number of categories. i denotes the index of subjects and j is index of

categories. Thus, n_{ij} is the number of raters who assigned the *i-th* subject to the *j-th* category.

$\bar{P}_e$ is given by

$$\bar{P}_e = \sum_{j=1}^{k} P_j^2, \tag{5.39}$$

where P_j is the proportion of all observations (n/N) that were assigned to the *j-th* category (i.e., column totals divided by all observations).

The interpretation of Fleiss's Kappa coefficient is like Cohen's coefficient (see Table 5.11) [14, 15].

Summary

In this chapter, we introduced the concepts of statistical inference for categorical (nominal variables). The general approach for any statistical inference test is to formulate the null/alternative hypothesis pair. Choose an appropriate statistic to test the hypothesis. Organize the data into appropriate format. Calculate the statistic and interpret the results based on the hypotheses and the significance level.

Table 5.13 summarizes the tests introduced in this chapter [16].

Table 5.13 Summary of statistical tests introduced in this chapter

Test	Hypothesis testing	Condition (s)	Example (s)
Pearson's chi-squared test	Test for independence Test for homogeneity Test for goodness of fit	The data should be organized into a contingency table. The measurements should not be paired or matched. Sample size should be sufficiently large	Are test results associated with a disease status? Do patients with and without disease X have similar qualitative test results?
McNemar's test	Testing for change in paired data	The data should be paired. Can only be used for 2×2 contingency tables	Do the results of a binary response qualitative test differ in a patient before and after an intervention?
Cochran-Mantel-Haenszel test	Test for association	The data is stratified or there are multiple repeated measures	Are test results associated with a disease status among both men and women?
Fisher's exact test	Test for association	The test is not chi-square distributed, so it can be used for small sample sizes	Are test results associated with a disease status?
Kappa coefficient	N/A	The contingency table is of raters' categories against each other	How much is the inter-rater reliability for a pathologic diagnosis?

References

1. Strike PW. Statistical methods in laboratory medicine. New York: Butterworth-Heinemann; 2014.
2. Elliott AC, Woodward WA. Statistical analysis quick reference guidebook: with SPSS examples. Thousand Oaks: Sage; 2007.
3. Agresti A, Kateri M. Categorical data analysis. Berlin Heidelberg: Springer; 2011.
4. Fisher RA. On the interpretation of X^2 from contingency tables, and the calculation of P. J R Stat Soc. 1922;85(1):87–94.
5. Simpson EH. The interpretation of interaction in contingency tables. J R Stat Soc Ser B Methodol. 1951;13:238–41.
6. Wilson EB, Hilferty MM. The distribution of chi-square. Proc Natl Acad Sci. 1931;17 (12):684–8.
7. Eisenhauer JG. Degrees of freedom. Teach Stat. 2008;30(3):75–8.
8. Sharpe D. Your chi-square test is statistically significant: Now what? Practical Assessment, Research & Evaluation. 2015;20:1–10.
9. Scheaffer RL, Yes N. Categorical data analysis: NCSSM Statistics Leadership Institute, USA; 1999. (online publication accessible at: http://courses.ncssm.edu/math/Stat_Inst/PDFS/Categorical%20Data%20Analysis.pdf)
10. Fleiss JL. Categorical Data Analysis. J Am Stat Assoc. 1991;86(416):1140–1.
11. Mantel N. Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. J Am Stat Assoc. 1963;58(303):690–700.
12. Trajman A, Luiz RR. McNemar X^2 test revisited: comparing sensitivity and specificity of diagnostic examinations. Scand J Clin Lab Invest. 2008;68(1):77–80.
13. Routledge R. Fisher's exact test. In: Encyclopedia of biostatistics. New York: John Wiley Publishing; 2005.
14. Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull. 1968;70(4):213.
15. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas. 1973;33(3):613–9.
16. Zhou XH, McClish DK, Obuchowski NA. Statistical methods in diagnostic medicine. John Wiley & Sons: New York; 2009.

Comparing Sample Means

6

Introduction

In the previous chapter, we learned about statistical inference for discrete variables. In pathology and laboratory medicine, however, there are many tests that provide quantitative and continuous results, examples of which include complete blood count, metabolite levels, enzyme activity, and so on. In fact, many qualitative, categorical test results are derived from quantitative results that are dichotomized for ease of interpretation.

Consequently, understanding of statistical inference for quantitative continuous variables is of utmost importance for practicing pathologists. They can use these statistical tests for many different applications, for example, for assessing the performance of one assay against another assay or for comparing the results of their lab to the results from the standard lab.

In this chapter, we explain some of the more important statistical tests that are used for continuous variables. We begin the chapter with introducing a few fundamental concepts such as continuous data, goodness of fit (also discussed in Chap. 4), and parametric and non-parametric testing and then move on introducing each of the tests.

Continuous Data

"Continuous variables" can assume any value in a real number interval. In continuous variables, there is no real limit to accuracy, i.e., they can assume any real number within their range. The data in continuous variables is only limited by the accuracy of measurement, documentation, and reporting. In other words, the range of values for continuous variables has no gaps no matter how infinitesimally small, and there are infinite numbers of real numbers between any two real numbers in the range of the variable, hence the term continuous.

A. Momeni et al., *Introduction to Statistical Methods in Pathology*,
DOI 10.1007/978-3-319-60543-2_6

As previously discussed in Chap. 3, measuring probabilities in continuous variables means measuring the probabilities of intervals, and this is done using the probability density function (PDF). For probability measurement in continuous intervals, a curve is fit to the data, and the area under the curve represents the probability which must always equal to 1, i.e., the entire area under the probability distribution curve has a probability of 1. Probabilities of intervals are calculated by PDF using the area under curve that corresponds to that interval. The cumulative distribution function (CDF) can calculate the cumulative probability of values smaller than a given cutoff. The CDF function has an important role in calculating *p*-values.

Mean and Median

Continuous data have two main summary location measures: "mean" and "median."

Mean or expected value (μ or $E(x)$) is the long-run average value of the random continuous variable and can be calculated using the following formula:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx \tag{6.1}$$

Example 6.1
Q: Assuming X is a random continuous variable with the following PDF, calculate the mean of X.

$$f(x) = \begin{cases} \dfrac{x}{5} & 0 \le x \le 3 \\ 0 & \text{for all other values} \end{cases} \tag{6.2}$$

A:

$$E(X) = \int_0^3 xf(x)dx = \int_0^3 x \times \frac{x}{5}dx = 0.2\int_0^3 x^2dx = \frac{9}{5} = 1.8 \tag{6.3}$$

Percentiles are values of a continuous random variable that bound a corresponding area under the probability distribution curve. For example, the fifth percentile is the value of the random variable, the cumulative distribution function of which is 0.05.

"Median" represents the 50th percentile of a random variable. We can say that the median is the middle value of the distribution.

Figure 6.1 shows the median and mean of a positively skewed continuous variable.

Median is a more stable statistic of a distribution compared to mean. Changes in the distribution and departures from normality tend to change mean more than

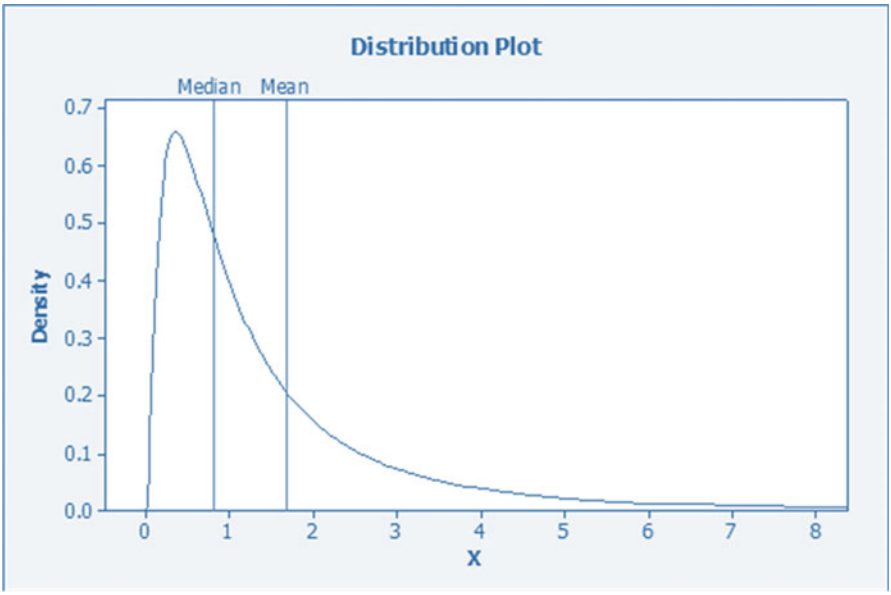


Fig. 6.1 Depiction of Median and Mean in a positively skewed distribution. Note that while in this distribution the mean is to the right of the median, this is not always true for skewed distributions

median. Hence, in normally distributed data, statistical tests that use mean as a metric are used, while in skewed distributions, median is preferred.

Note that in a perfectly symmetrical data distribution, the mean and median are equal.

Variance, Skewness, and Kurtosis

Variance, skewness, and kurtosis are high-degree central moments (see Chap. 3) of a random variable. These measures contain information about the shape of the distribution of the random variable. There are higher-degree central moments as well (e.g., hyperskewness and hyperflatness), but they are beyond the scope of this book.

Variance (σ^2 or s^2) is the second central moment and shows the dispersion of the data around the central location (mean). Variance of a random variable (X) is the expected value of the squared deviation from the mean, and it is always a nonnegative value:

$$\sigma^2 = E\left[(X-\mu)^2\right] \tag{6.4}$$

As the variance increases, the data will become more dispersed, and the distribution curve becomes flatter.

For continuous random variables, the variance is calculated using definite integrals:

$$\sigma^2 = \int x^2 f(x) dx - \mu^2 \tag{6.5}$$

The standard deviation (σ) is the positive square root of variance.

Skewness is a measure of asymmetry of the probability distribution and is the third central moment. A positively skewed distribution will either have the tail of the distribution longer on the right side or the bulk of distribution on the left side (or both). Conversely, a negatively skewed distribution will either have the tail of the distribution on the left side or the bulk of the distribution on the right side (or both). Symmetric distributions have a skewness of zero, while a skewness of zero does not necessarily imply a symmetrical distribution as the asymmetries on either side can cancel each other out.

Kurtosis is the fourth central moment and is a measure of the shape of the tails of the random variable distribution. High kurtosis means that the tails of the distribution are heavy, meaning that the peak of the distribution is narrower and more of the distribution falls under the tails of the distribution. Low kurtosis distributions have light tails, meaning that more of the distribution falls under the peak.

Calculation of skewness and kurtosis is beyond the scope of this book. The main applications of skewness and kurtosis are in some tests for normality (see below) [1, 2, 5–9].

Parametric Versus Non-parametric Tests

When deciding which statistical tests to use for continuous data, one of the first decisions is the choice between parametric and non-parametric tests. Parametric tests are based on assumptions about the data especially the distribution of the values; generally, parametric tests should be used when the data is known to follow a normal distribution. Non-parametric tests on the other hand make no assumptions about the distribution of the data.

The parametric tests that we will cover in this chapter include t-test and ANOVA.

Parametric tests analyze group means, and their hypothesis testing is about finding difference in the mean value. This has been shown to give parametric tests a greater statistical power compared to non-parametric tests, i.e., these tests are more likely to detect a significant effect when it truly exists.

Non-parametric tests analyze group medians. This is especially important in highly skewed data distributions in which median is a better measure of central tendency compared to mean (see above). In fact, many laboratory tests are highly skewed and follow log-normal distribution rather than normal distribution. Take liver function tests, for example. These tests have a lower threshold, but there is no

real upper limit, while in healthy individuals they are likely to be bound within a normal range, but some patients and healthy individuals have higher liver function tests. In other words, we are more likely to see a very high AST than a very low AST.

The non-parametric tests that we will cover in this chapter include Mann-Whitney test and Kruskal-Wallis test.

However, and because of the central limit theorem (see Chap. 3) as the sample size increases, the distribution of data approximates a normal distribution. Thus, even for data that is not normally distributed, we can treat the data as normally distributed and use parametric tests for their analysis. In fact, and as a rough guide, for a one-sample t-test, if the sample size is greater than 20, we can use abnormally distributed data as well. For a two-sample t-test and one-way ANOVA, the sample size in each group should be greater than 15 (in ANOVA if you have more than ten groups, the sample size in each group should be greater than 20). So, if the sample size is small and you are uncertain about the normal distribution of the data, you can use non-parametric tests; otherwise, you can still use parametric tests.

One of the problems with non-parametric tests is that their basic assumption is that the dispersion (spread) of the data should be similar in the groups. In parametric tests, however, different spreads (different variances) are tolerated.

On the other hand, only continuous data without significant outliers can be used with parametric tests, while some non-parametric tests can handle outliers and, in addition, ordinal and ranked data.

The parametric tests and their non-parametric counterparts (with respective null/alternative hypothesis) that are covered in this chapter are summarized in Table 6.20 in the summary section of this chapter [10–12].

Outliers

There are circumstances when some of the variables have values that are too distant from other observations to fit into the distribution of interest. For example, suppose we are measuring the sodium level in the sera of individuals and we have established that 95% of individuals have sodium values that fall between 135 and 145 mEq/L. If we have one sample that is diluted with water and has a sodium value of 115 mEq/L, then this sample is an outlier (Fig. 6.2).

It is very important to determine if the outliers are due to true variability or if they have occurred because of a measurement error. If the outlier is due to a measurement error, then it can be ignored. In laboratory medicine, if an outlier is identified, the routine practice is to repeat the measurement: if the value remains the same, then it is attributed to variability. If the value changes, then the original measurement is considered as an error and is ignored. Checking for outliers is either done using control samples (which have a known distribution) or by performing a "delta check", in which the patient's results are compared with their previous results. We will discuss outliers more in Chap. 10.

When the data contains outliers, it is usually more appropriate to use non-parametric tests.

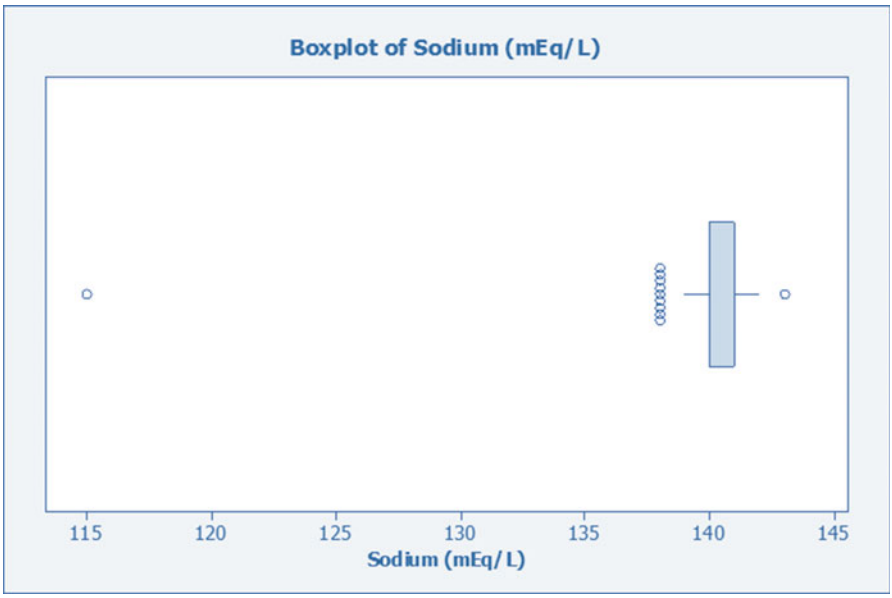


Fig. 6.2 Boxplot showing sodium concentration in the serum. Note the outlier at the extreme left of the diagram. A boxplot shows the distribution of continuous data. The box (*the rectangle*) usually contains the central three quartiles of data with the *horizontal lines* extending to 10th and 90th percentiles of the data. The *circles* represent outliers in the data

One-Tailed Versus Two-Tailed Testing

In the previous chapter, we explored the *p*-value; here, we will discuss a very important characteristic of tests that relates to *p*-value calculation and interpretation: one-tailed or two-tailed test. Ignoring this concept can lead to misinterpretation of test results and to misinformed conclusions drawn from the *p*-value.

When dealing with a one-tailed test, the test allots all the α to statistical significance in one direction of the test. Essentially, a one-tailed test is a directional test that not only shows that the sets of variables compared are different but that the difference is in one direction. For example, if we want to compare the performance of a new test with another test and our objective is to show that the new test is better (superior) than the other test, then a one-tailed test is desired. The types of studies that rely on one-tailed tests are said to follow a superiority design (see Chap. 12). If the effect of the test is only important in one direction, then a one-tailed test in that direction should be employed as one-tailed tests tend to have more power in detecting an effect. But this may cause effects in the opposite direction to the one-tailed test to be missed, and this can be serious, for example, when testing if a new test is better than another test, if a one-tailed test is used, detection of an effect means that the new test is better, but, if no effect is detected, it does not mean that the new test is equal to the other test; in fact, it may mean that the new test is performing worse than the other test.

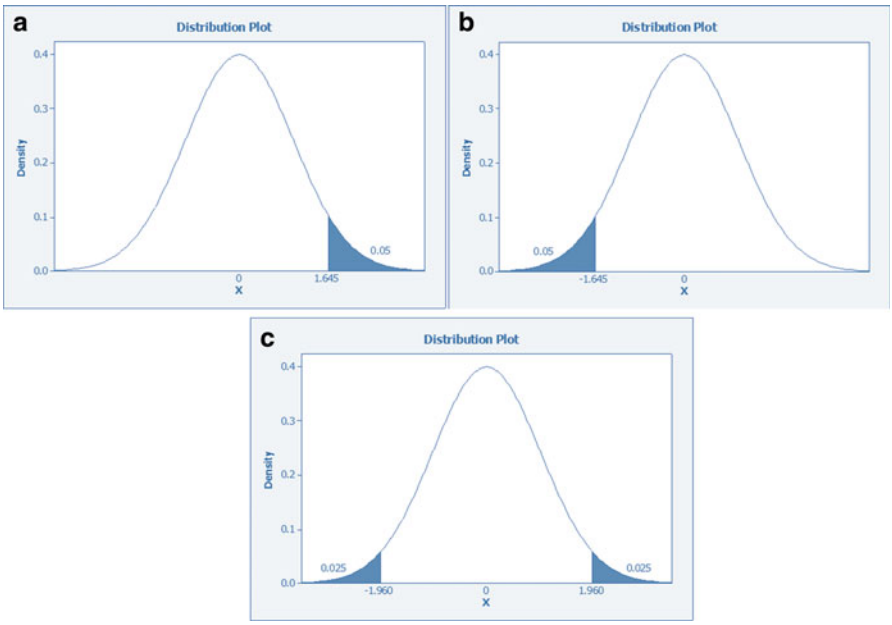


Fig. 6.3 One-tailed tests with significance level of 0.05 are shown with either left-sided or right-sided testing (**a**, **b**). A significance level of 0.05 for a two-tailed test divides the significance to either end of the test statistic distribution (**c**)

The direction of the one-tailed test can be in either side of test statistic distribution (Fig. 6.3). When the direction is toward the lower extreme of the distribution, the test is useful for noninferiority trial designs (e.g., where the objective is to show that a new test is not performing worse than a current test). When the direction is toward the higher end of the distribution, the test is useful for superiority trial designs.

The two-tailed test allots the α to the two sides of test statistic distribution. In effect, in two-tailed tests, if the test statistic value is in either extreme of the distribution, then the null hypothesis is rejected. If significance level is set at 0.05, then in a two-tailed test, 0.025 cutoff at either extreme of the distribution is considered as statistical significance. Two-tailed tests can be used in equivalence trials (e.g., where the objective is to show that the performance of two diagnostic tests is similar). However, they are generally preferred by statisticians, and, in most statistical software, the tests are set to two-tailed testing by default. Overall, two-tailed testing is a more robust statistical analysis.

Unless stated otherwise, we have assumed a two-tailed test in explaining the concepts in this chapter.

Testing for Normality

Sometimes, it is advisable to determine the distribution of the variable we want to use in a statistical test. While, in many instances, we can assume normal or near-normal distribution, testing for normality can verify if a variable is normally distributed. This is an important step before attempting to use parametric tests to see if the data satisfy the requirements of those tests. In pathology and laboratory medicine, this is even more critical as we determine reference intervals with the assumption that the diagnostic test results follow a Gaussian distribution in the population (see Chap. 3).

While the central limit theorem allows us to assume that the distribution is near normal for large sample sizes (>30), it is still a good practice to check for normality.

The easiest way to judge the normality of distribution is to use normality distribution plots (see Chap. 3) and check whether the plotted distribution forms a straight line. A Q-Q plot comparing the quantile distribution of the variable against quantile distribution of a variable which we know is normally distributed is another way to visually inspect for normal distribution. These methods, however, suffer from inaccuracy as subjective inspection of the probability distribution plots may not be able to detect small deviations from normality.

There are many statistical tests that test for normality. These include the Jarque-Bera test, D'Agostino-Pearson omnibus test, Shapiro-Wilk test, Kolmogorov-Smirnov test, and the W/S test. Here we will discuss the latter two tests.

"W/S test" is a simple test for kurtosis, q, that is the ratio of the range of values, W, to the standard deviation, s, i.e.,

$$q = \frac{w}{s} \tag{6.6}$$

q values for each sample size has a critical range, and, if the calculated q values fall within that range, then we can say that the data is normally distributed. Table 6.1 lists the critical values of q for different sample sizes (up to 50) and significance levels.

Example 6.2
Q: Table 6.2 shows the AST results from a sample of 14 patients. The standard deviation of the results is 7.35. Determine whether the results are normally distributed at significance level of 0.05.

A:

$$q = \frac{w}{s} = \frac{24}{7.35} = 3.265 \tag{6.7}$$

Going to Table 6.1, we can see that the critical range for a sample size of 15 at significance level of 0.05 is from 2.97 to 4.17, and since 3.265 is within this critical

Table 6.1 The critical values for "q values" for different sample sizes at different significance levels. "a" is the lower boundary of the critical range and "b" is the upper boundary of the critical range

	Alpha = 0.000		Alpha = 0.005		Alpha = 0.01		Alpha = 0.05	
Sample size	a	b	a	b	a	b	a	b
3	1.732	2.000	1.735	2.000	1.737	2.000	1.758	1.999
4	1.732	2.449	1.82	2.447	1.87	2.445	1.98	2.429
5	1.826	2.828	1.98	2.813	2.02	2.803	2.15	2.753
6	1.826	3.162	2.11	3.115	2.15	3.095	2.28	3.012
7	1.871	3.464	2.22	3.369	2.26	3.338	2.40	3.222
8	1.871	3.742	2.31	3.585	2.35	3.543	2.50	3.399
9	1.897	4.000	2.39	3.772	2.44	3.720	2.59	3.552
10	1.897	4.243	2.46	3.935	2.51	3.875	2.67	3.685
11	1.915	4.472	2.53	4.079	2.58	4.012	2.74	3.80
12	1.915	4.690	2.59	4.208	2.64	4.134	2.80	3.91
13	1.927	4.899	2.64	4.325	2.70	4.244	2.82	4.00
14	1.927	5.099	2.70	4.431	2.75	4.34	2.92	4.09
15	1.936	5.292	2.74	4.53	2.80	4.44	2.97	4.17
16	1.936	5.477	2.79	4.62	2.84	4.52	3.01	4.24
17	1.944	5.657	2.83	4.70	2.88	4.60	3.06	4.31
18	1.944	5.831	2.87	4.78	2.92	4.67	3.10	4.37
19	1.949	6.000	2.90	4.85	2.96	4.74	3.14	4.43
20	1.949	6.164	2.94	4.91	2.99	4.80	3.18	4.49
25	1.961	6.93	3.09	5.19	3.15	5.06	3.34	4.71
30	1.966	7.62	3.21	5.40	3.27	5.26	3.47	4.89
35	1.972	8.25	3.32	5.57	3.38	5.42	3.58	5.04
40	1.975	8.83	3.41	5.71	3.47	5.56	3.67	5.16
45	1.978	9.38	3.49	5.83	3.55	5.67	3.75	5.26
50	1.980	9.90	3.56	5.93	3.62	5.77	3.83	5.35

range, then we can conclude that the data is normally distributed. Fig. 6.4 is a normal distribution plot of the data.

"Kolmogorov-Smirnov test" (K-S test) can be used to determine if a sample has a specific distribution. The K-S test is the most commonly used test for testing goodness of fit for normal distribution. This test is based on the "empirical distribution function" (EDF). To define the EDF, all the data points in the sample should be ordered from smallest to largest value. For example, if X is a random variable with N values, the variable values should be ordered from x_1 to x_n. Then EDF is a stepwise function, and at each step, it can be calculated as

$$F_n(x) = \frac{n(i)}{N} \tag{6.8}$$

Table 6.2 The table of AST results for Example 6.2

Patient number	AST level (units/L)
1	32
2	28
3	33
4	34
5	40
6	28
7	33
8	37
9	50
10	36
11	52
12	43
13	29
14	37
15	42

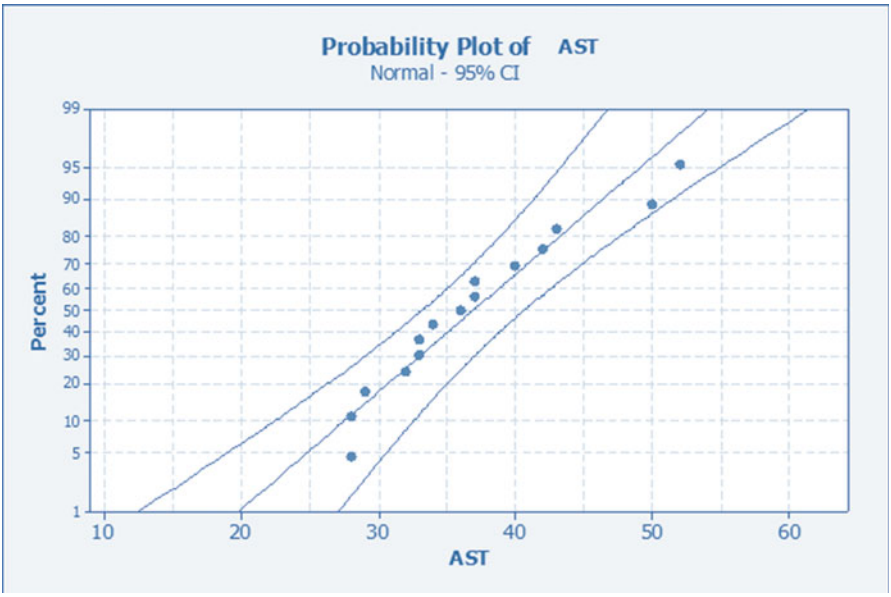


Fig. 6.4 The normal distribution plot for Example 6.2

where $n(i)$ for each x_i is the number of values less than x_i. The EDF is stepwise, in the sense that for each increase from one ordered value to the next, the value of EDF increases by $1/N$. EDF is similar to cumulative distribution function in the sense that the sum of all the EDF values over the range of the variable equals 1.

Now the K−S test (D_n) can be calculated by comparing the EDF function of the variable with the cumulative distribution function of the desired distribution (usually Gaussian normal distribution).

$$D_n = \max_{1 \le i \le N} \mid F_n(x) - F(x) \mid \tag{6.9}$$

where $F(x)$ is the CDF of the desired distribution. Simply stated, the computed value of the K−S test is the maximum positive value that the difference of estimated distribution function with CDF can take for all the points in the range of the sample. This calculated value can then be looked up in a table of critical values for the corresponding sample size and desired significance level (Table 6.3). If the test value is less than the corresponding critical value, then the sample has a distribution like our chosen distribution.

The K−S test should only be used for continuous variables (discrete variables should be adapted into a continuous distribution). One limitation of the K−S test is that it is more sensitive to differences in distribution near the center of the data distribution which renders the test less sensitive to problems in the tails. Also, it is generally recommended that the K−S test is used for larger datasets (where $N > 50$). For smaller datasets, the Shapiro-Wilks test can be used.

Table 6.3 The critical values for "D values" for different sample sizes at different significance levels

N	0.10	0.05	0.01
1	0.950	0.975	0.995
2	0.776	0.842	0.929
3	0.642	0.708	0.828
4	0.564	0.624	0.733
5	0.510	0.565	0.669
6	0.470	0.521	0.618
7	0.438	0.486	0.577
8	0.411	0.457	0.543
9	0.388	0.432	0.514
10	0.368	0.410	0.490
11	0.352	0.391	0.468
12	0.338	0.375	0.450
13	0.325	0.361	0.433
14	0.314	0.349	0.418
15	0.304	0.338	0.404
16	0.295	0.328	0.392
17	0.286	0.318	0.381
18	0.278	0.309	0.371
19	0.272	0.301	0.363
20	0.264	0.294	0.356
25	0.240	0.270	0.320
30	0.220	0.240	0.290
35	0.210	0.230	0.270
OVER 35	$\frac{1.22}{\sqrt{N}}$	$\frac{1.36}{\sqrt{N}}$	$\frac{1.63}{\sqrt{N}}$

The Anderson-Darling test is an improved version of the K−S test that is available in some statistical software and is usually more powerful than the K−S test.

A two-sample Kolmogorov-Smirnov test can be used to determine if two samples have a similar distribution by comparing the EDF functions of the two samples. The critical value for two sample K−S test is calculated based on the sample size and the desired significance level:

$$\text{Critical value} = C\sqrt{\frac{N_1 + N_2}{N_1 N_2}} \tag{6.10}$$

where C is a constant that equals 1.36 for a 0.05 significance level and 1.63 for a 0.01 significance level. If the calculated K−S test value is smaller than the critical value, then the two variables have a similar distribution [3, 4].

Example 6.3

Q: Table 6.4 shows the results of AST and ALT values from a sample of 20 patients. The results are ordered from the smallest to largest. Do AST and ALT have a similar distribution at a significance level of 0.05?

A: Table 6.5 lists the EDF values of the AST and ALT among the range of their sample, and for each pair of EDF values, the corresponding difference is shown. The maximum value (D_n) is in bold face.

Table 6.4 Table of values for Example 6.3

Sample number	AST (units/L)	ALT (units/L)
1	27	25
2	30	26
3	31	30
4	31	30
5	33	32
6	36	33
7	38	33
8	39	37
9	40	38
10	41	39
11	41	40
12	41	41
13	43	44
14	44	44
15	48	47
16	49	49
17	50	50
18	51	51
19	52	52
20	53	52

Table 6.5 EDF values for AST and ALT and the corresponding difference of the values. The maximum difference is in bold face

Sample number	AST (units/L)	ALT (units/L)	EDF_{AST}	EDF_{ALT}	$\lvert EDF_{AST} - EDF_{ALT} \rvert$
1	27	25	0	0	0
2	30	26	1/20	1/20	0
3	31	30	2/20	2/20	0
4	31	30	2/20	2/20	0
5	33	32	4/20	4/20	0
6	36	33	5/20	5/20	0
7	38	33	6/20	5/20	1/20
8	39	37	7/20	7/20	0
9	40	38	8/20	8/20	0
10	41	39	9/20	9/20	0
11	41	40	9/20	10/20	1/20
12	**41**	**41**	**9/20**	**11/20**	**2/20**
13	43	44	12/20	12/20	0
14	44	44	13/20	12/20	1/20
15	48	47	14/20	14/20	0
16	49	49	15/20	15/20	0
17	50	50	16/20	16/20	0
18	51	51	17/20	17/20	0
19	52	52	18/20	18/20	0
20	53	52	19/20	18/20	1/20

The K−S test value is 2/20 (0.1). The critical value for the sample size and significance level is given by

$$\text{Critical value} = C\sqrt{\frac{N_1 + N_2}{N_1 N_2}} = 1.36\sqrt{\frac{20 + 20}{20 \times 20}} \cong 0.430 \tag{6.11}$$

Since the calculated K−S value (0.1) is smaller than the critical value (0.43), then AST and ALT can be said to follow a similar distribution.

The EDF values can also be plotted. The plot is like a CDF plot, and in fact, in one-sample K−S test, the sample EDF is plotted against the CDF of the desired distribution. Figure 6.5 shows the plot for Example 6.3.

Parametric Tests

Parametric statistics is one of the most common statistical tests used; this is because they are easier to compute and have more statistical power. As with any other statistical test, the first step is to formulate the null/alternative hypotheses pair. Based on the hypotheses pair and nature of the data, a decision can be made over which statistical test to employ.

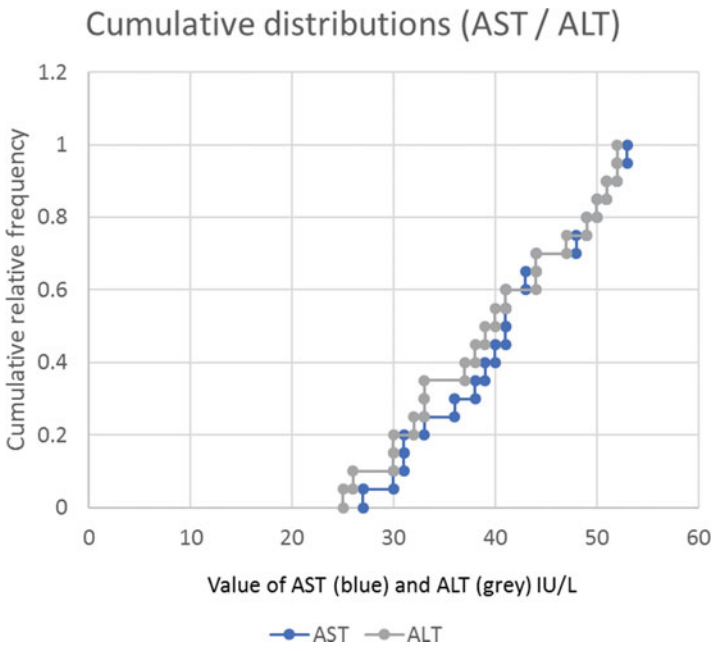


Fig. 6.5 Plot of EDFs for AST and ALT from Example 6.3

All parametric tests work by comparing the location (mean) of data. The null hypothesis generally is that the location of the data is similar between either the samples or the sample and a hypothetical distribution. The alternative hypothesis states that the samples have different locations. In other words, parametric tests aim to determine whether two or more sets of continuous data are significantly different from each other. These tests also make assumptions about data, for example, the assumption for the t-test is that the data are normally distributed.

For example, if we are interested in cholesterol levels in patients with and without coronary artery disease (CAD), we can use a parametric test. The null hypothesis would be that cholesterol levels are similar in patients with and without CAD. The alternative hypothesis here will be that there is a statistically significant difference in cholesterol levels in patients with and without CAD.

Here we will discuss two parametric statistical tests: "t-test" and "analysis of variance." In short, the t-test is applicable in cases where there are one or two groups. For comparisons of three or more groups, analysis of variance should be employed.

Student's t-Test

The so-called student's t-test was invented by William Sealy Gosset in the early twentieth century who published his method under the pseudonym, "student." This

"t-test" is used to determine whether the mean value of the data for a group is significantly different from the mean of another group or from a specific value. In cases where the mean of the group is compared to a specific mean (hypothesized mean), a one-sample t-test is used, and in comparisons between groups, two-sample t-test is used.

Student's t-test is based on calculation of a test statistic called "t-value." This value is a standardized variable with a standardized distribution. Thus, since the distribution of t-value is known and based on the location of the calculated t-value on the t-distribution curve, a *p*-value can be calculated. This concept is similar to computation of *p*-values from the chi-squared score (see Chap. 5).

Note that, if we are dealing with population-sized samples, i.e., large samples, then we can use z-test instead of the t-test. The fundamental difference is that we look for critical values of z-distribution instead of t-distribution.

Before we show you the calculations for t-test, we will introduce the concept of t-distribution.

T-Distribution

"T-distribution" is a well-known and documented distribution. The distribution of t-values is mainly determined by the degree of freedom (D.F. or v).

Simply stated, degree of freedom is dependent on the number of parameters in the equation and the sample size. As the number of parameters increases, you will have a lower degree of freedom meaning that the accuracy of your assumptions decreases. However, any increase in the sample size can offset the increase in the number of parameters. For example, a one-sample t-test, with N being the sample size, has one parameter, and thus one degree of freedom is spent calculating the mean, and the remainder of the sample ($N-1$) is used for estimating variability of the data. For a two-sample t-test, there are $N-2$ degrees of freedom for variability and error. Essentially, an increase in the number of parameters is a cost that needs to be offset from the sample size.

T-distribution resembles a normal distribution in that it has a bell-shaped distribution curve which is symmetrical around its mean (which is 0). The difference between a t-distribution and normal distribution is that the sample for a normal distribution is population sized, but, for the t-distribution, the sample is a subset of the population, i.e., the sample size is usually small. Thus, the t-distribution curve has a narrower peak and heavier tails compared to normal distribution. Figure 6.6 plots the t-distribution curve for different degrees of freedom.

As the sample size increases, the t-distribution plot will approximate normal distribution plot. With sample sizes of more than 20, the t-distribution is like a normal distribution for practical purposes.

Let us examine an example. We assume that the population mean for sodium is 140 mEq/L. We have a sample of 20 patients, and we want to see if the mean sodium of this sample is the same as the population mean (i.e., a one-sample t-test) at a significance level of 0.05. This means that we have a t-distribution with 19 degrees of freedom ($N-1$). This is a two-tailed test since we are interested if the means are equal (not if one is greater than the other). Thus, the significance level

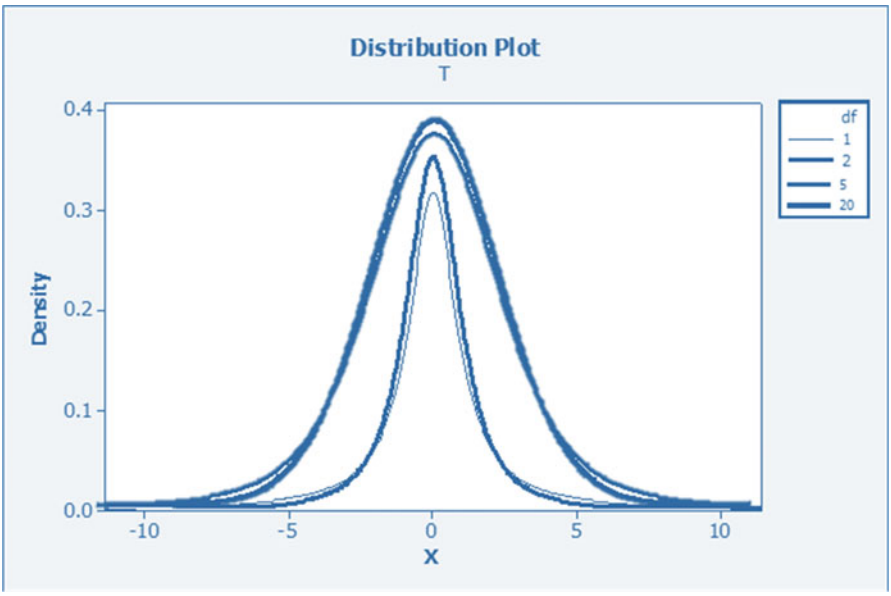


Fig. 6.6 Plots of t-distribution based on degrees of freedom

of 0.05 will mean the extreme 0.025 of either tail of the distribution curve; this translates to a t-value of 2.093 and −2.093 (these values are looked up in the t-score critical values table). If the t-value of the one-sample t-test is 1.5, then we can say that the sample mean is like the population mean, because to get a p-value of less than 0.05, we need t-values of either more than 2.093 or less than −2.093. This is shown in Fig. 6.7.

Cutoff values for the t-distribution for different degrees of freedom at 0.05 and 0.01 significance levels are provided in Table 6.6. These cutoff values essentially describe a 95% and 99% confidence interval for the t-distribution. As you can see, as the degree of freedom increases, the cutoffs get closer to that of a normal distribution (from Chap. 2, we know that 95% and 99% confidence intervals are 1.96σ and 2.58σ, respectively).

A comprehensive table of critical t-values is given in Appendix C.

One-Sample t-Test

One sample t-test is used when we want to determine if the mean of a sample (μ) is similar or different to a hypothesized mean (m_0). Thus, the null/alternative hypotheses pair can be stated as:

- H_0: the mean of the sample is equal to the hypothesized mean ($\mu = m_0$).
- H_1: the mean of the sample is different from the hypothesized mean ($\mu \neq m_0$).

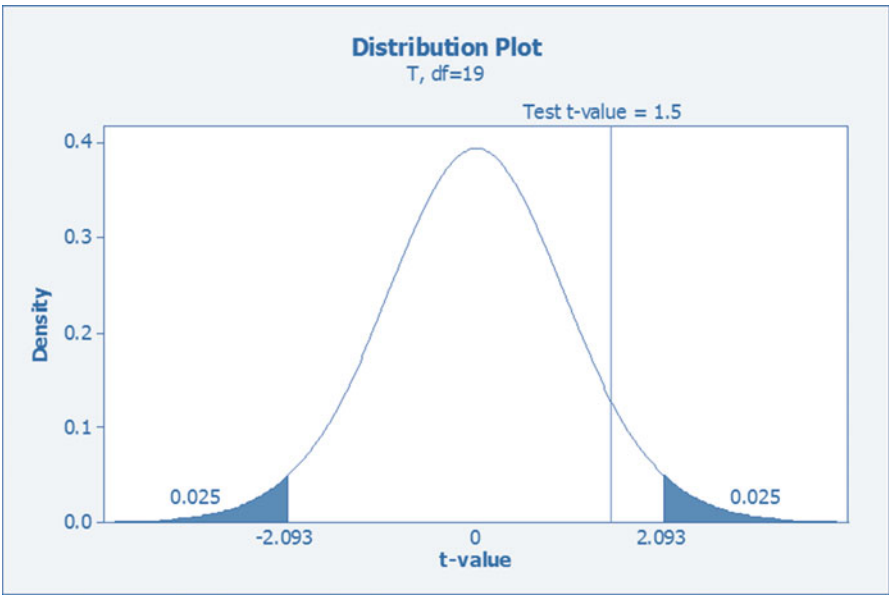


Fig. 6.7 A two-tailed α of 0.05 is shown (*shaded area*) on a t-distribution curve with 19 degrees of freedom. Since the calculated t-value is not within the *shaded area*, then we have failed to reject the null hypothesis. This means that the sample mean sodium level is like our hypothesized mean of 140 mEq/L

Table 6.6 Cutoff values of two-tailed t-distribution for different degrees of freedom at 0.05 and 0.01 significance levels

df	$\alpha = 0.05$	$\alpha = 0.01$
2	±4.303	±9.925
3	±3.182	±5.841
4	±2.776	±4.604
5	±2.571	±4.032
8	±2.306	±3.355
10	±2.228	±3.169
20	±2.086	±2.845
50	±2.009	±2.678
100	±1.984	±2.626

The next step is to calculate the sample mean (μ):

$$\mu = \frac{x_1 + x_2 + \ldots + x_n}{n} \tag{6.12}$$

where n is the sample size. Followed by calculation of the variance (σ^2):

$$\sigma^2 = \frac{\sum_1^n (x_n - \mu)^2}{n - 1} \tag{6.13}$$

The standard deviation (σ) is the square root of the variance:

$$\sigma = \sqrt{\frac{\sum_1^n (x_n - \mu)^2}{n-1}} \tag{6.14}$$

The t-statistic is then given by

$$t = \frac{\mu - m_0}{\sigma/\sqrt{n}} \tag{6.15}$$

$\sigma/\sqrt{n}$ is also known as the standard error of the mean (SE) and is also known as the noise (variation). The numerator of this formula is known as the signal. T-statistics, then, is really a signal-to-noise ratio; the bigger the ratio, the more certain we are that the difference or signal is the cause rather than noise (or random variability).

The t-value can then be looked up in a table of t-values for corresponding degrees of freedom ($n-1$) and significance level.

Example 6.4

Q: Table 6.7 lists the serum sodium levels of 11 patients. Determine if the mean sodium level of these patients is equal to 140 with a significance level of 0.05.

A: The mean of the sample is

$$\mu = \frac{x_1 + x_2 + \ldots + x_{11}}{11} = \frac{1565}{11} = 142.27 \tag{6.16}$$

The standard deviation of the sample is

$$\sigma = \sqrt{\frac{\sum_1^{11} (x_n - \mu)^2}{10}} = 4.78 \tag{6.17}$$

Table 6.7 Sodium levels for Example 6.4

Sample number	Sodium (mEq/L)
1	137
2	150
3	138
4	138
5	144
6	142
7	145
8	147
9	148
10	137
11	139

Now we can calculate the test statistic:

$$t = \frac{142.27 - 140}{4.78/\sqrt{11}} = 1.58 \tag{6.18}$$

Looking at Table 6.6, we can see that, for 10 degrees of freedom and a significance level, α, of 0.05, the cutoff value is 2.228; since the t-value of 1.58 is less than this value, we can conclude that the mean of the sample is equal to the hypothesized mean.

Independent Sample t-Test

In cases where our objective is to compare the mean value of a variable between two independent groups, we can use the two-sample independent t-test. The null/alternative hypotheses pair can be stated as:

- H_0: the mean of the two sets are equal ($\mu_1 = \mu_2$).
- H_1: the mean of the two sets are different ($\mu_1 \neq \mu_2$).

Alternatively, it can be stated that the t-test shows us whether the difference we observe between the two groups is a random occurrence or is there a true difference between the two sets.

It is imperative that, before using the t-test to test the hypothesis, we must make sure that our data fits the assumptions for parametric tests, namely, that it is a continuous variable that either follows a normal distribution or the sample size is sufficiently large.

After formulating the hypotheses, the next step is to calculate the mean and variance of the two groups.

The test statistic is given by

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{6.19}$$

To determine the p-value for the corresponding t-value, the degree of freedom of the t-test should be determined:

$$df = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{\frac{\sigma_1^2}{n_1}}{n_1 - 1} + \frac{\frac{\sigma_2^2}{n_2}}{n_2 - 1}} \tag{6.20}$$

The calculated degree of freedom should be rounded down to the nearest integer.

Example 6.5

Q: Compare the mean sodium level in Table 6.7 to the mean sodium level in Table 6.8, and determine if they are significantly different at α of 0.05.

Table 6.8 Sodium levels for Example 6.5

Sample number	Sodium (mEq/L)
1	140
2	142
3	138
4	148
5	144
6	132
7	155
8	137
9	149
10	157
11	141

A:

The mean and standard deviation of sodium level for the first table are 142.27 and 4.78, respectively. The mean and standard deviation of sodium level for the second table are 143.91 and 7.67, respectively.

The t-statistic is given by

$$t = \frac{142.47 - 143.91}{\sqrt{\frac{4.78^2}{11} + \frac{7.67^2}{11}}} = -0.60 \tag{6.21}$$

The *df* is given by

$$df = \frac{\left(\frac{4.78^2}{11} + \frac{7.67^2}{11}\right)^2}{\frac{\frac{4.78^2}{11}}{11-1} + \frac{\frac{7.67^2}{11}}{11-1}} \cong 16 \tag{6.22}$$

For 16 degrees of freedom and a significance level of 0.05, the cutoff values for t-statistic will be ± 2.120. Thus, because -0.60 does not reach the cutoff point, then we can say that the means of the two groups are equal. Figure 6.8 shows the boxplots for this example.

Paired t-Test

There are circumstances where we want to compare the means between two paired groups. For example, we want to measure the serum sodium levels in patients before and after giving them a diuretic. In paired samples, each patient is its own control. We can formulate the paired t-test hypotheses pair as:

- H_0: there is no difference before and after the treatment.
- H_1: there is a difference in sodium levels before and after the treatment.

Paired tests have more statistical power since the only source of variability is intra-patient variability (e.g., the difference in sodium levels of the patient before

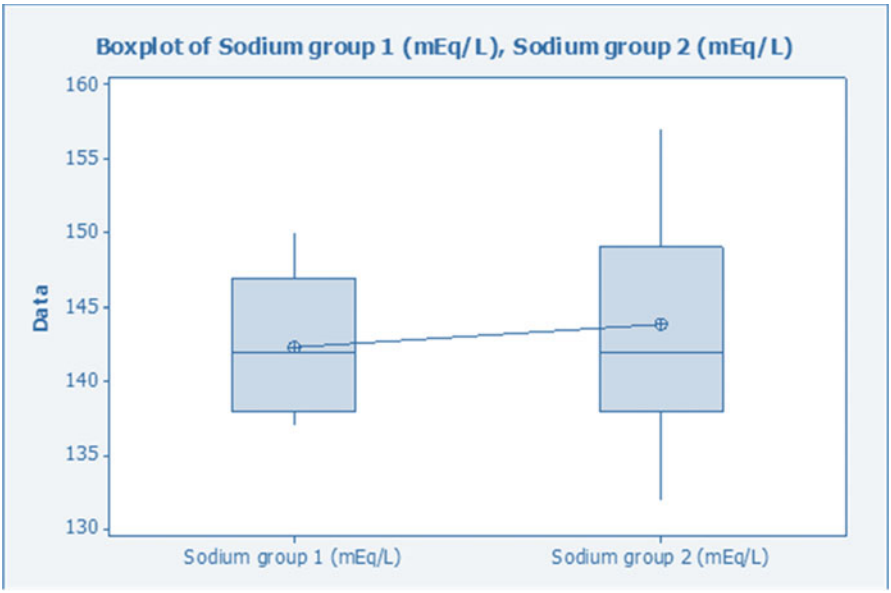


Fig. 6.8 Boxplot diagram for Example 6.5

and after the treatment) rather than inter-patient variability. The intra-patient variability is more likely to be due to true effect rather than random variation, and consequently the paired tests have higher statistical power.

The paired test can be thought of as a one-sample t-test; the difference of values for each pair of observations is determined, and the mean difference (μ_D) is calculated. For the null hypothesis to be true, we expect that the mean difference equals to 0. (In special circumstances we can set the hypothesized mean difference at a non-zero value, e.g., when we expect a natural difference in the before and after calculations.) Thus, now we can rewrite the null/alternative hypotheses pair as:

- H_0: the mean difference between two groups is zero.
- H_1: the mean difference between two groups is non-zero.

The degree of freedom for a paired t-test is like a one-sample t-test and equals $n-1$ with n being the number of paired observations.

Now we can calculate the t-statistic as

$$t = \frac{\mu_D}{\sigma_D/\sqrt{n}} \tag{6.23}$$

Comparing the t-statistic with critical t-values at corresponding degrees of freedom and significance levels can give us the p-value [13, 14].

Table 6.9 Serum sodium levels before and after administration of a diuretic. The difference between the values is written in the fourth column

Sample number	Sodium (mEq/L) before diuretic	Sodium (mEq/L) after diuretic	Difference
1	140	130	−10
2	142	137	−5
3	138	131	−7
4	148	142	−6
5	144	143	−1
6	132	135	3
7	155	142	−13
8	137	128	−9
9	149	142	−7
10	157	148	−9
11	141	140	−1

Example 6.6
Q: Table 6.9 shows the serum sodium levels of patients before and after a diuretic is given. Determine if the diuretic influences sodium levels at significance level of 0.05.

A: The mean difference of the sodium level before and after diuretic administration is 5.91, and the standard deviation is 4.66. Now we can calculate the t-statistic as

$$t = \frac{\mu_D}{\sigma_D/\sqrt{n}} = \frac{-5.91}{4.66/\sqrt{11}} \cong -4.66 \tag{6.24}$$

The degree of freedom of the test is 10 ($n-1$), and the cutoff values for t-statistic at significance level of 0.05 for this degree of freedom are ± 2.228. Since the calculated t-value is beyond this cutoff, then we can reject the null hypothesis and state that the sodium level after diuretic administration is different from sodium level before the diuretic.

One-Way ANOVA

With the t-test, we can only determine if the means of two groups are different. However, what if we have three groups or more and we want to see if the means of these groups are different? To answer this, we can use the "analysis of variance" test, also known as "ANOVA." Here we will discuss the concept of one-way ANOVA. In one-way ANOVA, there is only one grouping variable which can have two or more number of groups (e.g., the country of origin of patients). In two-way ANOVA, there are two grouping variables (e.g., the country of origin of patients and their gender).

The null/alternative hypotheses pair can be stated as:

- H_0: there is no difference in the mean of groups.
- H_1: there is a difference in the mean of groups.

ANOVA uses "F-statistics" which is the ratio of "mean squares" (MS). Alternatively, F-statistic is the ratio of explained variance to unexplained variance or in other words the ratio of between-group variability to within-group variability:

$$F = \frac{MS_{\text{effect}}}{MS_{\text{error}}} = \frac{\text{Explained Variance}}{\text{Unexplained Variance}} = \frac{\text{Between} - \text{group Variance}}{\text{Within} - \text{group variance}} \tag{6.25}$$

The explained variance ($SS_b^2/K-1$) is given by

$$\text{Explained variance} = \sum_{i=1}^{K} \frac{n_i(\mu_i - \mu)^2}{k-1} \tag{6.26}$$

where μ is the overall mean of data, K is the number of groups, μ_i is the mean of the ith group, and n_i is the size of the ith group.

The unexplained variance ($SS_w^2/[N-K]$) is given by

$$\text{Unexplained variance} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \frac{\left(x_{ij} - \mu_i\right)^2}{N-K} \tag{6.27}$$

where x_{ij} is the jth observation of the ith group and N is the overall sample size.

Thus, the F-statistic can be rewritten as

$$F = \frac{\sum_{i=1}^{K} \frac{n_i(\mu_i-\mu)^2}{k-1}}{\sum_{i=1}^{K} \sum_{j=1}^{n_i} \frac{\left(x_{ij}-\mu_i\right)^2}{N-K}} \tag{6.28}$$

The F-statistic follows the F-distribution with $K-1$ and $N-K$ degrees of freedom. The calculated F-value should be compared with the cutoff values for F-distribution with corresponding degrees of freedom and significance level. The critical values for F-distribution for different significance levels and degrees of freedom are provided in Appendix D.

The reasoning behind F-statistics is that if the means of the groups are equal (or almost equal), they will be clustered around the overall mean of the data (i.e., their variance is small); however, if one or two group means are different from the other means and the overall mean, then the variance will be larger. Thus, if the variability (variance) between the groups decreases, we can say that the means are more likely to be equal. Here, the within-group variance is the noise (which we need to cancel, hence being the denominator).

Table 6.10 Summary of F-statistics

	Sum of squares	Degree of freedom	Mean square	F	p-value
Between group	SS_b^2	$K-1$ (df1)	$\frac{SS_b^2}{k-1}$	$\frac{MS_{\text{effect}}}{MS_{\text{error}}}$	Calculated p-value
Within group	SS_w^2	$N-K$ (df2)	$\frac{SS_w^2}{N-K}$		
Total	$SS_b^2 + SS_w^2$	$N-1$			

Table 6.11 Serum levels of drug A in three groups

Sample number	Drug concentration in group 1 (mg/L)	Drug concentration in group 2 (mg/L)	Drug concentration in group 3 (mg/L)
1	30	33	35
2	28	34	37
3	32	29	37
4	34	27	34
5	24	25	30
6	27	31	33
7	30	31	35
8	31	34	33
Mean	29.5	30.5	34.25
Standard deviation	3.11	3.29	2.31
Overall mean:	31.41		

Software that runs ANOVA usually provides a table that summarizes the F-statistic (Table 6.10) [15].

Example 6.7

Q: Table 6.11 lists the serum concentration of drug A in three groups of individuals. At a significance level of 0.05, determine if the serum concentration of drug A in the three groups is equal or different.

A: The explained variance can be calculated as

$$\begin{aligned}\text{Explained variance} &= \sum_{i=1}^{3} \frac{n_i(\mu_i - \mu)^2}{2} = \frac{8(29.5 - 31.41)^2}{2} \\ &+ \frac{8(30.5 - 31.41)^2}{2} + \frac{8(34.25 - 31.41)^2}{2} = 50.167\end{aligned} \tag{6.29}$$

The unexplained variance can be calculated as

$$\text{Unexplained variance} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \frac{\left(x_{ij} - \mu_i\right)^2}{N - K} = \frac{68 + 76 + 37.5}{21} = 8.643 \tag{6.30}$$

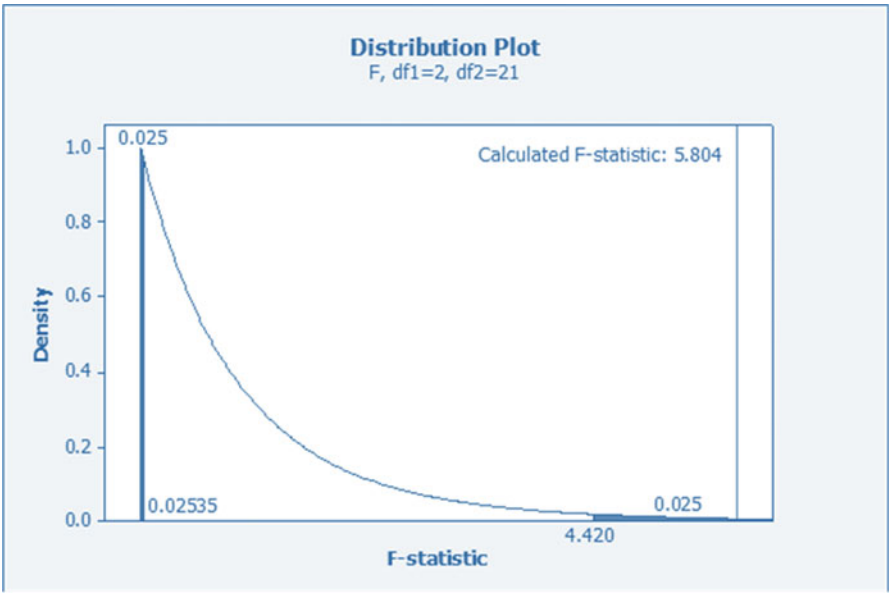


Fig. 6.9 F-distribution plot for Example 6.7

Thus, the F-statistic can be calculated as

$$F = \frac{\text{Explained Variance}}{\text{Unexplained Variance}} = \frac{50.167}{8.643} = 5.804 \tag{6.31}$$

The critical value for F-distribution with degrees of freedom of 2 and 21 and significance level of 0.05 is 3.4668. Thus, since the calculated F-statistic is greater than the critical value, we can reject the null hypothesis and state that at least one group mean is different from others (Fig. 6.9).

Note that ANOVA only shows if there is a difference between group means but it will not show us which groups are different from each other. In Example 6.7, a boxplot diagram can show us that the mean drug level in groups 1 and 3 is different (Fig. 6.10). But sometimes the identification of the source of difference is not so straightforward, and, to find this out, we need to conduct "post hoc tests"; these tests provide one-by-one comparisons between the groups to identify the source of difference. We will briefly introduce some of these tests, but the detailed explanation of the tests is beyond the scope of this book.

One of the tests that can be used is called the "Fisher's least significant difference" (LSD) test; this test performs one-by-one comparisons and finds the source of difference if the null hypothesis is rejected by ANOVA. The shortcoming of the LSD test is that it does not correct for multiple comparisons.

A general rule that you need to remember for running these post hoc tests is that if you are doing multiple one-by-one comparisons, the probability of falsely

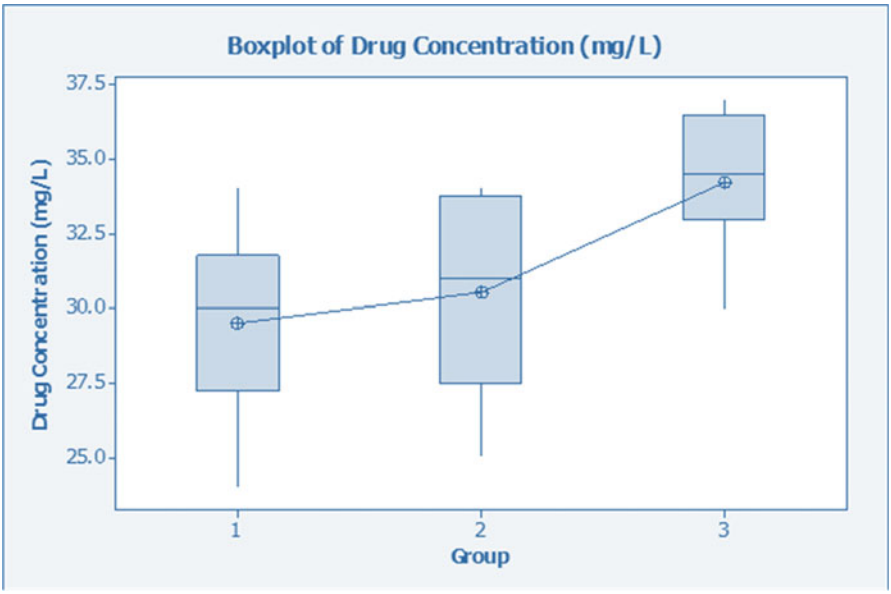


Fig. 6.10 Boxplot diagram for Example 6.7

rejecting the null hypothesis and finding a difference goes up, i.e., the possibility of type I error increases. In Chap. 5, we introduced the concept of Bonferroni correction for dealing with the issue of increased type I error in multiple comparisons, and in fact, one of the post hoc tests that you can use is to do multiple two-sample t-tests and correct the alpha level using a Bonferroni correction.

There are other post hoc tests and corrections. Most statistical software have these post hoc tests as an option when running ANOVA; the choice of which test to use is user and data dependent, but, in general, we recommend "Scheffé's method" which can be used in many different scenarios. Another option is the "Tukey's test" which compares all possible pairs of means and is a powerful post hoc test.

Non-parametric Tests

As we mentioned earlier, the parametric tests make assumptions about the data including its distribution. Non-parametric tests are useful when those assumptions are violated or when you feel that median might be a better summary statistic for your data than mean (e.g., there are many outliers in the data). However, the generalizability of non-parametric tests comes at the expense of their statistical power. Here we will explain two of the more common non-parametric tests: the "Mann-Whitney U test" and the "Kruskal-Wallis test." [16]

Mann-Whitney U Test

"Mann-Whitney U test" is a non-parametric test that can compare a continuous (or ordinal) variable between two groups, and in a sense, it is the non-parametric equivalent of the t-test. This test is one of the more powerful non-parametric tests and in fact can be used even for variables with normal distribution as it has a statistical power comparable to the t-test.

The problem that the Mann-Whitney test addresses is: given two sets of data, are the two sets the same or are they different? Mann-Whitney compares the *median* values of the two sets of data and tests the difference to determine if it is significant. Note the use of the median rather than the mean; means can be strongly determined by extreme values, while medians are much less influenced by extreme values. As part of the test, as we describe below, the values for each dataset are arranged in either increasing or decreasing orders or ranks and are then compared. Quantitation of the difference of the so-called sum of ranks enables us to determine whether the datasets are statistically similar or different by computing the so-called U-statistic.

Simply stated, the Mann-Whitney U test null hypothesis is that the two groups have a similar distribution of data and a similar median. The alternative hypothesis states that the median and distribution of the two data are different.

If the null/alternative hypothesis pair sounds familiar, it is because we introduced a similar concept when talking about the two-sample Kolmogorov-Smirnov test early in the chapter. In fact, the two-sample KS test is also a non-parametric test that you can use. The difference between the KS test and the Mann-Whitney test is that the former is more sensitive to any difference between the two distributions, while the Mann-Whitney test is more sensitive to differences in the median value.

One of the important assumptions of the Mann-Whitney U test is that the variable being compared has a scale, i.e., in comparing two values from the variable lists, we can clearly determine which is greater and which is smaller. This assumption means that continuous and ordinal variables are acceptable. Another assumption is that the observations are independent.

Mann-Whitney U test provides us with a statistic called the U-value. The U-value follows a distribution known, unsurprisingly, as the U-distribution under the null hypothesis. As we demonstrated with other tests above, we need to look up the U-value obtained from the test in the U-distribution and determine if the value is smaller than the cutoff set by our significance level. (For the U test to be significant, the U-value should be smaller than the cutoff value.) The cutoff values are based on the size of each group and the significance level selected. For sample sizes larger than 20, the distribution of U approximates a normal distribution.

The Mann-Whitney test is actually very simple, and we will demonstrate the test using an example [17, 18].

Table 6.12 Values of AST in the chronic hepatitis and healthy group for Example 6.8

AST (units/L)	Group
30	Healthy
35	Healthy
36	Chronic hepatitis
39	Chronic hepatitis
39	Healthy
40	Healthy
42	Healthy
44	Chronic hepatitis
44	Healthy
47	Healthy
50	Chronic hepatitis
52	Chronic hepatitis
53	Healthy
53	Healthy
55	Chronic hepatitis
56	Healthy
60	Chronic hepatitis
100	Chronic hepatitis
160	Chronic hepatitis
200	Chronic hepatitis

Table 6.13 Calculation of sum of ranks for the healthy group for Example 6.8

Healthy group AST value	30	35	39	40	42	44	47	53	53	56	Total
Number of AST values in chronic hepatitis group that are less than this value	0	0	1.5	2	2	2.5	3	5	5	6	27

Example 6.8

Q: Table 6.12 shows the ranked results of serum concentration of AST in a total of 20 patients, ten of whom have chronic hepatitis and ten of whom are ostensibly normal (labeled "healthy" in the table). Using Mann-Whitney test with a significance level of 0.05 (corresponding to a two-tailed U-statistic cutoff of 23), determine if the two have different distributions.

A: Here is how we calculate the U-statistic:

First, for each AST value in the healthy group, count how many AST values in the chronic hepatitis group is less than that value (i.e., rank the value). Count all the ties as 0.5. Sum all the ranks for the healthy group (R_1) (Tables 6.13 and 6.14).

Repeat this for the chronic hepatitis group to obtain the sum of ranks for that group (R_2).

Now the U-statistic for the test is the smaller of the two sum of ranks, which means that U equals 27. As we show below, there is a U-distribution, which is

Table 6.14 Calculation of sum of ranks for the chronic hepatitis group for Example 6.8

Chronic hepatitis group AST value	36	39	44	50	52	55	60	100	160	200	Total
Number of AST values in healthy group that are less than this value	2	2.5	5.5	7	7	9	10	10	10	10	73

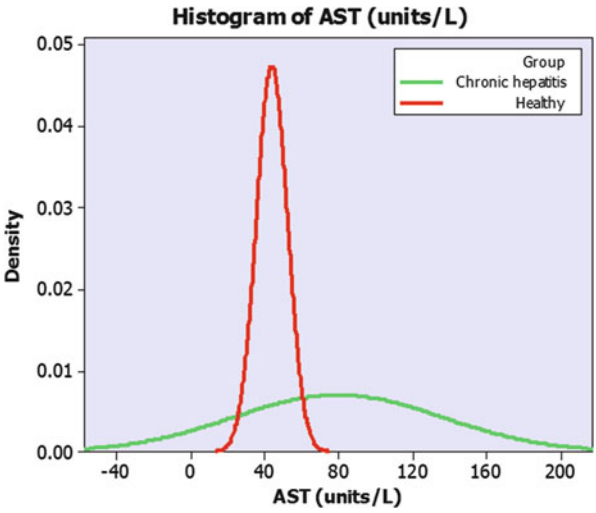


Fig. 6.11 Distribution of values and histogram for AST values in the healthy and chronic hepatitis groups. You can see that the fitted curves for the two distributions overlap

dependent on the number of degrees of freedom. As it happens, at an alpha value of 0.05, the critical value for the U-distribution is 23. Since 27 is greater than the cutoff value of 23, then we can say that we have failed to reject the null hypothesis, i.e., the two datasets are statistically the same. The underlying reason for this conclusion lies in the fact that the median values for the distribution of the values for each dataset are close to one another as we show in Fig. 6.11. We also show that the distributions of the values overlap strongly. We have shown the two distributions using two histograms in Fig. 6.11.

Alternatively, we can use the following formula for calculation of Mann-Whitney U-statistic (either of these two formulas can be used).

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} \text{ and } U_2 = R_2 - \frac{n_2(n_2+1)}{2} \tag{6.32}$$

$$U = \min(U_1, U_2) \tag{6.33}$$

where n_1 is the sample size of group 1 and n_2 is the sample size of group 2. To calculate the sum of ranks, order all the values (from both groups), and assign overall ranks to each value (breaking the ties by giving 0.5). Now, R_1 is the sum of ranks for values from group 1; R_2 is the sum of ranks for values from group 2.

Kruskal-Wallis Test

"Kruskal-Wallis test" is the equivalent of the ANOVA test for comparisons of a continuous variable between multiple groups when the basic assumptions of ANOVA (e.g., normal distribution) are violated. Just like ANOVA, in the Kruskal-Wallis test, the objective is to determine whether a difference between the values of the groups exists. Thus, the null/alternative hypotheses pair is like ANOVA with the exception that in Kruskal-Wallis instead of mean we use the term "stochastic dominance."

Stochastic dominance is a partial ordering. In simple terms, it means that values from one group are more likely to be greater (or lesser) than the values from another group.

The Kruskal-Wallis test uses H-statistics. H-statistics follows a chi-squared distribution with degrees of freedom corresponding to the number of groups minus 1.

The first step in the K-W test is just as Mann-Whitney test and involves ordering all the values from all groups and assigning ranks to them. The average overall rank ($\bar{r}$) and average rank for each group ($\bar{r}_i$) are then calculated.

Assuming there are no ties (or few ties) in the data, then the Kruskal-Wallis H-statistic is given by

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(\bar{r_{ij}} - \bar{r}\right)^2} \tag{6.34}$$

where g is the number of groups, N is the total number of values, n_i is the sample size for group i, and r_{ij} is the rank of the observation j of group i.

If there are many ties in the data, the H-value needs to be corrected:

$$\text{Corrected } H = \frac{H}{1 - \frac{\sum_{i=1}^{G}\left(t_i^3 - t_i\right)}{N^3 - N}} \tag{6.35}$$

where G is the number of groups that have tied ranks and t_i is the number of tie values in the ith group.

If $K-W$ test rejects the null hypothesis (p value $<$ alpha), and you want to determine which group dominates which group (s), then you can use the Dunn's post hoc test.

Calculations for $K-W$ and Dunn's post hoc tests are usually cumbersome, and we recommend that statistical software be employed for these calculations.

Effect Size

Unfortunately, in reporting of scientific studies, usually the *p*-value is the sole metric that is reported with a test. *P*-value on its own has little merit in showing the degree of association or difference, and it simply implies rejection of the null hypothesis. As the *p*-value gets smaller, it means that the certainty with which we can reject the null hypothesis increases but this does not necessarily mean that, for example, the two means in a t-test are getting further apart.

In fact, the pathologists should always be aware of the difference between statistical significance and clinical significance. In pathology and laboratory medicine, a successful diagnostic or prognostic test needs to accurately classify patients and individuals and be a guide in clinical decision making.

For example, we may conduct a study with very high statistical power that shows a very small yet statistically significant difference in mean values of a metabolite between two groups. This difference may be undetectable using less accurate analyzers or be of no importance to the clinicians treating the patients.

For these reasons, we highly recommend reporting of "effect size" in conjunction with reporting of statistical significance. Effect size is the size of the difference between groups. This can give useful insights into clinical applicability of the study results.

In Chap. 5, we introduced the association measures such as Cramer's V; these are essentially effect size measures for nominal variables. For continuous variables, there are multiple effect size measures available. Here we will introduce "Cohen's *d*" and "Cohen's *f*."

Cohen's *d*

We can use "Cohen's *d*" for describing the size of the effect in two-sample t-tests. Cohen's *d* is a concept similar to coefficient of variation: the calculation of Cohen's *d* consists of the ratio of the mean difference to the pooled standard deviation:

$$d = \frac{\mu_1 - \mu_2}{\sigma_{\text{pooled}}} \tag{6.36}$$

The pooled standard deviation can be given by

$$\sigma_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}} \tag{6.37}$$

where n_1 and n_2 are the size of the samples and σ_1^2 and σ_2^2 are their respective variances. Table 6.15 shows the attributed effect size based on d value:

Table 6.15 Effect size based on Cohen's d value

d value	Effect size
0.01	Negligible
0.20	Small
0.50	Medium
0.80	Large
1.20	Very large
2.0	Huge

Cohen's *f*

"Cohen's f" also known as "Cohen's f^2" is used to show the size of the effect in either ANOVA tests or multiple regression. Cohen's f is the ratio of the standard deviation of the means to the pooled standard deviation.

$$f = \frac{\sigma_{\text{means}}}{\sigma_{\text{pooled}}} \tag{6.38}$$

The standard deviation of the means is given by

$$\sigma_{\text{means}} = \sqrt{\frac{\sum_{i=1}^{k} (\mu_i - \mu)^2}{k}} \tag{6.39}$$

where k is the number of the groups, μ_i is the mean of each group, and μ is the overall mean. The pooled standard deviation is the square root of the mean squared error (see above) [19, 20].

Ordinal Variables

There are instances where observations of a variable are ordered or ranked. Ranked variables are in between categorical and continuous variables: they are assigned discrete numbers, but these numbers are part of a scale. The scale, however, may not show exactly how much one rank is different from another, just that it is larger (or smaller) than the other rank.

A common example of ordinal variables in pathology is cancer staging, where cancer patients are staged into pathologic stages I, II, III, and IV and clinical stages I, II, III, and IV. As the stage rank increases, the prognosis worsens, but the degree of worsening of the prognosis is not uniform across stages or different cancers.

Statistical evaluation of ordinal variables requires special non-parametric tests. Here, we will introduce two of such tests: "Kendall's Tau test" and "Spearman's rho test." For smaller sample sizes, use Kendall's Tau and for larger sample sizes use Spearman's rho test.

Kendall's Tau Test

"Kendall's Tau test" is used to identify associations between two ordinal variables. The null/alternative hypotheses pair in Kendall's Tau is stated as:

- H_0: the ranks are discordant, i.e., if the rank for observation i of one variable is larger (or smaller) than the rank for variable *j*th of that variable, then rank for observation *i*th of the other variable should also be smaller (or larger) than the rank for the observation *j*th of that variable (i.e., if $x_i > x_j$ then $y_i < y_j$ and $x_i < x_j$ then $y_i > y_j$). In other words, the variables are independent.
- H_1: the ranks are concordant, i.e., if the rank for observation *i*th of one variable is larger (or smaller) than the rank for variable *j*th of that variable, then rank for observation *i*th of the other variable should also be larger (or smaller) than the rank for the observation *j*th of that variable (i.e., if $x_i > x_j$ then $y_i > y_j$ and $x_i < x_j$ then $y_i < y_j$). In other words, the variables are dependent.

There are three Kendall's Tau tests: Tau-a, Tau-b, and Tau-c. The latter is more commonly used as it adjusts for ties. Here, we will explain Tau-a. Calculation of Tau-b is a bit more complicated with adjustments made for the ties. Tau-c is preferable in situations where the number of ranks in the two variables is unequal. Overall, manual calculation of Tau tests is a time-consuming effort (especially with large sample sizes), and we recommend using computers for this task as most statistical software can calculate all three Tau tests.

Tau-a statistic is given by

$$\tau_a = \frac{\text{Number of concordant pairs} - \text{number of discordant pairs}}{\text{Number of concordant pairs} + \text{number of discordant pairs}} \tag{6.40}$$

In Tau-a calculation, tied pairs are ignored as they are neither concordant nor discordant.

Tau, like Pearson's correlation, has values between -1 and 1 with 0 showing no association, -1 showing perfect negative correlation, and 1 showing perfect positive correlation. In order to test for significance, for smaller sample sizes, the Tau critical value tables can be consulted; if the calculated Tau is greater than the corresponding value for significance level and sample size (number of pairs), then reject the null hypothesis.

Example 6.9

Q: Table 6.16 shows the pathologic stage and nuclear grade of six breast cancer cases. Is pathologic stage dependent on nuclear grade? Critical value for Tau for this example is 0.733.

A: Table 6.17 shows the number of concordant and discordant pairs as we move down the ranks. Remember for each nuclear grade rank we should consider all ranks larger than that rank and check if the corresponding pathologic stage ranks are concordant or discordant (ties are counted only once).

Table 6.16 Table of nuclear grade of breast cancer versus pathologic stage

Sample number	Nuclear grade	Pathologic stage
1	I	I
2	II	I
3	II	II
4	III	III
5	III	IV
6	III	IV

Table 6.17 Counting of concordant and discordant pairs for Example 6.9

Sample number	Nuclear grade	Pathologic stage	Concordant pairs	Discordant pairs
1	I	I	2 [(II,II), (III, III)]	1 [(II,I)]
2	II	I	2 [(II,II), (III,III)]	0
3	II	II	Ignored tie	
4	III	III	1 [(III,III)]	0
5	III	IV	Ignored tie	
6	III	IV	Ignored tie	
Total			5	1

Now, we can calculate the Tau-a:

$$\tau_a = \frac{\text{Number of concordant pairs} - \text{number of discordant pairs}}{\text{Number of concordant pairs} + \text{number of discordant pairs}} = \frac{5-1}{5+1} = \frac{4}{6} = 0.667 \tag{6.41}$$

Since 0.667 is less than the cutoff (0.733), then we cannot reject the null hypothesis.

Spearman's Rho Test

Spearman's rho test is equivalent of Pearson's correlation coefficient for ordinal variables. In fact, Spearman's rho is also the non-parametric alternative of Pearson's correlation, which can be used in situations where the distribution of the two continuous variables is unknown or non-normal.

The Spearman's rho test like the Kendall's Tau test tests for dependence of two ordinal variables. However, where the Tau test counted the number of the concordant and discordant pairs, the Spearman's rho test is a measure of differences between assigned ranks. In this test, the first variable is ordered from smallest rank to largest rank (x_i), and then the difference between the corresponding rank of the second variable with the rank of the first variable is calculated ($d_i = x_i - y_i$). The next step is to calculate the sum of the squared differences ($\sum d_i^2$).

Now, the Spearman's rho is given by

$$\rho_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{6.42}$$

In order to test for significance, a t-value can be calculated from the rho statistics; this t-value follows a student's t-distribution with $n-2$ degrees of freedom under the null hypothesis.

$$t = \rho_s \sqrt{\frac{n-2}{1-{\rho_s}^2}} \tag{6.43}$$

Note that in the above formula for Spearman's rho, ties are not tolerated and the ranks should all be distinct integers [21, 22].

Example 6.10

Q: Two pathologists have ranked their preference for using immunohistochemistry or special stains in workup of renal cancers. Their rankings are given in Table 6.18. Are the two rankings dependent? The significance level for the corresponding t-value is 1.860.

A: The differences and squared differences of the ranks are shown in Table 6.19.

The rho value can be calculated as

$$\rho_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 28}{10(100-1)} \cong 1 - 0.17 = 0.83 \tag{6.44}$$

The corresponding t-value for this rho is

$$t = \rho_s \sqrt{\frac{n-2}{1-{\rho_s}^2}} = 0.83\sqrt{\frac{8}{0.3111}} \cong 4.21 \tag{6.45}$$

Table 6.18 Rankings of two pathologists for IHC markers for renal cancers

Stain	Pathologist 1 preference for IHC	Pathologist 2 preference for IHC
CA-IX	1	1
CK7	2	3
CD117	3	6
AMACR	4	4
CD10	5	5
PAX2	6	2
Colloidal iron	7	7
PAX8	8	8
RCC	9	10
TFE3	10	9

Table 6.19 Corresponding differencesand squared differences in ranking for Example 6.10

Stain	Pathologist 1 preference for IHC	Pathologist 2 preference for IHC	d_i	d_i^2
CA-IX	1	1	0	0
CK7	2	3	-1	1
CD117	3	6	-3	9
AMACR	4	4	0	0
CD10	5	5	0	0
PAX2	6	2	4	16
Colloidal iron	7	7	0	0
PAX8	8	8	0	0
RCC	9	10	-1	1
TFE3	10	9	1	1
Total				28

Table 6.20 Summary of statistical tests in this chapter

Parametric	Non-parametric	Examples
One-sample t-test	One-sample Wilcoxon	Is the mean of our sample equal to a hypothetical mean?
Two-sample t-test	Mann-Whitney U test	Are the liver function test results for patients with chronic viral hepatitis different from patients with autoimmune hepatitis?
One-way ANOVA	Kruskal-Wallis test	Are the liver function test results different between patients with acute viral hepatitis, chronic viral hepatitis, and autoimmune hepatitis?

Since 4.21 is larger than 1.860, then the null hypothesis is rejected, and we can say that the rankings of the two pathologists are similar (dependent). In fact, the calculated p-value is 0.0029.

Summary

In this chapter, we provided an overview of the commonly used statistical tests for continuous and ordinal variables. The choice of these tests depends on the null/alternative hypotheses pair as well as the nature of the data being tested. Remember that parametric tests while more powerful make assumptions about the data; if these assumptions are not met, then the results of a parametric test may be misleading. Non-parametric tests, on the other hand, make much less assumptions about the data but suffer from lower statistical power. Table 6.20 provides a summary of tests introduced in this chapter.

References

1. Strike PW. Statistical methods in laboratory medicine. New York: Butterworth-Heinemann; 2014.
2. Ore O. Pascal and the invention of probability theory. Am Math Mon. 1960;67(5):409–19.
3. Wilcox R. Kolmogorov–smirnov test. In: Encyclopedia of biostatistics. New York: John Wiley & Sons; 2005.
4. Massey FJ Jr. The Kolmogorov-Smirnov test for goodness of fit. J Am Stat Assoc. 1951;46 (253):68–78.
5. Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. BMC Med Res Methodol. 2005;5(1):13.
6. DeCarlo LT. On the meaning and use of kurtosis. Psychol Methods. 1997;2(3):292.
7. Borenstein M, Cooper H, Hedges L, Valentine J. Effect sizes for continuous data. Handbook Res Synth Meta Analy. 2009;2:221–35.
8. Norušis MJ. SPSS 14.0 guide to data analysis. Upper Saddle River, NJ: Prentice Hall; 2006.
9. Sheskin DJ. Handbook of parametric and nonparametric statistical procedures. Boca Raton: CRC Press; 2003.
10. Zimmerman DW. A note on the influence of outliers on parametric and nonparametric tests. J Gen Psychol. 1994;121(4):391–401.
11. Zimmerman DW, Zumbo BD: The effect of outliers on the relative power of parametric and nonparametric statistical tests. Perceptual and Motor Skills. 1990;71:339–349.
12. Sheskin DJ. Parametric versus nonparametric tests. In: International encyclopedia of statistical science. Berlin Heidelberg: Springer; 2011. p. 1051–2.
13. Portney LG, Watkins MP. Foundations of clinical research: applications to practice. Upper Saddle River, NJ: Prentice Hall; 2000.
14. Bailey NT. Statistical methods in biology. Cambridge: Cambridge university press; 1995.
15. St L, Wold S. Analysis of variance (ANOVA). Chemom Intell Lab Syst. 1989;6(4):259–72.
16. Gibbons JD, Chakraborti S. Nonparametric statistical inference. Berlin Heidelberg: Springer; 2011.
17. McKnight PE, Najab J. Mann-Whitney U Test. In: Corsini encyclopedia of psychology. Hoboken: John Wiley; 2010.
18. McKight PE, Najab J. Kruskal-Wallis Test. In: Corsini encyclopedia of psychology. Hoboken: John Wiley; 2010.
19. Cohen J. A power primer. Psychol Bull. 1992;112(1):155.
20. Rosenthal, R. Parametric measures of effect size. In: H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 231–244). New York: Russell Sage Foundation; 1994.
21. Fredricks GA, Nelsen RB. On the relationship between Spearman's rho and Kendall's tau for pairs of continuous random variables. J Statist Plann Inferen. 2007;137(7):2143–50.
22. Stuart A, Ord JK, Arnold S. Kendall's advanced theory of statistics. Vol. 2A. In: Classical inference and the linear model. 6th ed. London: Hodder Arnold; 1999.

Multivariate Analysis

7

Introduction

Thus far, we have been dealing with statistical tests that can handle two variables. This includes regression of two continuous variables or test of independence for two categorical variables. There are many situations, however, where we deal with multiple variables, and we want to understand their contribution to a desired outcome [1]. In pathology and laboratory medicine, this is commonly encountered in the form of multiple risks/hazards/exposures and an outcome like disease status. In these situations, one-by-one comparison (univariate analysis) between input variables and the response variables is problematic as we will explain below. In such situations, we can use another class of statistics called multivariate statistics.

Another goal in pathology and laboratory medicine is to create diagnostic/prognostic models. As our field has expanded, we now have a battery of tests at our disposal that can help in diagnosing patients. However, interpreting these tests at the same time and coming to a single conclusion about the patient can be difficult. In these situations, decision-making tools and criteria are helpful. One of the ways for creating such criteria is to apply multivariate statistics.

Thus, there are two advantages in running multivariate statistics. One is to create predictive models that can summarize multiple independent variables and allow the summary metric to be used for diagnostic/prognostic purposes. For example, the relation of mutational status which is composed of multiple genes with prognosis can be summarized in a single summary metric. Another advantage is to account for confounding factors.

When running statistical tests, there are variables or factors that correlate with the dependent variable and the independent variable, thus making the correlation and relation between the independent and dependent confounded. Ideally, in design of the studies, confounding factors should be identified and controlled using random sampling and stratification. Unfortunately, it is difficult to account for all confounding factors in trial design; hence multivariate statistical tests allow us to explain the remaining confounding factors and understand the true effect between

A. Momeni et al., *Introduction to Statistical Methods in Pathology*,
DOI 10.1007/978-3-319-60543-2_7

an independent variable and a dependent variable. For example, in calculation of glomerular filtration rate (GFR), we know that there is a correlation between serum creatinine and GFR; however, this correlation is affected by age, gender, and race. Thus, to calculate GFR using serum creatinine, adjustments are required. These adjustments can be made with the help of multivariate statistics.

As we pointed out, an alternative to multivariate statistics is to stratify the data and run the univariate analysis in each stratum. For example, we can compare the data in women and men separately. However, as we mentioned in the previous chapters, running multiple comparisons increases the chance of type I error, furthermore by dividing the sample into strata each stratum will have smaller sample size. Thus too many strata will lead to diminished statistical power. For all these reasons, multivariate analysis is often a more viable option.

Here we will talk about "generalized linear models." Unlike previous chapters, we will not provide in depth explanation of the solutions used for calculating these statistical tests as they require advanced understanding of mathematical notation, and, in reality, hardly anyone attempts to solve these equations without the aid of statistical software. Instead, in this chapter, we focus on providing an understanding of the concepts for each test, provide guidance on the appropriate context for each test, and helping you to interpret the results of these statistical tests.

Generalized Linear Model

The most common way to fit one (or more) variables to another variable and determine association or create predictive models is to create a linear regression model. The assumption of the linear regression model is that the response or dependent variable should be a continuous variable with a normal distribution. The linear regression creates a linear predictor model where the value of the response variable (Y) is predicted by the input variables (X_k). Each input variable in the model has a corresponding regression coefficient (β), which is the amount of change in the response variable if the corresponding input variable changes one unit and all other input variables are constant.

Generalized linear models expand the concept of linear regression to continuous and non-continuous response variables like nominal or ordinal variables. There are different categories of generalized linear functions including linear regression, logistic regression, and Poisson regression. The models with non-continuous response variables use "link functions" as proxies for the response variable. These link functions are commonly the logarithm of the odds of the response variable (called logit functions). The advantage of the logit functions is that they are continuous variables with near normal distribution and thus can be used for fitting models just as in linear regression [2].

We introduced the concept of linear regression for two variables in Chap. 4. Here we will explain multiple regression analysis (i.e., regression between multiple input variables and a single response variable). We will also talk about logistic regression and introduce different types of logistic regression.

Multiple Regression Analysis

We learned in Chap. 4 that we can form a regression line between an independent and dependent continuous variable pair. This regression line can also tell us about the correlation of the two variables. This regression and correlation are fundamental principle of many analyzers used in pathology where an indirect or proxy measure with high correlation with the true analyte level is used for measuring that analyte.

There are situations where the dependent variable is affected by multiple independent input variables, i.e., we want to model the relationship between a dependent variable (Y) with multiple predictors (X_k). The input variables can be continuous or categorical. For categorical variables, they must be recoded into a "dummy variable" or "indicator variable," i.e., for each category, a number should be assigned (for binary variables, a value of 0 and 1 should be assigned).

In biology and medicine, the relation between the predictors and outcome is hardly deterministic, i.e., the predictors can give an approximation of the dependent variable with an associated degree of error. The "multiple regression analysis" (also known as multiple linear regression) allows us to devise a general additive linear model that correlates a dependent variable to multiple input variables. This linear predictor function can be stated as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon, \tag{7.1}$$

where β_0 (alternatively called α) is the intercept of the model. β_k is the regression coefficient for each input variable and ε is the error. The assumption in these models is that the error is normally distributed with a mean value of 0.

The standard deviation of error (σ) will also be the standard deviation of the dependent variable for fixed values of input variables (as summarized in Eq. 7.18 in Chap. 4). If the standard deviation of error is large, then the confidence interval for Y will also be large. As the standard deviation of error decreases, the confidence interval for Y also narrows. The regression function essentially provides the mean predicted value of Y for any set of values for input variables (in fact, comparison of the predicted values with the observed values of Y allows us to check for goodness of fit of the model).

The coefficient of regressions can be shown in plots as well. For input variable x_k if all the other variables are considered as constants, then the regression coefficient is the slope of the plot of x_k versus Y (Fig. 7.1). The intercept is the value of Y if all the predictors are set to 0.

Model fitting in multiple regression analysis is like simple linear regression. The simplest method involves using ordinary least squares (OLS) principle to estimate the regression coefficients. This principle is based on minimizing the sum of squared deviation of the model from the observed values. For each observation, we have an observed dependent variable value; the model also predicts a value for the dependent variable by estimating the value from the input variables. The difference between the two values is the residual of the model at each point. An ideal model will have the smallest sum of residuals. This can be stated as

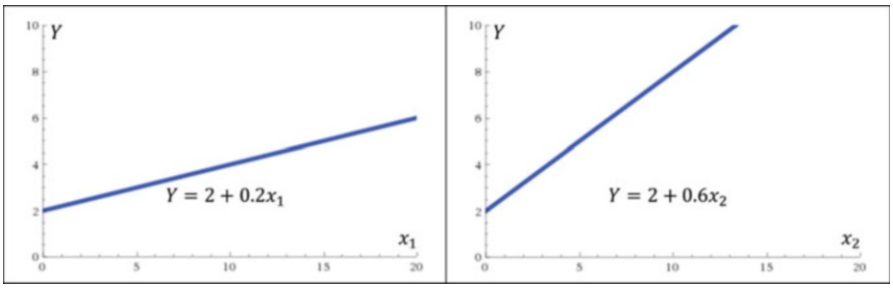


Fig. 7.1 The plots for regression coefficients for the predictor function: $Y = 2 + 0.2x_1 + 0.6x_2$. The *left* panel shows the plot for the first input variable and the *right panel* shows the plot for the second input variable versus the response variable while the other variable is constant (and set to 0)

Table 7.1 Table of parameter estimates for multiple regression analysis

Input variable	B	SE_B	T	$p-$value
Intercept	B_0	SE_{B0}	T_0	$p-\text{value}_0$
x_1	B_1	SE_{B1}	T_1	$p-\text{value}_1$
x_2	B_2	SE_{B2}	T_2	$p-\text{value}_2$
...	...	...	...	...
x_k	B_k	SE_{Bk}	T_k	$p-\text{value}_k$

$$Q = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}))^2, \tag{7.2}$$

The regression coefficients for the model are ones that allow for the above sum to be minimized. This essentially means solving $k + 1$ equations with $k + 1$ unknowns (in this case the intercept and the regression coefficients). Statistical software follow solutions similar to what was discussed in Chap. 4 for solving this problem. It mainly involves deriving the normal equations for the predictor function and inverting the matrix of the resultant normal equations and using this matrix for solving the equations.

After running a multiple regression analysis in a statistical software, a table like Table 7.1 will be provided which summarizes the parameter estimates.

In this table, the first row (sometimes last row) will be the intercept of the model. Subsequent rows will be the predictors of the model. Commonly, the first column will have the regression coefficients for each variable and the second column will contain the corresponding standard error of the regression coefficient. This essentially is an estimate of the variability of the regression coefficient (as the sample size increases, the standard error of the coefficients will be smaller).

The third column in this table usually is a test statistic that allows us to determine if the input variable is a statistically significant predictor of the dependent variable. A commonly used statistic is the t-value. This value is ratio of the regression coefficient and its standard error:

$$t = \frac{B_k}{SE_{Bk}}, \tag{7.3}$$

t-value follows the t-distribution; simple interpretation of t-value will be that if t-value is larger than the corresponding cutoff for a significance level of 0.05 or alternatively if the 95% confidence interval of the regression coefficient excludes 0, then we can say that input variable is a predictor of the dependent variable (i.e., the null hypothesis that the regression coefficient for that input variable is 0 is rejected). The corresponding *p*-value for the t-test is shown in the next column [3, 4].

Assessing Utility of the Fitted Model

The next question to ask is how good the predictor function fits the observed data. Assessing the goodness of fit is done using the R^2 statistics. The "R-squared" is also known as the "coefficient of determination." This statistic determines the proportion of the variability in the dependent variable that is explained from the independent variable(s). For simple linear regressions with intercept, the r-squared statistic is the square of the correlation coefficient (r^2). For multiple correlations, the R^2 is the sum of all correlation coefficients adjusted for correlations between the input variables. R^2 statistic can have values between 0 and 1. As the statistic nears one, the prediction power of the model increases, with 1 being the perfect score (meaning that all the variations in the response variable are explained by the input variable).

The calculation of R-squared is done through calculation of the "residual sum of squares" ($SS_{residual}$) and "total sum of squares" (SS_{total}) as discussed in Chap. 4, Eq. 4.19:

$$SS_{total} = \sum_{i=1}^{n} (y_i - \bar{y})^2, \tag{7.4}$$

$$SS_{residual} = \sum_{i=1}^{n} \left(y_i - y_{i\ predicted}\right)^2, \tag{7.5}$$

where $\bar{y}$ is the mean of the observed values of the dependent variable. This equation has $(n - (k + 1))$ degrees of freedom, where n is the sample size and k is the number of input variables.

Now a measure called "R-squared (R^2)" can be calculated that will have a perfect score of 1 (showing a perfect model) and minimum score of 0:

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}, \tag{7.6}$$

This test essentially means if the total deviance of the predicted model is small compared to the variation of the observed values of Y, then the model predictions are good.

The problem with R-squared is that, if you increase the number of predictors in the model, R-squared invariably increases as the model has more ways to fit the data. Yet when the model is tested on another sample, the model's prediction will not be accurate. Because of this, it is advisable to use "adjusted R-squared" statistics to assess the model's fit.

$$\text{Adjusted } R^2 = 1 - \left(\left(\frac{n-1}{n-(k+1)} \right) \times \frac{SS_{\text{residual}}}{SS_{\text{total}}} \right), \tag{7.7}$$

F-Test

Multiple regression creates a model with the estimated linear function that predicts the dependent variable. The assumption is that all the dependent variables are at least predicted by one of the independent input variables, i.e., at least one of the regression coefficients should be non-zero. This can be stated as a null/alternative hypotheses pair and then tested:

- H_0: All the regression coefficients are zero, i.e., there is no relationship between the dependent variable and the input variables ($B_1 = B_2 = \ldots = B_i = 0$).
- H_1: At least one of the regression coefficients is non-zero. (i.e., model's prediction with inclusion of at least one predictor is superior to prediction using only the intercept).

This hypothesis can be tested using the "F-test". This test is similar to ANOVA and compares the explained and unexplained variance in the data. The test statistic can be derived from the R-squared measure:

$$F = \frac{R^2 / k}{(1-R^2) / (n-(k+1))}, \tag{7.8}$$

The test statistics follows an upper-tailed F-distribution (i.e., significance means that the model's prediction with predictors is better than the intercept only model) with df1 of k degrees and df2 of $(n-(k+1))$ degrees. If the test is significant (i.e., p-value is less than alpha), then we can reject the null hypothesis and state that a model can be fitted to data and that there is at least one predictor in the model.

Example 7.1

A study has been done to understand the predictors of glomerular filtration rate (GFR) or its equivalent creatinine clearance. In the study, creatinine clearance (response variable), age, weight, gender, and serum creatinine levels (input variables) are measured in 20 individuals. The results of the study have been

Table 7.2 The results for the study in Example 7.1

Sample number	Serum creatinine (mg/dL)	Weight (kg)	Age (years)	Gender indicator	GFR (mL/minute)
1	1.0	60	25	0	94
2	1.0	62	40	0	88
3	1.3	70	56	0	62
4	2.0	80	60	0	44
5	2.0	73	72	0	33
6	1.6	90	53	0	68
7	1.2	58	40	0	66
8	3.0	65	76	0	16
9	2.5	120	30	0	73
10	2.5	74	60	0	32
11	2.0	80	60	1	38
12	2.0	66	31	1	43
13	1.0	66	70	1	55
14	3.5	110	41	1	37
15	1.9	58	34	1	38
16	1.7	66	21	1	56
17	1.2	55	51	1	48
18	2.0	43	70	1	20
19	1.5	45	30	1	39
20	2.5	61	60	1	24

Table 7.3 The parameter estimates for multiple regression analysis of Table 7.2

Parameter	*B*	Std. error	*t*	Sig.
Intercept	95.172	1.076	88.468	.000
Serum creatinine	−26.317	.461	−57.095	.000
Weight	.526	.016	33.875	.000
Age	−.562	.013	−42.909	.000
Gender = female	−13.356	.480	−27.838	.000

summarized in Table 7.2. Note that the gender (male/female) has been altered into an indicator function (male, 0; female, 1).

We have run a multiple regression analysis to determine whether age, weight, gender, and serum creatinine are predictors of creatinine clearance or not. The parameter estimates are summarized in Table 7.3.

The results show that all the calculated regression coefficients for the input variables are significant (i.e., they do not include 0 in their confidence interval). The linear function for the prediction model can be written as

$$\begin{aligned}\text{GFR} = {} & 95.17 - 26.17(\text{serum creatinine}) + 0.526(\text{weight}) - 0.526(\text{Age}) \\ & - 13.356(\text{if female}),\end{aligned} \tag{7.9}$$

This predictor function is similar to the Cockcroft-Gault Formula used in calculating GFR in patients. In fact, that formula was devised using a similar statistical approach.

The R-squared measure for the model is 0.912 with adjusted R-squared of 0.911. This shows that the model is a very good fit of the creatinine clearance with small residuals. The F-test for the model returns a value of 38.85 with df1 of 4 and df2 of 15; this translates to a *p*-value of ~0.000, i.e., the model can be fitted into the data.

Residual Plots

Some of the measures of regression analysis can be plotted in what are known as the "residual plots." Among them "residual versus fit" plots are more important.

Residuals versus fit graph plots the residuals (Y-axis) for each fitted (predicted) value (X-axis). Preferably, the points should fall randomly on both sides of 0. The points should not have any recognizable patterns or fall on one side of the 0 line. Outlier points are immediately recognizable in the plot and show that some of the observations vary considerably from the predicted model. Also, patterns like fanning of the points show that the variance is not constant throughout the model. If any patterns are identified, then they should be investigated and the model adjusted accordingly (e.g., outliers may indicate measurement or sampling error).

Figure 7.2 shows the residual versus fit plot for Example 7.1.

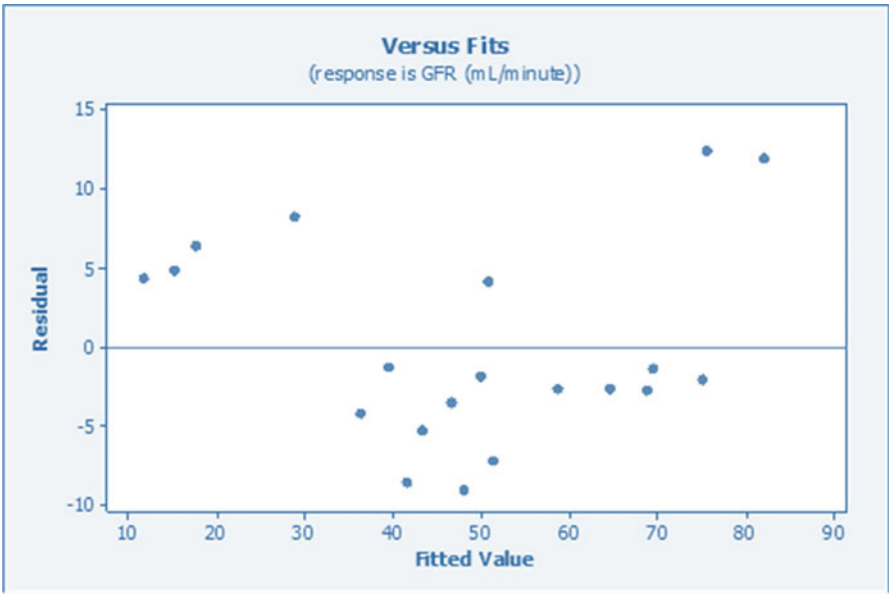


Fig. 7.2 The residual versus fits plot for Example 7.1. The points show no recognizable pattern and fall on both sides of the 0

Interaction and Collinearity

The assumption of linear regression is that the input variables are independent, i.e., they do not correlate with each other or the correlation of one variable with the dependent variable is not affected by another of the input variables. The first argument of independence can be restated as lack of collinearity and the second argument can be stated as lack of interaction.

"Interaction" occurs if the changes in the dependent variable are affected not only by an additive linear function (e.g., $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$) but also by multiplicative interaction between the input variables (e.g., $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$). Thus, if any interaction exists, multiple linear regression model will not explain all the variability in data and lead to underestimation of the effect of the predictors on the dependent variables.

If an interaction is known to exist between the input variables, then alternative model fitting processes like higher-order models should be used (e.g., a full quadratic model can be used for a model with two variables assuming full interaction between the predictors (i.e., each variable with itself as well as the two variables with each other)). Alternatively, all possible interactions can be formed and entered into a linear regression model (this will only work for small number of predictors) to check if any of the interactions have a significant regression coefficient.

Generally, it is advisable that, if an interaction is suspected, the data is tested for interaction. Most statistical software have options for checking for interaction. These tests are usually stated as either "R-squared change" or "F-test change." Simply stated, the software iteratively adds possible interaction terms and at each step checks for changes in either R-squared measure or the F-value. If the changes in these statistics are statistically significant from one iteration to the other, it shows that the interaction term should be incorporated into the prediction model.

Collinearity occurs when one input variable has significant correlation with another variable. For example, weight and body mass index are significantly correlated. In general, collinear variables should not be included in the model simultaneously and inclusion of one should exclude the other. While, the overall prediction model may remain valid, the individual weight (coefficient) given to predictors may be erroneous and lead to misinterpretation of the relation of the predictor with the outcome. Despite the tolerance of overall model to collinearity, multicollinearity can lead to decrease in model fit and should be avoided in regression analysis. Statistical software can calculate a measure known as the "variance inflation factor" (VIF) which is the inverse of the percent of variance in the predictor that cannot be accounted for by the other predictors; hence large values indicate that a predictor is redundant and can be excluded from the model (VIF values of more than 10 should prompt you to investigate that predictor and perhaps leave it out of the model).

Logistic Regression

"Logistic regression" is a predictive model that predicts the relation of the independent variables to a dependent variable: this dependent variable is either a nominal or ordinal variable with distinct categories. If the dependent variable has only two categories, then the model is called a "binary logistic regression." For categorical dependent variables with more than two responses, a "multinomial regression" is used, and finally for ordinal variables, "ordinal regression" is used. Logistic regression in fact is a generalized linear model with binomial response [5].

Binary Logistic Regression

There are situations when we want to determine a binary outcome based on multiple inputs. A common example in pathology is when multiple tests, inputs, or criteria are used to predict whether the patient has a disease or not. For example, combinations of size, nuclear grade, mitotic activity, invasion, and architectural patterns are used for histopathologic diagnosis of many cancers. The design of such criteria is a crucial aspect of anatomic pathology and allows for reproducibility of diagnosis between different pathologists. In these situations, "binary logistic regression" can be used for modeling the data.

The input variables in binary logistic regression are also known as explanatory variables. These can be used to estimate probability of one of the two events of the binary response variable, e.g., combining multiple variables in order to determine if cancer is present. Binary logistic regression is also useful in finding the relation of a continuous independent variable and a binary categorical response variable. For example, we can use binary logistic regression to determine the correlation of level of serum procalcitonin with presence of sepsis.

For a single explanatory variable (x) and a binary response variable (y), the logistic regression can be expressed as

$$y = \begin{cases} 1 & \beta_0 + \beta_1 x + \varepsilon > 0 \\ 0 & else \end{cases}, \tag{7.10}$$

where β_0 (sometimes written as a) is a constant and is the intercept of the model. β_1 is the "regression coefficient" and ε is the error. The purpose for running logistic regression is to determine the constant and the coefficient.

Regression analysis using multiple explanatory variables is an extension of the above expression:

$$y = \begin{cases} 1 & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \varepsilon > 0 \\ 0 & \text{else} \end{cases}, \tag{7.11}$$

Logistic regression is based on the "standard logistic function," which is an S-shaped probability distribution. In standard logistic function, the input (t) is a real

number that can take any value from -∞ to ∞, yet the output ($\sigma(t)$) will always assume a value between 0 and 1. The formula for the standard logistic function is

$$\sigma(t) = \frac{1}{1 + e^{-t}}, \tag{7.12}$$

The input for the logistic function can be a function as well. For logistic regression, we can use $\beta_0 + \beta_1 x$ as the input and we can rewrite the formula as

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}, \tag{7.13}$$

As the $F(x)$ can only assume values between 0 and 1, we can treat this value as a probability; in fact, this is the probability of the response event occurring given the input value (e.g., the probability of a patient having sepsis given the value of the serum concentration of procalcitonin).

The "logit function" (g) of the above expression is equal to the linear regression expression:

$$g(F(x)) = \ln \frac{F(x)}{1 - F(x)} = \beta_0 + \beta_1 x, \tag{7.14}$$

The reasons for this transformation is that the logit function, unlike the probability expression, is not bound by the limits 0 and 1 and can assume any value between -∞ and ∞. Furthermore, logit function is a linear expression that is easier to discover and interpret. Finally, the logit function can be exponentiated to give the odds; the odds of the response event occurring given the input variable will be:

$$\text{Odds} = \text{Exp}\big(g(F(x)) = \frac{F(x)}{1 - F(x)} = e^{\beta_0 + \beta_1 x} \tag{7.15}$$

The regression coefficients can be estimated using the "maximum likelihood estimator." In general, there are no perfect solutions to this equation, and this model tests different estimates generated by iterative algorithms (either Newton-Raphson or iteratively reweighted least squares) to find the best estimates of the coefficients. Thankfully, with advents of statistical software, this process is done by computers. The calculation of coefficients is beyond the scope of this book.

Let us explore regression coefficients further. β_0, as we said, is the intercept, and it can be stated as the odds that the response event occurs if all the input variables are zero. β_i is the coefficient for input variable x_i, and the odds of the response variable change by $\beta_i x_i$ for changes in the input variable. If the $\beta_i > 0$, then as x_i increases the odds of response event increase, and conversely if $\beta_i < 0$, then as x_i increases the odds of response event decrease.

Example 7.2
Q: A study was done in order to determine the utility of serum procalcitonin levels (pct) in detection of sepsis (sep). The following logit expression is the result of the study. Interpret the regression coefficients. If a patient has a procalcitonin level of 25 μg/L, calculate the odds and probability of that patient having sepsis.

$$g(\text{sep}) = -2.7 + 0.15(\text{pct}), \tag{7.16}$$

A: The β_0 is -2.7. This means that the odds of a patient having sepsis if the procalcitonin level is zero equals:

$$\text{baseline odds} = e^{-2.7} \cong 0.067, \tag{7.17}$$

This translates to a probability of 0.062, i.e., even if the procalcitonin level is undetectable, there is still a 6% chance that the patient has sepsis.

The β_1 is 0.15. This means that for every unit increase in procalcitonin level, the logit function of sepsis increases by 0.15. We can calculate the change in the odds ratio as well:

$$\text{Odds Ratio} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1(x)}} = e^{\beta_1} = e^{0.15} \cong 1.16, \tag{7.18}$$

The odds of sepsis for the patient with procalcitonin level of 25 μg/L is given by

$$\text{Odds}(\text{sep}) = e^{\beta_0+\beta_1 x} = e^{-2.71+0.15(25)} \cong 2.83, \tag{7.19}$$

The probability of the patient having sepsis is given by

$$\text{Probability of sepsis} = \frac{1}{1+e^{-(-2.71+0.15(25))}} \cong 0.74, \tag{7.20}$$

This means that there is a 74% probability that a patient with procalcitonin level of 25 μg/L has sepsis.

In cases where the input variables are all categorical or ordinal, we can test for association between the input variables and the response variable using a chi-squared test as well. However, the chi-squared test will only tell you if the variables are related and do not provide a predictive result whereby, based on the input variables (risk), the response can be predicted.

Testing for hypothesis in binary logistic regression is to test if each input variable contributes to the model. In other words, the null/alternative hypothesis for each input variable x_i can be expressed as:

- H_0: The variable does not contribute to the model and has no effect on outcome; in other words, the regression coefficient of the variable is zero ($B_i = 0$).
- H_1: The variable contributes to the model and exerts an effect on outcome; in other words, the regression coefficient of the variable is non-zero ($B_i \neq 0$).

The basis for hypothesis testing is that the distribution of estimates for the regression coefficient produced by the maximum likelihood estimator follows a (near) normal distribution. Thus, we can simply say that if the 95% confidence interval of the estimated regression coefficient does not contain 0, then we have rejected the null hypothesis with a significance level of 0.05.

This testing is either done using the "Wald test" or the "likelihood ratio test." Here, we will explain the Wald test as the latter is computationally intensive (the likelihood ratio test, however, is considered more robust).

Wald test is the ratio of the squared regression coefficient to the squared standard error of the coefficient regression:

$$W_i = \frac{B_i^{2\,2}}{SE_{B_i}}, \tag{7.21}$$

The standard error of the coefficient of regression can also be extracted from the maximum likelihood estimator. Wald statistics follows a chi-squared distribution with one degree of freedom. Thus, if the Wald statistics is greater than 3.84 (for significance level of 0.05), then we can reject the null hypothesis.

Binary logistic regression for multiple input variables can be performed using the likelihood ratio test or Wald test. It must be noted that choosing input variables should be based on possible causality as random pairings may sometimes lead to incorrect models. The model can be run in a stepwise hierarchical method or with all input variables entered at the same time. The latter choice is often better since it can provide a better model as well as provide meaningful statistical significance information about each variable. In the stepwise approach, input variables are added (or removed) one by one with model's prediction power estimated at each step, and improvements in the prediction power are then attributed to the input variable changed.

Statistical software usually provides a summary table after running binary logistic regression (Table 7.4). These tables usually include information such as intercept, B coefficient, Wald score (or likelihood ratio score), significance levels, and odds for each input variable (B coefficient exponentiated) [6, 7].

Table 7.4 Parameter estimates table for binary logistic regression

Input variable	B	SE_B	W	$p-\text{value}$	Odds (Exp(B))
x_1	B_1	SE_{B1}	W_1	$p-\text{value}_1$	e^{B_1}
x_2	B_2	SE_{B2}	W_2	$p-\text{value}_2$	e^{B_2}
...	...	...	...	...	...
x_i	B_i	SE_{Bi}	W_i	$p-\text{value}_i$	e^{B_i}
Constant	B_0	SE_{B0}	W_0	$p-\text{value}_0$	e^{B_0}

Example 7.3

Q: In a study the association of mutations in three genes in endometrial tissue with occurrence of endometrial cancer is evaluated. Table 7.5 provides the summarized results of the study.

Our goal is to determine the association of these mutations in these genes with endometrial cancer status. One approach will be to perform three univariate analyses, one for association of each gene with cancer. Table 7.6 shows the results of the univariate analyses.

Univariate analysis shows P53 and PIK3CA mutations to be significantly associated with cancer status. This does not account for the fact that some cases may have both mutations and the *p*-values do not reflect this. One way to correct this is to run stratified chi-squared tests (in this case Fisher's exact test since the sample size for each stratum is very small), but there are too many strata, and sample size in some strata is so small that prevents calculation of test statistic (e.g., P53-positive, PTEN-positive, PIK3CA-negative stratum with only 2 cases).

Our other option is to run a binary logistic regression. The results of this regression are shown in Table 7.7.

Table 7.5 Summary of results for Example 7.3

						Cancer		
						No	Yes	
						Count	Count	Total
P53	No	PTEN	No	PIK3CA	No	7	3	10
					Yes	3	3	6
			Yes	PIK3CA	No	5	2	7
					Yes	1	2	3
	Yes	PTEN	No	PIK3CA	No	4	7	11
					Yes	1	7	8
			Yes	PIK3CA	No	0	2	2
					Yes	0	3	3
Total						21	29	50

Table 7.6 Univariate analysis results for Example 7.3

Gene	Chi-squared value	*p*-value
P53	8.489	0.004
PTEN	0.035	0.851
PIK3CA	3.955	0.047

Table 7.7 Binary logistic regression parameter estimates for example 7.3

Input variable	B	SE_B	W	$p-$value
P53	1.965	0.710	7.658	0.006
PTEN	0.651	0.754	0.746	0.388
PIK3CA	1.267	0.702	3.260	0.071
Constant	−1.198	0.604	3.934	0.047

After running the logistic regression, we can see that only P53 has remained statistically significant and in fact some of the association of PIK3CA has been explained away because of co-occurrence of its mutations with mutations of the other genes.

Goodness of Fit

Binary logistic regression creates a model that predicts the response variable using the input variables. Consequently, one important piece of information that we need from the model is how good is the model in predicting the response variable. In other words, how well do the predictions fit the observed data.

The goodness of fit is usually measured using variations of the R^2 statistic. We previously mentioned how the R-squared statistic is calculated. One major problem with R-squared statistic for multiple correlations is that it is always additive, i.e., as you add input variables, the R-squared will increase. Take the above example; if the model's R-squared is calculated with intercept, P53, PTEN, and PIK3CA as part of the model, the R-squared will be higher than a model with only P53 and intercept. Yet, this increase in R-squared is false since we showed that PTEN and PIK3CA should not actually be part of the predictive model. For this reason, the adjusted R-squared statistic is usually used which adjusts the statistic as more variables are included in the model.

Since binary regression has a binary response, direct calculation of R-squared is not possible. In binary logistic regression, the goodness of fit is measured using the following formula:

$$D_{\text{null}} - D_{\text{fitted}} = -2 \ln \frac{\text{likelihood of the null model}}{\text{likelihood of the fitted model}}, \tag{7.22}$$

In this calculation, the nominator is the likelihood of the null model which means the difference in likelihood between the null model (the response variable is only dependent on the intercept) and a perfect model (saturated model), and the denominator is the difference in likelihood between the fitted model (with all the input variables) and the saturated model. This can be used to calculate a pseudo-R-squared for the model:

$$\text{Likelihood ratio } R^2 = \frac{D_{\text{null}} - D_{\text{fitted}}}{D_{\text{null}}} = 1 - \frac{D_{\text{fitted}}}{D_{\text{null}}}, \tag{7.23}$$

Let us interpret the above equation. If the fitted model is near the perfect model, it means that the difference between the perfect model and the fitted model is small, much smaller than the difference between the null model and the perfect model; thus, the ratio of the differences is close to 0, and consequently, the R-squared will be close to 1.

Other R-squared values that are usually reported include the "Cox and Snell R-squared" and "Nagelkerke R-squared." Interpretation of the Cox and Snell R-squared is difficult since the value for R-squared even for a perfect fit does not

reach 1. Statistical software also report *p*-values with goodness-of-fit statistics. For the model to fit the observations, the *p*-value of the R-squared measures *should not be* statistically significant.

Multinomial Logistic Regression

In previous section, we evaluated logistic regression with a binary nominal response. What if the categorical response has multiple responses? For example, we may wish to categorize a pathologic lesion into three or more diagnostic categories using a series of input variables. In these situations, we can use "multinomial logistic regression." For multinomial regression, logistical and loglinear models can be used. Generally, we recommend using logistical models as they are easier to interpret and formulate.

Multinomial logistic regression (also known as polytomous logistic regression) is an extension of the binary logistic regression. The basis of the binary logistic regression is to calculate the logit function. In binary logistic regression, the logit function is solitary, i.e., there is only one logit function for the model. In multinomial regression with a dependent variable with k number of response categories, there will be *k-1* non-redundant logit functions (out of a total of $k(k-1)/2$ possible logits).

The model predicts that observation i has the outcome k using a linear predictor function ($f(k, i)$) of M input variables. This predictor function can be stated as

$$f(k,i) = B_{0,k} + B_{1,k}x_{1,i} + B_{1,k}x_{1,i} + \ldots + B_{M,k}x_{M,i} + \varepsilon, \tag{7.24}$$

Thus, an independent variable (x_m) will have a separate regression coefficient ($B_{m,k}$) for each response category (Y_k). The regression coefficient can be interpreted as the increase in log odds of falling into category Y_k versus all other categories, resulting from a one-unit increase in the mth covariate (input variable), if other covariates are constant.

In order to estimate the regression coefficient, an extension of the maximum likelihood estimator called the "maximum a posteriori" estimation is used, which provides the best estimates of the coefficient using iterative processes.

There are instances where we can change the multinomial responses into a sequence of binary choices, and, instead of running a multinomial regression, we can run $k-1$ binary logistic regressions. Figure 7.3 shows the multinomial responses for histopathologic diagnosis of breast lesions in the left panel. The right panel shows the multinomial response transformed into a sequence of binary choices. This approach is only useful in instances where the multinomial responses are sequential. For these models, each input variable will have k-1 regression coefficients (one for each binary regression).

The multinomial regression provides two useful sets of information: first, the model will provide the overall association of the independent variables with the response variable. These are usually calculated using a likelihood ratio test and

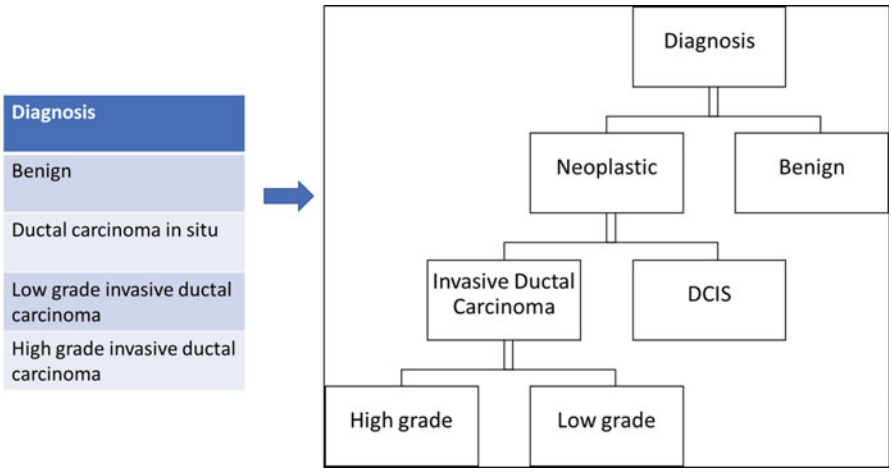


Fig. 7.3 A multinomial variable can be transformed into a sequence of binary variables

provide the test statistic (chi-squared with appropriate degrees of freedom) and the corresponding *p*-value for each independent variable. The second information set consists of regression coefficients for each independent variable as it pertains to a response in the multinomial dependent variable.

The goodness of fit is measured usually using either a Pearson chi-squared test or a deviance test; for both tests, if the resulting chi-square score is less than the significance level for corresponding degrees of freedom, then it can be said that the model fits the observations (i.e., no statistically significant difference between the model and observed data is present). Pseudo-R-squared metrics such as "Cox and Snell R-squared" and "Nagelkerke R-squared" will also be reported [8].

Example 7.4

A study aims to evaluate three immunohistochemical stains (CyclinD1, ER, and CK5/6) in diagnosis of intraductal breast lesions (normal, usual ductal hyperplasia (UDH), atypical ductal hyperplasia (ADH), and ductal carcinoma in situ (DCIS)). The results of the study are summarized in Table 7.8.

The table provides the regression coefficients for each independent variable for each category. For example, the regression coefficients for CK5/6, for ADH, DCIS and UDH, are, respectively, 3.267,3.840 and 0.168. The regression coefficients for CK5/6 are only significant for DCIS and ADH. Since the input variables are all categorical with binary responses, the model assigned regression coefficients only to one of the responses, since the other response is redundant (hence, a regression coefficient of 0).

Note that the normal response category is missing from the table as it has been set as the reference response, i.e., the model is designed based on the ability to distinguish each response category from the normal category. Looking at the UDH

Table 7.8 Summary of results for Example 7.4. Running a multinomial logistic regression will provide us with the following results (Table 7.9)

						Diagnosis				
						ADH	DCIS	Normal	UDH	Total
CK5/6	No	ER	No	CyclinD1	No	1	2	3	1	7
					Yes	3	0	0	0	3
			Yes	CyclinD1	No	1	1	0	2	4
					Yes	4	9	0	0	13
	Yes	ER	No	CyclinD1	No	0	0	3	4	7
					Yes	1	1	1	4	7
			Yes	CyclinD1	No	1	0	3	0	4
					Yes	1	1	1	2	5
Total						12	14	11	13	50

Table 7.9 Parameter estimates from multinomial regression analysis of Example 7.4

Diagnosis		*B*	Std. Error	Wald	Sig.	Exp(*B*)
ADH	Intercept	.589	1.088	.293	.588	
	[CyclinD1 = 0]	−3.663	1.407	6.779	.009	0.026
	[CyclinD1 = 1]	0				
	[ER = 0]	−.331	1.064	0.097	0.756	0.719
	[ER = 1]	0				
	[CK5_6 = 0]	3.267	1.358	5.789	0.016	26.228
	[CK5_6 = 1]	0				
DCIS	Intercept	.566	1.131	0.250	0.617	
	[CyclinD1 = 0]	−3.704	1.436	6.655	0.010	0.025
	[CyclinD1 = 1]	0				
	[ER = 0]	-1.223	1.109	1.216	0.270	0.294
	[ER = 1]	0				
	[CK5/6 = 0]	3.840	1.426	7.252	.007	46.541
	[CK5/6 = 1]	0				
UDH	Intercept	0.889	0.991	0.806	0.369	
	[CyclinD1 = 0]	−1.411	.999	1.996	0.158	0.244
	[CyclinD1 = 1]	0				
	[ER = 0]	0.313	0.903	0.120	0.729	1.367
	[ER = 1]	0				
	[CK5/6 = 0]	0.168	1.000	.028	.866	1.183
	[CK5/6 = 1]	0				

diagnosis, we can see that none of the IHC stains has a significant regression coefficient, and this essentially means that the model will be ineffective in separating UDH from normal breast tissue.

The overall significances assigned to each IHC stain based on the likelihood ratio test are 0.006, 0.499, and 0.001 for CyclinD1, ER, and CK5/6, respectively. This shows that overall CyclinD1 and CK5/6 are associated with diagnosis.

Ordinal Logistic Regression

"Ordinal logistic regression" also known as ordered logistic regression is a variation of the multinomial regression that is used when the response variable is ordered. For example, tumor pathologic stage is an ordered variable, and, if a researcher is interested in designing a model where a set of input variables can predict the cancer stage, then an ordinal logistic regression model can be used. While a multinomial regression can be used is these instances, the ordering information will be lost and usually the ordering information is important. For example, in case of tumor staging, different stages of the tumor carry different weights in clinical decision making, and thus the order of the stages is a very important information.

The ordinal logistic regression makes important assumptions about the data which is that the orders of the response variable follow "proportional odds." The

proportional odds assumption means that the relationship between each pair of the outcome categories is the same, e.g., in context of tumor staging, the relationship of stage I to stages II, III, and IV is the same as the relationship of the stage II to stages III and IV. This assumption is fundamental and allows for one set of coefficients to be calculated for the input variables.

For example, in cancer staging (with four stages and the proportion of cancer patients in each stage are represented by T_1, T_2, T_3, and T_4), the odds of these stages must remain proportional, i.e., the number added to log of odds of each stage compared to higher stages must remain constant:

$$\log\frac{T_1+T_2+T_3}{T_4}=\log\frac{T_1+T_2}{T_3+T_4}+Q=\log\frac{T_1}{T_2+T_3+T_4}+2Q, \tag{7.25}$$

In fact, it is prudent to either test for this assumption before running an ordered regression or most commonly check for the assumption as part of the ordered regression. Many statistical software programs include an option with ordinal regression to check for the proportional odds assumption (also known as parallel lines assumption). When the software checks for this assumption, you need the null hypothesis to be true, i.e., for odds to be proportionally distributed, the p-value for the parallel lines test should be insignificant.

When the ordered logistic regression is run, statistical software will provide a table which contains the regression coefficient for each input variable and its corresponding p-value, as well as threshold levels for the different ranks in the ordered response variable. These thresholds (cut points) show where the latent variable ($\beta_0+\beta_1x_1+\beta_2x_2+\ \ldots\ +\beta_ix_i$) is cut for each rank of the ordered response variable.

The goodness of fit of the model is calculated like other regression models, i.e., using -2 log likelihood ratio as well as the R-squared statistic [9].

Example 7.5
In a study, the aim is to construct a predictive model based on the modified Bloom-Richardson score (composed of nuclear, tubular, and mitotic activity scores) to predict the stage of breast cancer. Each of the input variables can assume values of 1, 2, or 3. The response variable (cancer stage) can assume values of 1, 2, 3, or 4. Table 7.10 summarizes the results of the study.

In order to create a model, we can run an ordered logistic regression, the results of which are provided in Table 7.11.

In Table 7.11 we see the regression coefficients for nuclei," "tubules," and "mitosis," as well as their standard errors, the Wald test, p-values, and the exponents of the coefficients (i.e., odds). The interpretation of these numbers is similar to what we explained in the binary logistic regression section. Both "nuclei" and "mitosis" are statistically significant, while "tubules" is not. For example, for nuclear grade, we can interpret the regression coefficient such that for a one-unit increase in nuclear grade, we expect a 1.525 increase in the ordered log odds of

Table 7.10 Summary of results for Example 7.5

						Stage				
						1.00	2.00	3.00	4.00	Total
						Count	Count	Count	Count	Count
Tubules	1.00	Nuclei	1.00	Mitosis	1.00	3	0	0	0	3
					2.00	2	1	0	0	3
			2.00	Mitosis	1.00	0	3	0	0	3
					2 3.00	0	0	0	1	1
	2.00	Nuclei	1.00	Mitosis	1.00	1	1	0	0	2
					2 3.00	0	1	0	0	1
			2.00	Mitosis	1.00	1	0	1	1	3
					2.00	0	1	2	0	3
					3.00	0	0	1	1	2
			3.00	Mitosis	1.00	0	0	1	0	1
					2.00	1	1	1	1	4
	3.00	Nuclei	1.00	Mitosis	2.00	1	0	1	0	2
			2.00	Mitosis	1.00	1	1	0	0	2
					2.00	0	0	0	1	1
					3.00	0	0	1	1	2
			3.00	Mitosis	1.00	0	0	1	1	2
					2.00	0	1	1	2	4
					3.00	0	0	0	1	1
Total						10	10	10	10	40

Table 7.11 Parameter estimates from ordinal regression analysis for Example 7.5

		Estimate	Std. error	Wald	df	Sig.
Threshold	[Stage = 1.00]	4.321	1.341	10.374	1	0.001
	[Stage = 2.00]	6.078	1.525	15.891	1	0.000
	[Stage = 3.00]	7.737	1.707	20.540	1	0.000
Location	Nuclei	1.525	0.501	9.246	1	0.002
	Tubules	0.434	0.451	.926	1	0.336
	Mitosis	1.165	0.458	6.461	1	0.011

being in a higher cancer stage, if all the other variables in the model remain constant.

The thresholds are shown at the top of the parameter estimates. Threshold for going from a stage I cancer to stage II cancer is 4.321, i.e., if the value of the latent variable reaches 4.321, the stage increases from I to II. Similarly, if the threshold of 6.078 is reached, then the model will increase the stage of the cancer to III.

Internal and External Validity

One of the benefits of regression models is that they can be used to estimate (or predict) the response variable using the input variables. This is very useful in pathology where sets of diagnostic and/or prognostic criteria are needed to diagnose as well as prognosticate the diseases. The use of regression models as diagnostic/prognostic criteria, however, requires validation of the results.

Linear regression models make important assumptions about data, including existence of a linear function that can fit the response variable with addition of predictive variables and the distribution of the results. The first step in validation of the model is to look at the fit of the model to the data that was used to estimate the model. This is called, "apparent validation." Ideally, if a model is to be used as diagnostic criteria, there needs to be small residuals and high R-squared values. If the model fit is subpar, then attempts at nonlinear model fitting or introduction of interactions can be made. However, there might be predictors present that have not been included in the model (or the study) that could have increased the model's efficacy. Thus, a first step would be to identify an acceptable regression model.

In Chap. 4, we introduced the concept of calibration. In regression models, especially regression models with a response variable that is either continuous or ordinal or follows a Poisson distribution, calibration can be employed to evaluate models: the calibration line of the fitted model versus the observed variable should form a straight line. The slope of the fitted line is equal to the R-squared metric of the model. Figure 7.4 shows the correlation plot of the fitted model versus the observations for Example 7.1.

For binary logistic regression, discrimination power of the model is of more concern. The discrimination power can be shown using a receiver operating

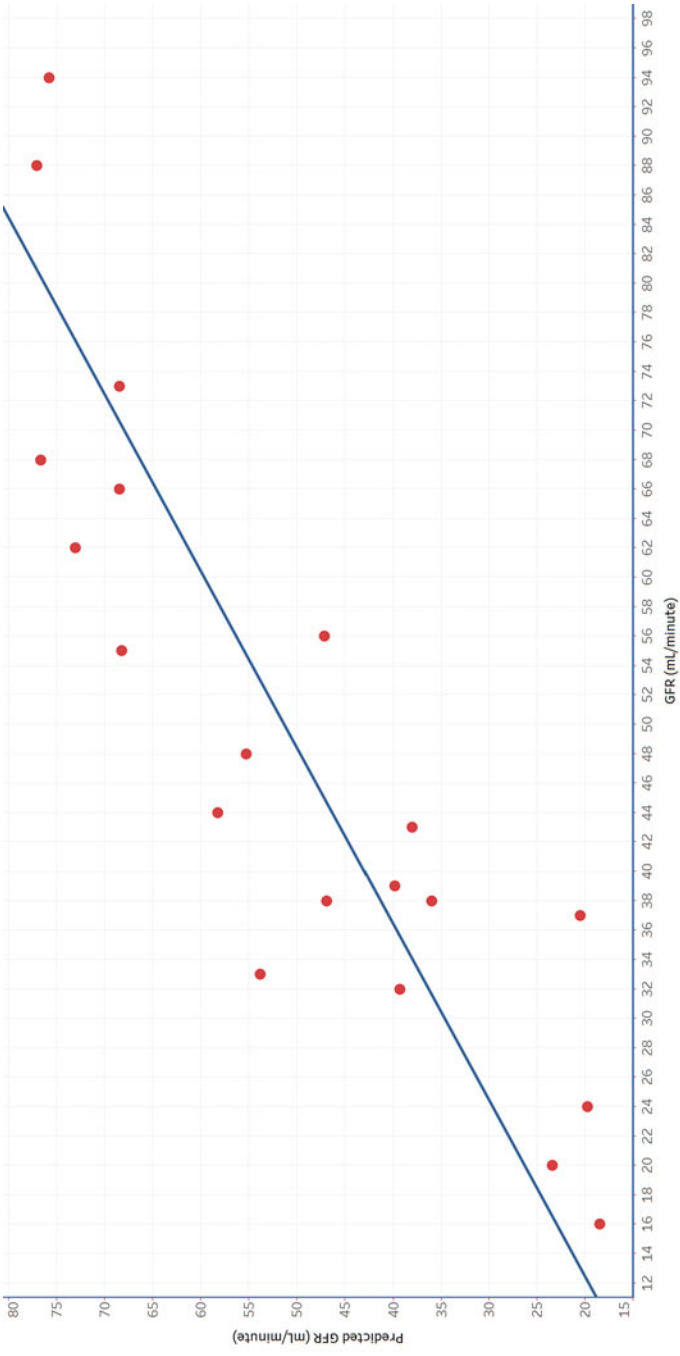


Fig. 7.4 The correlation plot of GFR (mL/min) versus predicted GFR from Example 7.1. The fitted line is straight and has a slope of 0.911 (which equals the R-squared statistics of the model)

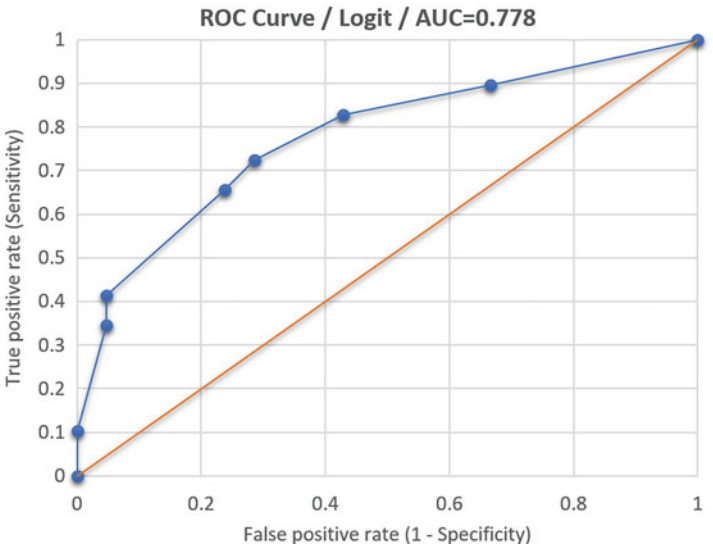


Fig. 7.5 ROC curve for Example 7.3. The area under curve is 0.778

characteristics (ROC) curve. The logit values for each observation can be used to construct a ROC curve. Larger area under curve (AUC) indicates a better fit of the model to the observed values, i.e., the combined sensitivity and specificity of the model will be higher. Figure 7.5 shows the ROC curve for Example 7.3.

In the end, even the best fit models have been designed to fit the currently observed data that was included in the design of the model. Thus, if a model was found which fitted the data, then you would expect the model to pass the apparent validity step. The real test of a model validity is to check for "internal" and "external validity." In fact, many models with apparent validity will later fail the external validity step. Factors such as randomization, large study sample, and controlling for bias can boost the chances of a model passing the validation step.

Internal validity refers to testing the model in new data drawn from the observed population, i.e., the observations are randomly (or sometimes non-randomly) divided into two or more blocks with one block for designing (training) and creating the regression model and the rest of the blocks for testing (validating) the predicted model. Internal validation should show the test data to have a prediction performance as good as (or near) the performance of the model for the design (training) data.

The simplest method for internal validation is a "split-sample method." In split-sample method, the observations are are divided into development and validation datasets either randomly or using a variable not in the model. The model is estimated using the development sample and then it is tested on the validation sample. If the predictions of the model for the validation dataset matches the observations of the validation sample, then we can claim internal validity of the model. This approach, however, is the least robust of the internal validation

methods, especially since only part of the data is used for model development decreasing its overall accuracy (unless the sample is large).

An alternative approach is the "cross-validation method," whereby the training and testing datasets are alternated, i.e., a subset is used for developing the model and then the model is tested on the other subset(s), and then the other subset is used for developing a model and the model is tested on the remaining data. In the end, the estimated model that has the best fit (or an average of calculated regression coefficients is used to calculate the final regression coefficient) is chosen as the internally valid model for the data. Common cross-validation methods include k-fold cross-validation where the sample is broken into k subsets; one subset is used for validation and the remaining subsets are used for developing data, and this process is iterated k times, each time a different subset being chosen as the validation sample. Common number for subsets is 10, but researcher may choose different numbers based on the data they have.

An extreme form of cross-validation is called "leave-one-out-approach." In this method, all the samples minus one are used for estimating the model, and that model is tested on the left-out sample, and this is then repeated for each observation (each observation gets to sit out from the development once).

The best approach to internal validation, however, is "bootstrapping." While in cross-validation and split-sample methods, the sample chosen for development or validation is not replaced and the data is effectively split into two or many parts, in bootstrapping, the sample chosen for development is drawn with replacement. This essentially means that one patient can be drawn for model development sample up to N times (with N being the sample size) because bootstrapping is done with replacement. Each time a sample is drawn and a model is fitted. This process is repeated many times (at least 100 times) until a stable regression model emerges. The final model (which is a combination of all iterations) is then validated using the entirety of the original sample (with each observation only represented once). Bootstrapping is a robust approach to model estimation and internal validation and it is highly recommended in design of diagnostic/prognostic criteria.

True validation of the model can only be achieved through testing the model on new data (i.e., new patients). This is the concept behind external validation: to use a new set of observations, a separate study setup for the validation of the model is needed. A perfect validation would show the model estimations of the external validation data to be as good as the estimations for the internal validation data.

External validation can be in the form of temporal validation, whereby the same researchers use data obtained at a different time from the original observations to validate their model. It can also have a spatial validation form, where the same investigators validate their model in a different setting (perhaps another laboratory or hospital). The best form of external validation, however, is the fully external form, where a different set of investigators in other centers validate the results. It is recommended that the sample size for external validations studies is at least as large as the original study.

The external validation is relatively simple once the new data is obtained. For binary dependent variables, a 2×2 contingency table can be formed, and false

negative, false positive, true negative, and true positive rates can be determined, and the specificity and sensitivity of the model should be recalculated. Measures of agreement such as Kappa coefficient can determine whether the model should be validated based on the new data or not. For continuous dependent variables, correlation tests can be used with the model expected to have high positive correlation with the new observations [10–12].

Summary

As pathologists, we employ many criteria in diagnosing pathologic conditions. Some of these criteria are arbitrarily set by expert panels which may sometimes lack clinical relevance. However, many criteria are designed using statistical methods that combine several observations and form models that can help determine disease states, measure physiologic or functional status of the patient, or predict prognosis. Many of these criteria are designed using generalized linear models (GLMs). As such, the knowledge of GLM not only allows you to better understand how criteria that you use were designed but also guide you in designing models of your own.

Here we introduced you to more commonly used GLMs including binary logistic regression, multinomial regression, ordinal regression, and multiple linear regression. For each we discussed how a model is designed and what are the parameters of a model. In the last section, we briefly explained the concept of validation which is essential if a model is to become clinically applicable.

References

1. Agresti A, Kateri M. Categorical data analysis. Berlin Heidelberg: Springer; 2011.
2. McCullagh P. Generalized linear models. Eur J Oper Res. 1984;16(3):285–92.
3. Stolzenberg RM. Multiple regression analysis. Handbook Data Analy. 2004;25:165–208.
4. Mason CH, Perreault WD Jr. Collinearity, power, and interpretation of multiple regression analysis. J Mar Res. 1991;1:268–80.
5. Rodríguez G. Logit models for binary data. Mimeo: Princeton University; 2002.
6. Tranmer M, Elliot M. Binary logistic regression. Cathie Marsh Census Survey Res Paper. 2008;20:3–42.
7. Maroof DA. Binary logistic regression. Statistical methods in neuropsychology. New York: Springer; 2012. p. 67–75.
8. Böhning D. Multinomial logistic regression algorithm. Ann Inst Stat Math. 1992;44 (1):197–200.
9. Bender R, Grouven U. Ordinal logistic regression in medical research. J R Coll Physicians Lond. 1997;31(5):546–51.
10. Sedgwick PM. Statistical significance and confidence intervals. BMJ: Br Med J (Online). 2009;339:b3401.
11. Proctor RW, Capaldi EJ. Internal and External Validity, in Why Science Matters: Understanding the Methods of Psychological Research, Blackwell Publishing Ltd, Oxford, UK. doi: 10.1002/9780470773994.
12. Slack MK, Draugalis JR. Establishing the internal and external validity of experimental studies. Am J Health Syst Pharmacy. 2001;58(22):2173–84.

Imputation and Missing Data

8

Introduction

Many statistical tests assume that the dataset is complete and the variables don't have missing values. The presence of missing data is a big challenge, especially if the distribution of the missing values is not completely random. For example, if a point-of-care device can measure blood glucose concentration, but cannot return values higher than 400 mg/dL, then any sample with a blood glucose of more than 400 mg/dL will be returned as missing or as an error term; in this case running statistical tests on the results, while ignoring the missing values introduces a considerable bias into the data that can lead to wrong interpretation of the data.

Missing data is a common occurrence in the field of pathology and laboratory medicine as most analyzers have multiple points of failure which can lead to errors or measurement failures. This is usually offset by repeat measurements or use of backup analyzers. However, there are still situations where datasets have missing data, and dealing with missing data is a necessary skill for a pathologist.

There are different solutions for dealing with missing data. They can be as simple as dropping the observation with missing data to more complex solutions such as "imputing" the missing data. In this chapter, we will explain some of these solutions [1].

Missing Data

A general definition for missing data is a variable for an observation that has no value assigned. In other words, the cell for the variable for that observation is either empty or contains terms such as N/A, missing, or so on.

Missing data may occur because of a general unresponsiveness of the observation or subject. For example, a clotted blood sample from a patient can lead to a general unresponsiveness where multiple tests on the clotted sample may fail to return values or when the genomic material extracted from a paraffin-embedded

A. Momeni et al., *Introduction to Statistical Methods in Pathology*,
DOI 10.1007/978-3-319-60543-2_8

material is degraded to an extent that many genomic studies may fail. In laboratory medicine and as part of preanalytical checks, general unresponsiveness is rare as such samples or observations will be filtered out with a repeat sample being tested. In general, the solution for observations with multiple missing values is to drop that observation; even variables for such observations that have values may represent significant error as the sample generally may have failed the required quality metrics.

The missing values can occur vertically as well, i.e., a variable may commonly have missing values. This is a common occurrence with less robust and extremely sensitive instruments. It may also occur because of the nature of the test. Again, systematic failures of a test or instrument are critical, and a decision regarding exclusion of the variable from analysis (or using an alternative metric or repeating measurement using a new instrument) should be made.

More common occurrences are instances where a single or few variables have missing values. These missing values if infrequent and random have a limited effect on the inferences from the dataset, and these are the types of missing data that we will focus on. These missing data points can have three main types: "missing not at random," "missing at random," and "missing completely at random." Missing not at random can itself be broken down to "missingness" that depends on unobserved predictors or "missingness" that depends on the missing value itself.

Types of Missing Data

Understanding the type of missing data we have in the dataset is the first step in dealing with these missing data. Many of the solutions used for missing data are dependent on the type of missing data.

Nonrandom Missingness

When the missing values have a pattern, or occupy a defined portion of a distribution that is not identifiable by the data in the dataset, then we have missing that is not at random. Nonrandom missingness introduces serious bias into the observations and causes inferential problems.

Nonrandom missingness can occur when the missing values depend on unobserved predictors. For example, assume that we are studying the serial brain natriuretic peptide (BNP) measurements in outpatients in regular 1-month intervals. Then patients with worsening heart failure especially if they become decompensated are less likely to be seen in an outpatient setting or to show up for their regular interval check. The missing data on BNP are then dependent on the severity of heart failure, and, if we have failed to measure and account for the severity of heart failure, then our missing data represent a nonrandom missingness depending on an unobserved predictor. If the reason for this study is to assess the effectiveness of a new drug, then we will have a bias introduced in our data in favor of the new drug which can have serious healthcare implications. Or, for example, if in a study of immunohistochemical (IHC) staining pattern of a tumor on biopsy,

when no tumor cells are seen on the IHC slide, then the value is recorded as missing. If this is dependent on the size of the biopsy and this is not recorded, then a nonrandom missingness has occurred.

There are times when the nonrandom missingness is dependent on the missing value itself. This is a serious problem and is often difficult to identify or to solve. For example, if point-of-care analyzers are more likely to show "N/A" or fail to show a result for glucose measurements for patients with high glucose levels, then this type of missingness has occurred. If this missing is absolute, e.g., all patients with glucose levels more than 400 mg/dL will have missing values, then we can say that "censoring" has occurred.

Censoring is common in laboratory medicine since many measurement instruments have an effective range, and values outside the range are not recorded. This concept is also common in survival studies, and, in fact, many survival statistics take censoring into account. We will talk about censoring in survival studies in the next chapter.

Please note that censoring should not be confused with "truncation" which is also a form of bias and is especially problematic in medicine. Truncation occurs when the measurement device returns values that are always within its measurement range. For example, a glucose measurement device with an effective range of 400 mg/dL will return serum glucose concentrations of 400, 500, and 1000 mg/dL as 400 mg/dL. While in censoring, the presence of missing data can alert you to the possibility of values outside of the effective range, in truncation this does not happen which can lead to serious clinical implications.

Random Missingness

Missingness at random implies that the missing value is not random, but the missing value is only dependent on the available, observed, and recorded data. For example, if in a study on glucose measurement, men are 20% less likely to show up for glucose measurement and gender is recorded in the dataset and the only missing values are for men, then a random missing has occurred. Many statistical software programs, when dealing with missing data, assume that missingness is random (either random or completely random).

To identify random missingness, a binary logistic regression is employed. The dependent binary response variable will be whether the data is missing or non-missing. If the other recorded variables can account for the missing values in the dataset, then the model predicted by the binary logistic regression will have a very good fit (i.e., very high R-squared value). Despite the results of the binary regression, we cannot be sure that the missing data is truly random. In fact, it is nearly impossible to ascertain if the missing data is random because the data is missing and we can only make an assumption about the nature of those missing data. In general, it is advisable that the binary logistic regression model be as inclusive as possible with many predictors to have a higher possibility of fitting the missing data [2].

Table 8.1 Summary of results for Example 8.1. The right column indicates whether the BNP variable has missing values

Case number	Gender	NYHA class	BNP	Missing
1	0	1	210	0
2	0	1	130	0
3	0	1	160	0
4	0	1	200	0
5	0	2	180	0
6	0	2	200	0
7	0	2	320	0
8	0	2	300	0
9	0	2	360	0
10	0	3	372	0
11	0	3		1
12	0	3	510	0
13	0	3		1
14	0	4		1
15	0	4	500	0
16	1	1	221	0
17	1	1	154	0
18	1	1	203	0
19	1	2	220	0
20	1	2	210	0
21	1	2	330	0
22	1	2	320	0
23	1	2	300	0
24	1	2	320	0
25	1	3	400	0
26	1	3	430	0
27	1	3	410	0
28	1	4		1
29	1	4		1
30	1	4	600	0

Example 8.1

Q: Table 8.1 lists the results of a study for measurement of brain natriuretic peptide (BNP) in outpatient setting. Some of the values for BNP are missing. Is the missingness at random?

A: We can run a binary logistic regression (see Chap. 7) with the response variable being the last column and the gender and NYHA class being the inputs.

The results show that NYHA class is a significant predictor of missingness (B, 2.39; p-value, 0.023). The model also has a respectable Nagelkerke R-squared of 0.546, yet still the model's classification table makes a classification error in 4 out of 30 cases. This suggests some nonrandom element in missingness.

The researchers then looked up the age of the patient from the charts and added them to the table (Table 8.2).

Table 8.2 Summary of results for Example 8.1 with age included

Case number	Gender	NYHA class	Age	BNP	Missing
1	0	1	60	210	0
2	0	1	54	130	0
3	0	1	58	160	0
4	0	1	70	200	0
5	0	2	82	180	0
6	0	2	70	200	0
7	0	2	62	320	0
8	0	2	48	300	0
9	0	2	54	360	0
10	0	3	61	372	0
11	0	3	88		1
12	0	3	65	510	0
13	0	3	87		1
14	0	4	78		1
15	0	4	60	500	0
16	1	1	54	221	0
17	1	1	58	154	0
18	1	1	70	203	0
19	1	2	82	220	0
20	1	2	70	210	0
21	1	2	62	330	0
22	1	2	48	320	0
23	1	2	54	300	0
24	1	2	61	320	0
25	1	3	82	400	0
26	1	3	60	430	0
27	1	3	56	410	0
28	1	4	78		1
29	1	4	75		1
30	1	4	61	600	0

Running the binary logistic model with age included as a predictor causes the Nagelkerke R-squared to equal 1. All the missingness is accounted for by the variables in the dataset. This implies that the missingness is random.

Completely Random Missingness

For missing values to be completely random, they should be independent of both observed and unobserved data, i.e., they should be truly random. Completely random missingness does not introduce a bias and thus minimally affects statistical inferences. This form of missingness, however, rarely happens in real life.

Graphical Visualization of Missing Data

Graphical visualization of missing data can always provide insights into the nature of missing data and how to effectively deal with that missing data. The easiest step is to employ the abilities of many statistical software programs (including Microsoft Excel) that allow for conditional formatting and sorting of the columns. In Excel, for example, apply conditional formatting in the format of color scales to each column. Sorting the columns may let you find a pattern in the missing data. For example, we have used this approach for the table from Example 8.1 (Fig. 8.1).

Another visualization tool that can help is to create a missing value pattern matrix along with a bar chart showing the frequency of each pattern. This is especially helpful in situations where more than one variable has missing values. For example, Table 8.3 shows missing values in multiple variables.

Each pattern in the pattern analysis matrix shows the variables that have missing values with the first row usually being no missing values. The corresponding bar chart should show that the non-missing pattern should be the most common followed by patterns with one variable missing, then two variables missing, and so on. If one pattern has a high frequency, then it should be investigated since it may be due to nonrandom missing. The visual inspection of these graphics will give you clues for how to deal with the missing data. Figure 8.2 shows the matrix and bar chart for Table 8.3 [3].

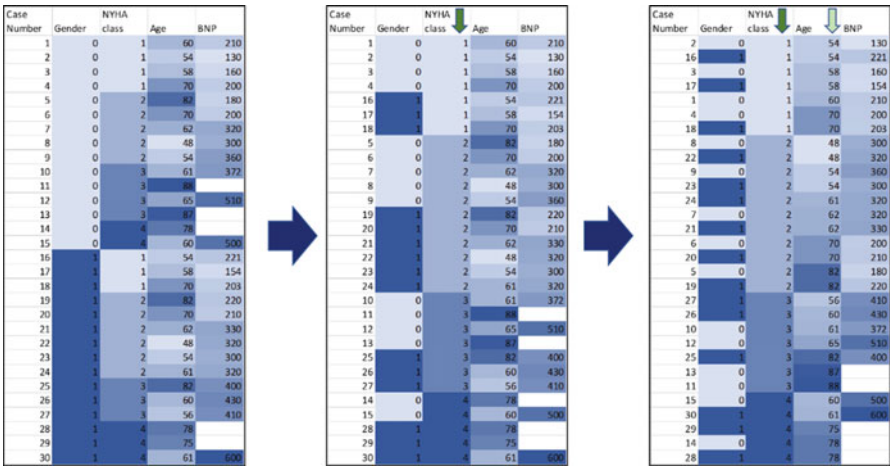


Fig. 8.1 Conditional formatting of Table 8.2 with color scales. The *left panel* shows the color scales applied to the table (darker colors indicate larger numbers within column). In the *middle panel*, the data has been sorted based on NYHA class. You can start to see the missing values of BNP cluster where NYHA class is higher, suggesting a correlation between NYHA class and missingness in BNP values. In the right panel, a second layer of sorting by age is added (first sorting based on NYHA class and then cases within each NYHA class sorted by age). Now you can see a better clustering of missing BNP values. The color pattern also suggests that a combination of high age and high NYHA class may account for the observed missingness

Table 8.3 Table of results from the BNP study with more variables having missing values

Case number	Gender	NYHA class	BNP	Age
1	0	1	210	60
2	0	1	130	54
3	0	1	160	58
4	0	1	200	70
5		2	180	82
6	0	2	200	70
7	0	2	320	
8	0	2	300	48
9	0	2	360	54
10	0	3	372	61
11	0	3		88
12	0	3	510	65
13				87
14	0	4		78
15	0	4	500	60
16	1	1	221	54
17	1	1	154	
18	1	1	203	
19	1	2	220	82
20	1	2	210	70
21	1	2	330	62
22	1	2	320	48
23	1	2	300	54
24	1			61
25	1	3	400	82
26	1	3	430	60
27	1	3	410	56
28	1	4		78
29	1	4		
30	1	4	600	61

Dealing with Missing Data

Currently, there is no consensus on dealing with missing data. Many papers fail to report missing data or how the missing data was dealt with, and this can be a cause for concern as missing data can introduce bias in the studies. Overall, it seems that a small proportion of missing data (5% or less) especially random missing can be ignored in statistical inferences. When the proportion reaches 10%, noticeable bias will be introduced in the inferences unless the missing data is missing completely at random. As the proportion of missing values increase, the statistical power and reliability of the statistical test decrease. Databases with missing data proportions of 20% and more should be examined carefully with acknowledgment of the possibility for major bias in the data.

In this section, we will discuss solutions for dealing with missing data starting with the concept of "robust statistics."

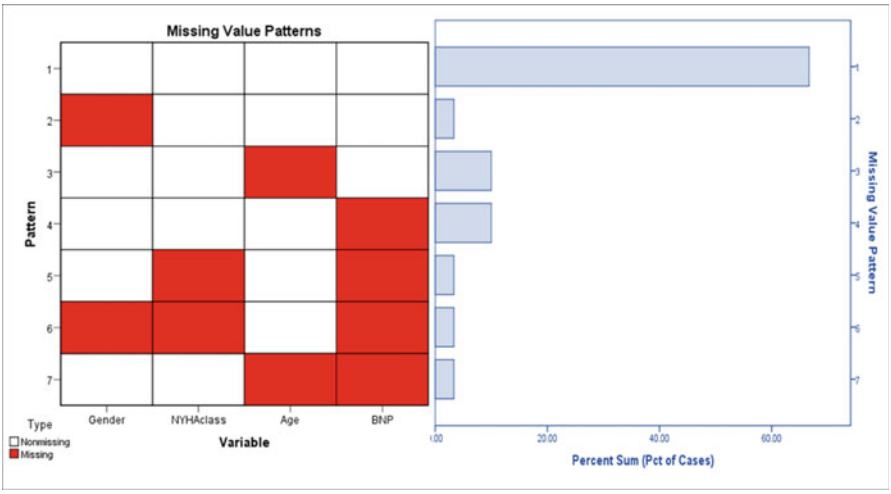


Fig. 8.2 Missing value pattern matrix and associated bar chart for Table 8.3

Robust Statistics

When the dataset has missing data, statistical inference should use statistical tests that are less prone to be affected by the missing data. It is important to remember that it is nearly impossible to know the exact nature of the missing data. For example, if we are measuring troponin levels from 100 healthy individuals and we have one missing value, the assumption is that, since the individuals are healthy, the missing value is within the distribution of the observed values (especially if the results follow a normal or near-normal distribution). However, the missing value may in fact be a significant outlier. Thus, tests that are less likely to be affected by outliers are more robust in dealing with missing data.

The more assumptions a statistical test makes about the nature of the data, the less robust it will be in dealing with missing data. Parametric tests, for example, make many assumptions about the data and are especially affected by outliers, and nonparametric tests are more robust in dealing with missing data.

Thus, while it is possible to identify and account for missing data in a dataset, it is still advisable to apply robust statistical measures to further reduce the possible bias in the statistical inference process.

Data Discarding Solutions

If the proportion of missing data is small and the sample size is large enough, then one solution to the missing data problem is to disregard observations that have missing value. Especially if the missingness has occurred at random, then the missing data can be "ignorable." One problem with discarding methods is that the

sample size will decrease leading to the loss of statistical power and increase in the standard errors of measurement. Furthermore, a principle in clinical trials is to do statistical analysis on an intention-to-treat basis, i.e., analyzing all the patients who were enrolled in the study. Dropping observations with missing data violates this principle.

On the other hand, data discarding solutions are often the easiest way for dealing with missing data, and, as such, they are widely employed. Most statistical software will employ data discarding solutions by default unless the user opts for an alternative approach to missing data. Data discarding includes two main solutions: "complete-case analysis" and "available-case analysis."

Complete-Case Analysis

In this approach (also known as "list-wise deletion"), only observations (or cases) that are complete with no missing data are analyzed. As we said before, this can introduce significant bias if the excluded cases would have had values that were statistically different from the complete cases. Also, this approach has the potential of excluding a big portion of the data and distorting the results of the statistical analysis.

Nonresponse Weighting

One way to reduce the bias in the list-wise deletion of cases is to make the remaining cases more representative of the entire set (to give them high weights). This is done through nonresponse weighting. The principle is that the difference between complete cases and cases with missing values can be used to calculate a weight for the remaining cases with the variables reweighted based on this calculated weight.

There are different weighting approaches; here we will briefly discuss a weighting approach called "propensity cell method." As we mentioned in the previous section, for missing at random, a binary logistic regression using an indicator missingness variable can be formed with all the remaining variables serving as predictors. We can ask the binary logistic regression to assign a log probability of responding (being complete) for each case. The weighting adjustment (or inflation) factor is then calculated as the inverse of this value. Many statistical software programs allow for inclusion of this weighting adjustment factor into the statistical test (e..g., WLS weight option in the commercially available IBM Statistical Package for the Social Sciences or SPSS for general linear models).

Available-Case Analysis

This approach is also known as "pair-wise deletion." In this approach, the observation is only excluded for the analytical test that uses that variable. For example, if we have three variables and observation *i*th has a missing value for the second variable, in pair-wise deletion, the *i*th observation is still included for comparison of

the first and third variable. For example, if we have three variables and there are ten samples, the data for variable 1 is complete, the data for variable 2 is missing for first sample, and data for variable 3 is missing for the last sample. In this case if you are comparing variables 1 and 2, then you can use observations 2 through 10. However, for comparison of variables 1 and 3, we can use observations 1 through 9. The problem with this approach is that different statistical tests in this context will use different subsets of the data, and thus the results may not be consistent or even comparable with each other.

While this approach addresses loss of statistical power to an extent, it still has the potential to introduce bias into the inferences.

Imputation

Discarding missing data, while simple, is often problematic. Another option for dealing with missing data is to "impute" their value or, in other words, fill in their value. These approaches are attractive since they do not change the sample size and avoid some of the bias introduced by case exclusion. They, however, may introduce some bias of their own.

Single Imputation

Single imputation implies that the missing value is filled in using data from other variables and the analysis is done on the completed dataset. There are different methods of single imputation which we will discuss below. It must be noted that single imputation tends to have small standards of error for the imputed value, but conversely this does not imply that the filled-in value is accurate; it only reflects the fact that we are making significant assumptions about the data and the missing values.

Adjacent Value

One way to fill in the missing values is to use previously available data for the subject. For example, if in a study of serum sodium levels, one of the subject's test result is missing but we have a result from the same patient from a previous observation, then we may fill in the missing value with that result. This approach is especially useful in longitudinal studies where a test subject may have multiple results over time, and thus the nearest temporal result can be used to fill in the missing value.

This approach is a conservative method which often underestimates the treatment (or exposure) effect, but in some situations, it can have an opposite effect.

A variation of this approach is to calculate the mean of the two adjacent measurements and use it as the missing value, but this requires that an observation is made before and after the missing observation.

Mean Imputation

This is perhaps the easiest imputation method and involves replacing the missing value with the mean of that variable for all cases (and not within group). This can lead

to reduced statistical power as the imputed value is the same across groups and any differences between them will be attenuated by the inclusion of mean imputation.

Random Imputation

In this approach, the missing value is replaced with a randomly drawn value of the variable from the dataset. One of the random imputation approaches is known as a "hot-deck" approach. In this approach the observations are ordered based on the variables with non-missing values, and the variable value for the observation next to the missing value is carried over to the missing value. This approach while simple has minimal utility as again it erodes the statistical power of the test and may also introduce bias.

Regression Imputation

"Regression imputation" involves fitting a regression model to the data with the response variable being the variable with the missing data and the predictors being all the other variables in the dataset. The regression is performed only on the complete cases. If a model with good fit is found, then the model can be used to estimate the missing values. If the values predicted are deterministic and the associated uncertainty with regression modeling is not included in the estimates created by the model, then this model tends to overestimate the statistical effects between groups since the data fits perfectly along the regression line, augmenting any difference between the groups. To address this, stochastic regression is used where an error term is introduced in the regression estimates (usually the average regression variance). This reduces the overestimation bias but still does not nullify it. Hence, more complex approaches such as multiple imputation are needed [4].

Example 8.2

A study has been conducted looking at different variables to determine that the patient has an acute coronary event (ACS). The occurrence of myocardial infarction was determined using coronary angiography. The results of the study are summarized in Table 8.4.

From the 30 patients in the study, five patients did not undergo coronary angiography, and, as a result, their group variable value is missing. We can use regression modeling to impute their values based on the other predictors. Since the missing variable is binary, we will run a binary logistic regression on the observation with complete values. The predicted values for observations 18, 19, 20, 26, and 27 based on regression are 1, 1, 1, 2, and 1, respectively. A closer look at the data, however, may cause you to question the findings of the regression model; for example, case number 27 has a troponin value of 0.74 μg/L, in the study, and the laboratory cutoff for troponin was 0.04 μg/L, hinting that the patient might indeed have had a myocardial infarction. Yet the regression model determined that the patient did not have a myocardial infarction.

Thus, regression is not always the best solution to imputing the missing values, especially in cases like our example where either the size is small for a model with good fit to be found or there are no predictors in the data.

Table 8.4 Summary of results for Example 8.2. Notice that five patients have missing values for group (highlighted in red)

Case Number	Group (1 : No ACS, 2: ACS)	Troponin level (μg/L)	Chest Pain	Sweating	Dyspnea	History of coronary disease	Hypertension	Congestive heart failure	Diabetes
1	1	.02	1	0	0	0	1	0	1
2	1	.02	0	0	0	0	1	0	1
3	1	.02	0	0	0	0	1	0	1
4	1	.02	1	1	1	0	0	0	0
5	1	.02	1	0	0	0	1	0	0
6	1	.02	1	0	0	0	1	0	0
7	1	.02	0	0	1	0	0	0	0
8	2	15.65	1	0	0	1	1	1	0
9	2	.03	0	0	0	1	1	1	0
10	2	19.08	1	1	1	0	1	1	1
11	2	.19	1	1	1	0	1	0	0
12	2	.07	0	0	1	0	1	1	1
13	2	.19	1	1	0	1	1	1	0
14	1	.04	1	0	0	1	1	1	1
15	1	.02	0	0	0	0	1	0	1
16	1	.02	0	0	0	0	0	0	0
17	1	.02	0	0	1	0	1	0	0
18		.24	0	0	0	0	1	0	1
19		.07	0	0	0	0	0	0	1
20		.02	0	0	1	0	1	0	1
21	1	.02	1	0	0	0	0	0	0
22	1	.02	1	0	0	0	0	0	0
23	1	.02	1	0	1	0	1	1	0
24	1	.02	1	0	0	0	1	1	0
25	1	.02	1	1	1	0	1	0	1
26		.02	0	0	1	1	1	1	1
27		.74	0	0	0	0	1	0	1
28	1	.02	1	0	0	0	1	0	0
29	1	.02	0	0	0	1	1	0	0
30	1	.02	0	0	0	0	1	0	1

In fact, in some situations (like our example), there may be proxy measures based on which missing value can be guessed. Thus, in our example, the missing values that have a troponin level of more than 0.04 μg/L can be given a value of 2 with the rest given a value of 1.

Multiple Imputation

In regression imputation, the imputed values lack the variance that the actual data could have had; this approach will only serve to reinforce the underlying patterns of data that the non-missing values had and thus leads to an overestimation of

statistical tests. An effective way to handle missing data, which minimizes bias, is called "multiple imputation." This method has three steps: imputation, analysis, and pooling. As the name implies, this method imputes multiple probable values for the missing data and each time runs the test statistic with one set of the imputed values. Thus, multiple sets of statistical tests will be run, each with their own test statistic. The next step will be to reconcile these multiple tests and summarize them into a single test statistic: this step is called pooling.

The main challenging step in this approach is the imputation whereby multiple completed datasets are created. Most statistical software currently employ "multiple imputation by chained equations method" (MICE) or "Markov chain Monte Carlo method" (MCMC) for this step. This method assumes that the missingness is at random. For the MCMC an additional assumption is that the missingness is monotonous. This means that if a variable is missing a value for an observation, all subsequent variables are also missing values for that observation. Other approaches employ multivariate regression, propensity scoring, or combinations of these to calculate the missing values for each imputation run.

The statistical test is then run on each completed dataset. The statistical tests do not need to adjust for missing values anymore and for each completed dataset; a test value as well as significance level will be recorded. The pooling of these tests is simple and involves calculating the mean of the test statistic over the multiple completed sets as well as its variance and p-value.

The variance and subsequently standard error of the test statistic (b) is calculated using the within-imputation variance (U_b) and between-imputation variance (B_b). These two measures are combined to form total variance (T_b) from which the pooled standard error of the test statistic (SE_b) is calculated.

The within-imputation variance is calculated by summing the mean of squared standard error of test calculation for each imputation averaged over the number of imputations (m):

$$\frac{U_b = \sum_{i=1}^{m} {SE_{bi}}^2}{m,} \tag{8.1}$$

where m is the number of imputation runs and SE_{bi} is the calculated standard error of test statistic in each run.

The between-imputation variance is calculated from the variation of the statistic in the different imputation runs:

$$B_b = \sum_{i=1}^{m} \frac{(b - \bar{b})^2}{m - 1}, \tag{8.2}$$

where b is the test statistic for each run (e.g., t-test statistics for each run) and $\bar{b}$ is the average test statistic over the multiple runs.

The total variance can be calculated as

$$T_b = U_b + \left(1 + \frac{1}{m}\right) B_b, \tag{8.3}$$

And the SE_b is the square root of the total variance.

In order to calculate the p-value, the average test statistic ($\bar{b}$) is divided by its overall standard error (SE_b). This value is a t-value that follows a t-distribution with the degrees of freedom calculated from the following formula:

$$df = (m-1)\left(1 + \frac{mU_b}{(m+1)B_b}\right)^2, \tag{8.4}$$

This number is truncated to the nearest integer. An important decision in multiple imputation is how many imputation runs are needed. Usually, 3–5 imputation runs are enough to produce robust statistical inferences. It has been shown that after the first few imputations, possible gains in efficiency of multiple imputation diminish, and thus higher number of imputations are deemed unnecessary. The efficiency of the imputation model can be calculated:

$$\text{Efficiency} = \frac{1}{1 + \frac{\gamma}{m}}, \tag{8.5}$$

where γ is the fraction of missing information:

$$\gamma = \frac{r + 2/(df+3)}{r+1}, \tag{8.6}$$

where r is the relative increase in variance due to nonresponse and can be calculated using the following equation [5–8]:

$$r = \frac{\left(1 + \frac{1}{m}\right) B_b}{U_b}, \tag{8.7}$$

As the number of imputations increase, the fraction of missing information becomes smaller, and thus the gains in efficiency will become smaller. However, if the number of missing values is high, then more imputations are needed, and the efficiency of the model must be checked to see if additional imputations will lead to better efficiency or not [9, 10].

Example 8.3

Going back to Table 8.4, now we run a multiple imputation with ten imputations, to calculate the missing values. The model predictions for the missing values for each imputation are shown in Table 8.5.

Now, we can run 10 binary logistic regressions (Chap. 7) to determine if the measured variables can be a predictor of whether the patient has had an acute

Table 8.5 Imputed values for the missing values of group for 10 imputation runs

Case number	Imputation 1	Imputation 2	Imputation 3	Imputation 4	Imputation 5	Imputation 6	Imputation 7	Imputation 8	Imputation 9	Imputation 10
18	1	2	2	1	2	1	2	1	1	2
19	1	2	2	1	2	1	2	1	1	2
20	1	2	2	1	1	1	2	1	2	1
26	2	1	1	2	1	2	2	2	2	1
27	1	2	2	1	2	1	2	1	1	2

myocardial infarction. For each binary logistic regression run, one set of imputed values is used. Finally, the results from the 10 runs are pooled into a single statistical output. In this case, the pooled data showed that none of the input variables are good predictors of the outcome.

Summary

Most statistical tests assume completeness of the data and will ignore observations with missing data. Missing data is a major source of bias in statistics and can result in incorrect inferences from the data. The missing data can be dealt with by discarding the observations with missing value which, unless the sample size is sufficiently large, can lead to significant loss of statistical power and bias. Another solution is to impute the missing data. Single imputation is relatively easy and imputes the missing values based on the other values in the dataset. This, however, will lead to increased bias of the test results. A more attractive solution is to use multiple imputations where multiple completed datasets are created and the tests are run on each with the results pooled into a single statistical value.

The question of how missing data affects the statistical validity of results is very important, and appraising studies and literature require the reader to be able to identify in a paper how missing data was handled and whether the assumptions of the authors for missing data were valid. We will address in further depth appraisal of the literature in Chap. 12.

References

1. Dong Y, Peng CY. Principled missing data methods for researchers. SpringerPlus. 2013;2 (1):222.
2. Shara N, Yassin SA, Valaitis E, Wang H, Howard BV, Wang W, Lee ET, Umans JG. Randomly and non-randomly missing renal function data in the strong heart study: a comparison of imputation methods. PLoS One. 2015;10(9):e0138923.
3. Zhang Z. Missing data exploration: highlighting graphical presentation of missing pattern. Annals Transl Med. 2015;3(22):356.
4. Little RJ, Rubin DB. Single imputation methods. In: Statistical analysis with missing data. 2nd ed. Chicester: John Wiley and Sons; 2002. p. 59–74.
5. Gelman A, Hill J. Data analysis using regression and multilevel/hierarchical models. Cambridge: Cambridge university press; 2006.
6. Royston P. Multiple imputation of missing values. Stata J. 2004;4(3):227–41.
7. Little RJ, Rubin DB. Bayes and multiple imputation. In: Statistical Analysis with Missing Data. 2nd ed. Chicester: John Wiley and Sons; 2002. p. 200–20.
8. Yuan YC. Multiple imputation for missing data: concepts and new development (Version 9.0), vol. 49. Rockville: SAS Institute Inc; 2010. p. 1.
9. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prev Sci. 2007;8(3):206–13.
10. van Ginkel JR, Kroonenberg PM. Analysis of variance of multiply imputed data. Multivar Behav Res. 2014;49(1):78–91.

Survival Analysis

9

Introduction

While pathology and laboratory medicine are the cornerstones of diagnostic medicine, they also play a crucial and central role in prognostication. The role of anatomic pathology in determining disease stage and prognosis is well known. Clinical pathology also contributes greatly to the prognostication process: Many clinical decisions and predictive models depend on laboratory values, for example, the Childs-Pugh score used for assessment of prognosis of chronic liver disease uses a combination of clinical criteria (ascites and hepatic encephalopathy) with laboratory criteria (total bilirubin, serum albumin, and prothrombin time). Practicing pathologists' design and analysis of survival data may not occur in your routine clinical practice; however, many of the clinical decisions that you make are based on survival analysis data, and this requires you to understand the fundamental basics of survival statistics. In this chapter, we will explain some of the pertinent subjects relating to survival analysis. We will start with defining incidence.

Incidence

"Incidence" is the number of cases of a disease or condition that occurs in a defined area, over the course of a defined time period, usually 1 year. For example, a statement such as "10 cases of malaria were recorded last year" is stating an incidence of malaria in a 1-year period. However, such statements can be better stated as a probability or proportion; stating the number of occurrences without stating the denominator does not allow you to judge the magnitude of the health outcome. Taking the above statement, for example, if we rephrase the statement as "10 cases of malaria were recorded last year in a 100-person village," then it becomes a more meaningful statement.

Understanding incidence concepts is important in survival analysis, because incidence is a general term relating to the probability of an event occurring

A. Momeni et al., *Introduction to Statistical Methods in Pathology*,
DOI 10.1007/978-3-319-60543-2_9

(which can be metastasis, cirrhosis, etc.). Thus, "death" can also be an event that can be expressed as an incidence.

The "cumulative incidence" (incidence proportion) is a better outcome measure than incidence alone which is expressed using "incidence rate": The number of new cases per population at risk in a specified time period is called incidence rate. The denominator is the size of the population at risk at the beginning of the time period, and the numerator is the number of occurrences of the disease in that population during that time period. For example, the above statement can be stated as "Malaria has an incidence rate of 10 per 100 in a year."

It is important to know that with the exception of population-based epidemiologic studies, where it is possible to follow an entire population for occurrence of a disease, for most small-scale studies, it is better to use "incidence density rate" also known as the "person-time incidence rate." Here, the denominators will not be "per person" but "per person-time." This denominator can be calculated using:

$$\text{Person-time} = \sum_{i=1}^{N} t_i \tag{9.1}$$

where N is the number of at-risk people in the population who were observed and t_i is the amount of time that person was observed (or followed). This is especially useful in cases where people were not followed up for equal amounts of time; for example, person A was followed for 6 months and person B was followed for a year. Thus, incidence density rate statement will be something like "10 malaria cases per 1000 person-years were reported."

It must be noted that incidence density rate assumes a constant rate of occurrence of an event over time, i.e., the statement of "10 malaria cases per 1000 person-years" can be interpreted as 10 cases occurring in a population of 1000 in a 1-year period or 100 cases occurring in a population of 100 over a 10-year period. This basic assumption is sometimes wrong. For example, in survival statistics of cancer, stating 10 deaths in a population of 1000 in a 1-year period is not the same as 10 deaths in a population of 100 over a 10-year period; if 100 individuals with cancer are followed for 10 years and only 10 deaths occur in 10 years, then that cancer has a very indolent almost benign course. In fact, cancer survival statistics show that as time passes by, the probability of mortality caused by the cancer increases. For example, in the first year of follow-up, 10 people out of each 1000 individuals with cancer die, but in the second year of follow-up, 30 people out of each 1000 individuals with cancer die.

Thus, survival statistics in general benefits from a cumulative incidence approach. However, we must still account for different lengths of follow-up in patients, and this is solved through a concept known as censoring which we will discuss later in this chapter.

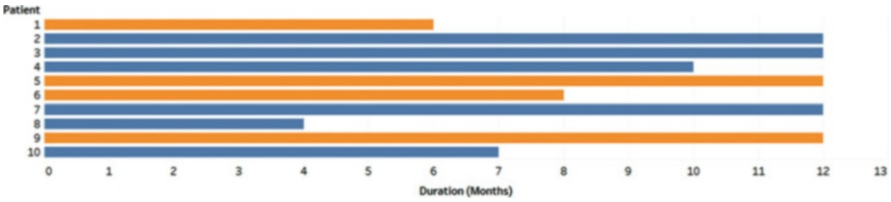


Fig. 9.1 Incidence and follow-up time of ten patients for Example 9.2. The orange bars signify the patients in whom disease X occurred

Example 9.1
Q: Figure 9.1 shows incidence of disease X in a population of ten who were followed up to a year. What are the cumulative incidence and incidence rate for this disease?

A: Four patients developed disease X during a 1-year period. Thus, the cumulative incidence is four per ten persons in a year.

The sum of all the follow-up time in these groups was 95 months (7.91 years). Consequently, the incidence density rate can be stated as 4 cases per 95 person-months or 50.56 cases per 100 person-years.

Survival Analysis

Survival is the time from onset of a disease (or more commonly from time of diagnosis) to the patient's death. While survival statistics generally refers to mortality, it may also refer to other events such as recurrence, metastasis, readmission, and so on. Thus, a better term for designating the events related to survival analysis is "failure." Survival analysis then is a study of time to failure data. In this sense, survival analysis needs at least two measures (variables) for each individual: event variable (showing whether the outcome has occurred or not) and time variable (either stored as duration or stored as a start date and an event date).

Survival analysis uses sets of statistical tests different from the usual linear regression models that we introduced in previous chapters; survival data is dependent on time thus making simple regressions less accurate. Furthermore, survival data tends to be incomplete (have censoring), making simple regression even further unreliable.

In this chapter, we will discuss three statistical methods used in survival analysis: Kaplan-Meier curves, Log-rank test, and Cox-proportional hazards regression. Before that, however, we need to explain the concept of "censoring" and "survival/hazard functions." These two functions are the theoretical basis for many of the statistical methods used in survival analysis. However, you may skip them and move to the next sections of this chapter where we have focused more on the practical aspects of survival analysis.

Censoring

As we saw with incidence density rate, the length of follow-up or observation is not uniform among all subjects of a study. For close-ended survival analysis (i.e., 5-year survival), ideally all patients are followed either for the entirety of the study period or until the time the event occurs. For example, if we are studying the 5-year survival of colon cancer, then, ideally, we would like all our patients to be followed for 5 years or until the time of their death. In other words, an ideal dataset would be complete without any dropouts. If we follow a cancer patient for a year and then he/she is lost to follow-up, then we cannot be sure whether the patient has survived for the remainder of the 5 years or whether he/she has died. Unfortunately, most survival studies will not have complete datasets.

Subjects who have incompletely observed responses are called "censored." Censoring is similar to missing data (see Chap. 8). In censoring, time to the last observation is recorded instead of time to event.

Censoring is usually right sided; that is, it is known that the time of event in the patient is after a certain date, but until the last follow-up date, the event has not occurred. For example, in survival analysis, we know that an alive subject who was lost to follow-up will inevitably die, yet we do not know the exact time it will occur, only that it should happen sometime after the last follow-up (called random type I censoring). Close-ended survival analysis is a type of right-sided censoring (called fixed type I censoring) in which the study is designed to end after a certain amount of time of follow-up; thus, anyone who does not experience the failure event and completes the study is said to be censored at N years (with N being number of years of follow-up). Another survival study design follows a type II censoring where the study ends after a specified number of failure events have occurred.

There are instances where censoring is left sided: for example, when a subject's lifetime is known to be less than a certain amount, but the exact time is unknown. For example, if we want to study time to metastasis in patients and if a patient has metastasis at the onset, then that patient is left-censored, i.e., the event has occurred sometime before the beginning of the study.

Censoring is noninformative, i.e., we cannot assume anything about the patient after the patient is censored. For censored observations, we can only use the data up to the point of censoring.

Another concept is truncating; in truncation, we are not aware that an event has occurred. For example, in a study of cancer survival, we study the survival from the time of diagnosis to death. If a patient dies from cancer before a diagnosis is even made, then that patient is truncated.

Survival Data

Survival data consists of four parameters (variables): For each patient, time to failure event and/or time to censoring is recorded (sometimes failure event is not death; thus, the patient is actually not censored after the event occurs, and you may

have both censoring time and failure time). The smaller of these two times is the observed response time. For each patient, an indicator variable is also needed: This variable assumes a value of 1 if the event has occurred or a value of 0 if the patient is censored before the event occurred. The survival analysis is usually performed using the observed response time and indicator variable [1, 2].

Survival Function

"Survival function" ($S(t)$) is the probability that a patient survives for more than a specified time (t). t or time is a positive continuous variable with a range of [0,∞]. At the beginning, the patient is alive ($t = 0, S(t) = 1$) as the time increases toward ∞, and the probability of survival decreases toward 0 ($t = \infty, S(t) = 0$). The survival function can be stated as:

$$S(t) = P(\{T > t\}) = \int_t^{\infty} f(u)du = 1 - F(t) \tag{9.2}$$

T here represents event time; thus, the probability that the patient is alive at t is the same as the probability that $T > t$. This can be restated as 1 minus the cumulative distribution function of t ($F(t)$). As you may recall from Chap. 3, cumulative distribution function (or distribution function) is the probability that a value (t) is smaller than a set value (T). That is, the cumulative probability of survival at each t translates to the probability that the patient has died before that time ($T > t$). This means that at each moment, the probability of a patient surviving is 1 minus the cumulative probability of dying up to that point.

The survival function can be shown as a graph, where the Y-axis shows probability of survival and the X-axis shows time. As time increases, the probability of survival always decreases (Fig. 9.2).

Hazard Function

"Hazard function" ($h(t)$) is the instantaneous rate of occurrence of the failure event. In simple terms, the hazard function is determined by calculating the changes in failure rate in ever smaller intervals:

$$h(t) = \lim_{\Delta t \to 0} \frac{S(t) - S(t + \Delta t)}{S(t) \times \Delta t} \tag{9.3}$$

Probability of survival is a continuous probability distribution that is highest at the start of the time period and continually decreases (see Chap. 3). In fact, survival probability follows what is known as "exponential distribution" which is a type of gamma distribution. Thus, just like any other continuous probability, it has a probability distribution function ($f(t)$) and a cumulative distribution function $F(t)$:

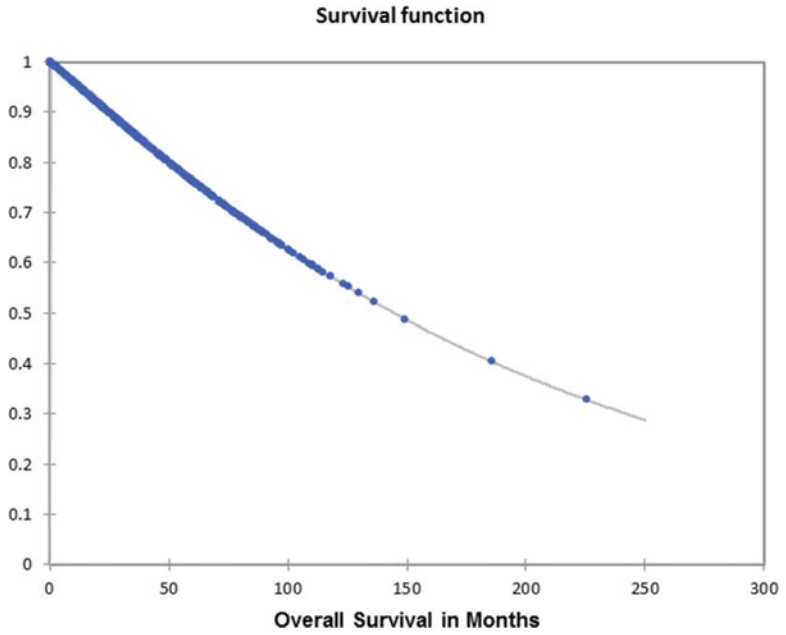


Fig. 9.2 The survival distribution function for a disease is plotted in this figure

$$f(t) = \lambda e^{-\lambda t} \qquad for\ t \geq 0 \tag{9.4}$$

where λ is the gamma which determines the shape of the distribution, e is the mathematical constant (approximately equal to 2.71828), and t is time.The cumulative distribution function of survival can be given by:

$$F(t) = \int_0^t \lambda e^{-\lambda t} dt = 1 - e^{-\lambda t} \tag{9.5}$$

We can actually state the hazard function using these two terms: The hazard function is the ratio of the probability distribution function of time ($f(t)$) to the survival function ($S(t)$):

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)} \tag{9.6}$$

Combining Eqs. 9.4, 9.5, and 9.6, we can see that the shape of survival distribution (λ) is actually the hazard function. Thus, in simple terms, hazard rate (hazard function) determines the shape of the survival distribution curve:

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)} = \frac{\lambda e^{-\lambda t}}{1 - (1 - e^{-\lambda t})} = \lambda \tag{9.7}$$

Hazard function can also be derived directly from the survival function (Fig. 9.3):

$$h(t) = -\frac{\partial \log(S(t))}{\partial t} \tag{9.8}$$

Where $\partial \log(S(t))/\partial t$ is the partial derivative of the logarithm of survival function with respect to time.

Next is the cumulative hazard function ($H(t)$) which is the accumulated risk of failure up to the time t. The cumulative hazard function has a reverse logarithmic relation with survival:

$$H(t) = \int_0^t h(v)dv = -\log(S(t)) \tag{9.9}$$

As you can see, all components of survival including survival function, hazard function, and cumulative hazard function are related and can be derived from each other. Thus, if only one of the functions is known, the others can be calculated using the known function.

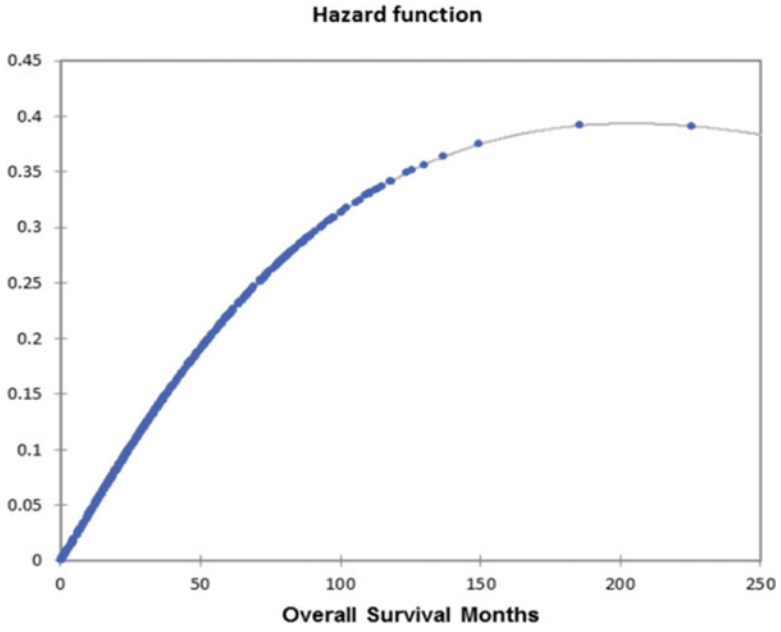


Fig. 9.3 Hazard function corresponding to the survival distribution function plotted in Fig. 9.2

The importance of these functions is that they are the foundations of survival statistics. In fact, survival analysis involves using statistical methods to estimate the survival and hazard functions (assuming that every observation or patient follows the same survival function) [3, 4].

Kaplan-Meier Estimator

The goal of survival analysis is to estimate the survival function of the patients; it is usually important to know the probability of survival at different intervals as well as changes in survival statistics based on some parameters (e.g., cancer stage) or interventions.

If the observations are complete without any censoring (i.e., we know the time of failure for all patients), then we can estimate the survival function by calculating the cumulative distribution function of the data ($F(t)$) and use that as a non-parametric estimator of the survival function ($S(t) = 1 - F(t)$).

In reality, however, most survival datasets have some patients who are censored. In these instances, we can use "Kaplan-Meier estimator" (product limit estimator) statistics to estimate the survival probability function. This method can show the proportion of patients surviving for a certain amount of time after an initiating event (e.g., after a diagnosis of cancer is made).

The Kaplan-Meier estimator is usually shown as a plot with a series of stepwise declining horizontal lines which resemble a survival distribution function plot in that the Y-axis is the proportion of patients alive and the X-axis is time. The right-censored patients are shown as small vertical tick marks on the survival curve. If the observations are grouped, then the Kaplan-Meier estimator can be used to show the estimated survival function for each group [1].

Example 9.2

Q: The survival data of ten patients who were diagnosed with pancreatic cancer and were followed for up to a year is shown in Table 9.1. What is the Kaplan-Meier estimator corresponding to these data?

Table 9.1 Survival data for Example 9.2. Event indicator variable shows whether patients died (1) or were censored (0)

Patient number	Event indicator	Observed response time
1	1	3
2	1	4
3	0	12
4	1	2
5	0	12
6	1	5
7	1	8
8	0	10
9	0	12
10	1	10

Table 9.2 The cumulative proportions of survival, for example, are shown in this table. Values with "a" indicate that the observations were censored

Event indicator	Observed response time (months)	Cumulative proportion surviving at the time
1	2	0.90
1	3	0.80
1	4	0.70
1	5	0.60
1	8	0.50
1	10	0.40
0	10	0.40[a]
0	12	0.40[a]
0	12	0.40[a]
0	12	0.40[a]

A: To draw the Kaplan-Meier estimator, first, we need to order the observed response times and estimate the proportion of patient surviving at each observed time. The results are shown in Table 9.2.

In this table, note that the first observed response time is one (1) patient who survived for only 2 months. Thus, right after 2 months, the cumulative survival rate is 0.9 (nine of ten patients are still alive). Right after 3 months (observed response time 3), another patient expired, leaving eight out of the original ten patients or a cumulative fraction of survivors that is 0.8.

Now we can use the cumulative proportion of surviving and the observed response time to draw the survival estimator (Fig. 9.4). Censored observation will be shown by a vertical tick on the plot.

Kaplan-Meier survival curves are easy to interpret and understand. The rate of decline and curve median (the point where half of the patient have failure events) are some of the useful metrics that can be extracted from the curve. For example, in Fig. 9.4, we can see that the median survival is 8 months: That is, half of the patients will die within 8 months of diagnosis, i.e., the cumulative survival (cum survival on the Y-axis) of 0.5 corresponds to an observed response time of 8 months on the X-axis. This means that, just after 8 months, only half of the patients are still alive.

Kaplan-Meier is a non-parametric estimator of survival function, and as such it makes no assumption about the data or its distribution. While this is advantageous in that it can be globally applied to survival data, it also prevents us from extracting some useful information from the estimator. For example, if we want to estimate the expected failure time for a patient, then we must use parametric estimators. Parametric estimators provide smoothed survival curves in comparison with the stepwise survival curves of Kaplan-Meier estimators and can be more accurate than non-parametric approaches if the assumptions of parametric distribution are correct.

Parametric survival estimators include Weibull, exponential, log-normal, and logistic methods. Out of these, the exponential model is the simplest form and is

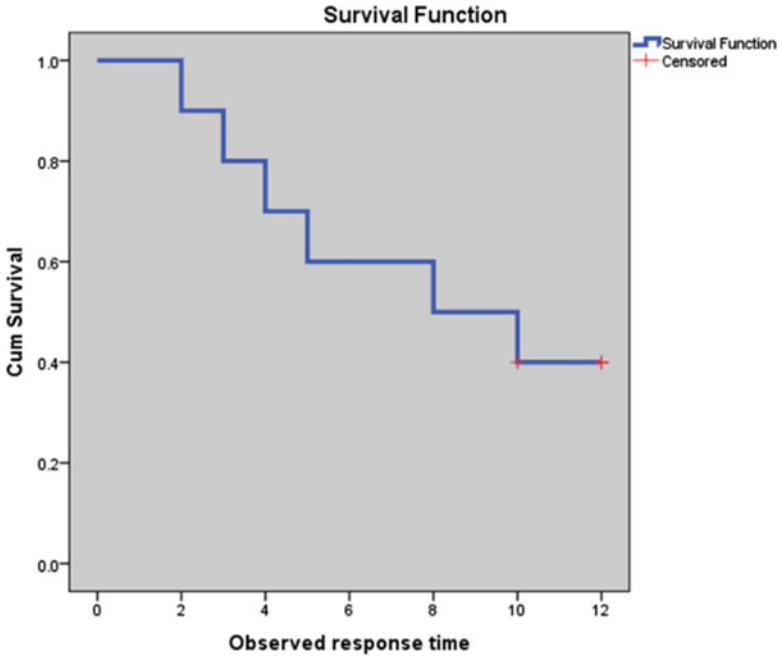


Fig. 9.4 Kaplan-Meier curve for Example 9.2

Table 9.3 Different parametric models used in survival estimation

Parametric model	Survival function ($S(t)$)	Hazard function ($h(t)$)
Exponential	$e^{-\lambda t}$	λ
Log-logistic	$\frac{1}{1+\lambda t^p}$	$\frac{\lambda p t^{p-1}}{1+\lambda t^p}$
Weibull	$e^{-\lambda t^p}$	$\lambda p t^{p-1}$

easiest to compute. The exponential model assumes that the hazard function is a constant ($h(t) = \lambda$) (an assumption that may be wrong). We explained in the previous section how we can derive the survival function from just knowing the hazard constant (see Eq. 9.9). Table 9.3 shows three common parametric distributions used in survival estimation and the corresponding equations for survival function and hazard function.

In Weibull and log-logistic models, in addition to the λ, we also have a p which determine the distribution of survival and its shape. For example, for the Weibull estimation, we can state that time (t) is a continuous variable that has a Weibull distribution with parameters λ and p ($W(\lambda, p)$).

Weibull distributions are continuous and have two parameters: a scale parameter (λ) and a shape parameter (p). Based on these two parameters, the Weibull distribution can have different curves (Fig. 9.5). This distribution is especially well suited for description of survival functions, since it allows us to define how

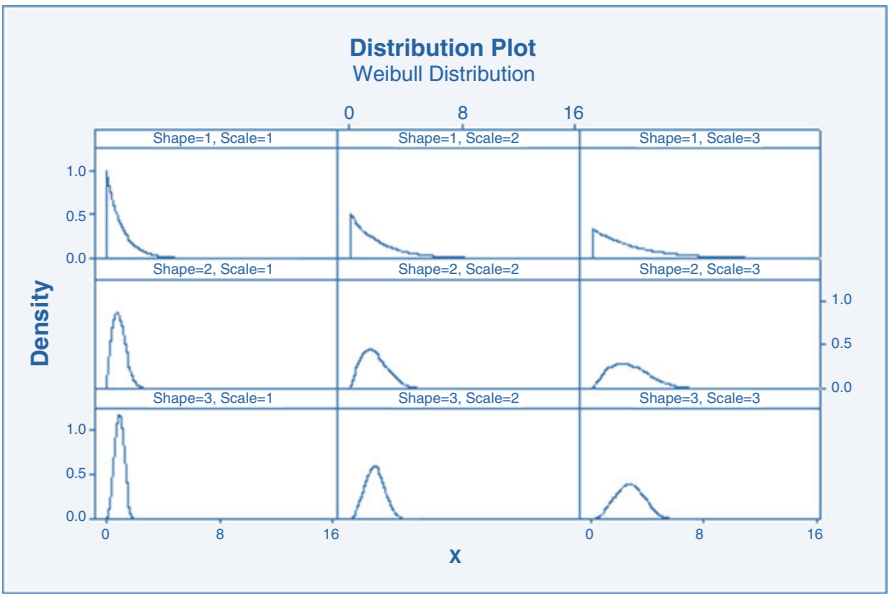


Fig. 9.5 Weibull distribution with different shape and scale parameters

hazard changes over time by using the shape parameter. In Weibull estimators if the shape parameter (p) equals to one, then it means that the failure rate is constant over time; if the shape parameter is between 0 and 1, it means that the failure rate decreases over time (e.g., infant mortality rate); and if the shape parameter is greater than 1, then failure rate increases over time.

Another important property of the Weibull estimation is that the following equation holds true for Weibull distributions:

$$\log(-\log(S(t)) = \log(\lambda) + p\log(t) \tag{9.10}$$

In other words, $\log(-\log(S(t))$ has a linear correlation with logarithm of time. We can use this property to check the fit of the Weibull estimator by plotting the log $(-\log(S'(t))$ versus the logarithm of time and check for linearity (where $S'(t)$ is a Kaplan-Meier survival estimate) (an example is shown in Fig. 9.6).

Log-Rank Test

One of the goals in survival analysis is to compare the survival functions between two or more groups. For example, we may ask questions such as "Is prognosis different between patients with endometrial endometrioid carcinoma and patients with endometrial serous carcinoma?" Answering such questions requires comparison of the survival functions for each group.

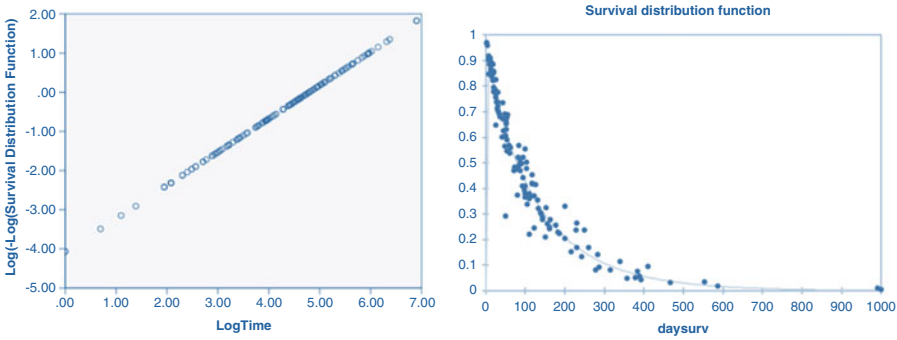


Fig. 9.6 The panel on the left plots the log(-log(survival distribution function)) versus the log (time) which shows a linear relation between the two; this signifies that the survival distribution follows a Weibull distribution. The panel on the right shows the survival distribution function (and the fitted Weibull distribution)

Table 9.4 Contingency group for t_i for two groups

	Group 1	Group 2	Total
Failure event has happened at t_i	d_{1i}	d_{2i}	d_i
Failure event has not happened at t_i	$n_{1i} - d_{1i}$	$n_{2i} - d_{2i}$	$n_i - d_i$
Total	n_{1i}	n_{2i}	n_i

One of the most common approaches to comparison of survival functions is the "log-rank test" which is also known as "time-stratified Mantel-Haenszel test" (see Chap. 5). This test computes the difference using a contingency table for those at risk at each event time.

For example, if we want to compare survival between two groups, then the first step is to order all the event times (any time a failure event occurs). Now for each t_i which is the ith ordered event, we can form a contingency table (Table 9.4).

The null hypothesis will be that the survival function is the same in two groups; thus, the proportion of individuals in each group having a failure at each t_i should be the same. Based on this, we can calculate the expected value of failure events for each group ($\widehat{e}$):

$$\widehat{e}_{1i} = \frac{d_i n_{1i}}{n_i} \quad \text{and} \quad \widehat{e}_{2i} = \frac{d_i n_{2i}}{n_i} \tag{9.11}$$

If there are more than two groups, this equation can be generalized to:

$$\widehat{e}_{ji} = \frac{d_i n_{ji}}{n_i} \tag{9.12}$$

The variance of the expected value will be:

$$\mathrm{Var}(\widehat{e}_{1i}) = \mathrm{Var}(\widehat{e}_{2i}) = \frac{n_{1i}n_{2i}d_i(n_i - d_i)}{{n_i}^2(n_i - 1)} \tag{9.13}$$

These calculations should be made for all the observed event times. Using these values, we can calculate the log-rank test:

$$Q = \frac{\left(\sum_{i=1}^{n} d_{1i} - \sum_{i=1}^{n} \widehat{e}_{1i}\right)^2}{\sum_{i=1}^{n} \mathrm{Var}(\widehat{e}_{1i})} \tag{9.14}$$

Q statistics follows a chi-squared distribution with (k−1) degrees of freedom (k being the number of groups). Thus, for two groups if the Q statistics is more than 3.84, then the test is significant at alpha level of 0.05. That is, it shows that survival of the two groups is different.

Often, we face stratification within our groups; in pathology, a common context where stratification is seen is in tumors which can be stratified based on their clinical stage (or histologic grade). For example, running an unstratified log-rank test, we might see a statistically significant difference between two cancers; however, this difference may occur because the clinical stages of these tumors were different leading to the observed difference. In these situations, running a stratified analysis will account for the difference in clinical stage. As we mentioned in Chap. 5, stratification has its own problems as well; mainly, stratification reduces the statistical power of the test requiring a larger sample size.

For a stratified log-rank test, the denominator remains the same as the unstratified test. For the numerator, the difference between observed values and expected values ($d_{1i} - \widehat{e}_{1i}$) is calculated and summed up within each stratum, and then the results are pooled (summed) across all strata and squared.

Example 9.3

Q: A study compared the survival of 20 patients in a 12-month period, 10 with endometrioid endometrial carcinoma (END, d_{2i}) and 10 with serous endometrial carcinoma (SER, d_{1i}). The results are shown in Table 9.5. Is survival different between the two cancer types?

A: To calculate the log-rank statistics, we can rewrite the table to show the number of events per ordered time for each cancer type (Table 9.6). We can then calculate the expected number of failure events for each ordered time.

Now we can calculate the log-rank test:

$$Q = \frac{\left(\sum_{i=1}^{n} d_{1i} - \sum_{i=1}^{n} \widehat{e}_{1i}\right)^2}{\sum_{i=1}^{n} \mathrm{Var}(\widehat{e}_{1i})} = \frac{(7 - 4.497)^2}{2.299} = 2.725 \tag{9.15}$$

Since 2.725 is less than the critical value (the critical value for a chi-squared distribution with one degree of freedom is 3.84 (See Chap. 5.)), then we conclude

Table 9.5 Table of results for Example 9.3

Patient number	Cancer type	Follow-up time	Event indicator	Patient number	Cancer type	Follow-up time	Event indicator
1	SER	2	Dead	11	END	12	Censored
2	SER	12	Censored	12	END	4	Dead
3	SER	4	Dead	13	END	10	Censored
4	SER	6	Dead	14	END	12	Censored
5	SER	10	Censored	15	END	8	Censored
6	SER	4	Dead	16	END	12	Censored
7	SER	12	Dead	17	END	12	Censored
8	SER	8	Dead	18	END	6	Censored
9	SER	8	Censored	19	END	6	Dead
10	SER	10	Dead	20	END	8	Dead

Table 9.6 Reformulated results for Example 9.3. The expected values and variance for each ordered time are included

Time	Number of failure events		Number at risk		Expected failure events		Variance of expected values
t_i	d_{1i}	d_{2i}	n_{1i}	n_{2i}	$\widehat{e}_{1i}$	$\widehat{e}_{2i}$	$\text{Var}(\widehat{e}_{1i})$
2	1	0	10	10	0.5	0.5	0.250
4	2	1	9	10	1.421	1.579	0.665
6	1	1	7	9	0.875	1.125	0.459
8	1	1	6	7	0.923	1.077	0.456
10	1	0	4	5	0.444	0.556	0.247
12	1	0	2	4	0.333	0.667	0.222
Total	7	3			4.497	5.503	2.299

that survival in two cancer types is not different. The Kaplan-Meier survival curves are shown in Fig. 9.7. As you may notice, the survive curves look different, yet the log-rank test failed to show a statistical significance. Part of the reason for this can be the small sample size. In fact, if the sample size increases, we may observe a statistically significant result.

Cox-Proportional Hazards Regression

There are instances where we want to estimate the effect of some predictor variables on survival (just as regression analysis (Chaps. 4 and 7) where we have a dependent variable and a number of predictors). For example, we may want to determine the effect of histologic grade on the survival of cancer patients.

In survival statistics, we can fit a regression model to survival; this is usually achieved through a so-called Cox regression model. You may recall that, in logistic

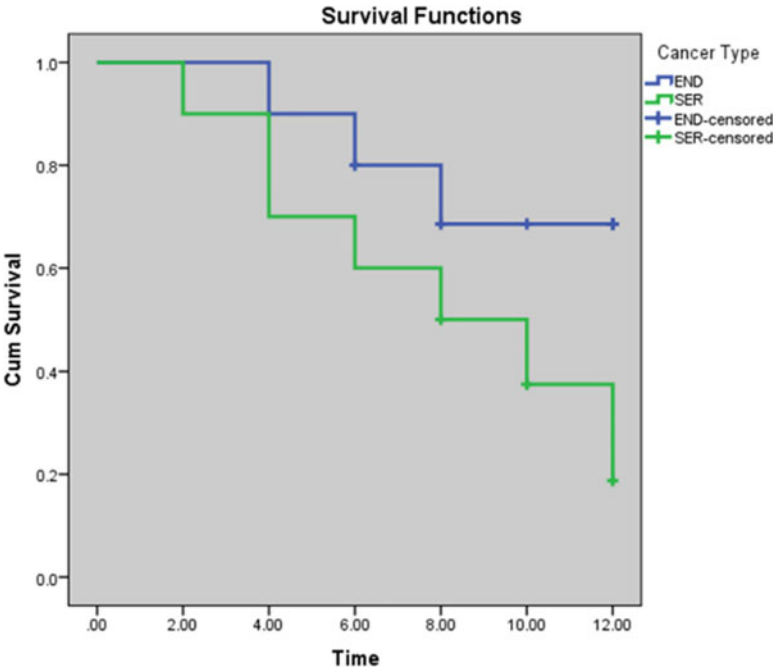


Fig. 9.7 Kaplan-Meier curves for Example 9.3

regression, we use odds ratio for fitting a model and calculating the contribution of each predictor to the regression model. In Cox regression, we use hazard ratio instead (ratio of incidence rates).

Cox regression is a non-parametric model, and thus it does not make assumptions about the probability distribution of the survival data; this makes Cox regression a robust approach to fit a model to survival data. Predictors in Cox regression can be nominal and ordinal.

A general formula for regression models is provided below:

$$y = \alpha + \beta x + \epsilon \tag{9.16}$$

In Cox-proportional hazards model, we can write the regression equation as:

$$\log\left(\frac{\lambda(t|x_{1i}, x_{2i}, \ldots, x_{ki})}{\lambda_0(t)}\right) = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_1 x_{ki} \tag{9.17}$$

where $\lambda(t|x_{1i}, x_{2i}, \ldots, x_{ki})$ is the hazard function for the ith individual at time t, and $x_{1i}, x_{2i}, \ldots, x_{ki}$are the corresponding predictors for that individual. $\lambda_0(t)$ is the baseline hazard function at time t (i.e., $x_{1i} = x_{2i} = \ldots = x_{ki} = 0$).

The hazard ratio represented in the left side of the Eq. 9.17 is the relative risk of the failure event occurring at time t, i.e., the ratio of the risk of the event for a

patient where all predictors contribute to the hazard and the risk of the event for a patient where all predictors are zero. A linear regression model can be fitted to the logarithm of the hazard ratio using the predictors and a Beta-coefficient for each predictor. In this model, changes in the predictors have a multiplicative effect on the baseline risk of the patients:

$$\lambda(t|x_{1i}, x_{2i}, \ldots, x_{ki}) = \lambda_0(t) \times e^{(\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_1 x_{ki})} \tag{9.18}$$

Thus, in Cox regression, the β-coefficient for each predictor is the logarithm of the hazard ratio attributable to that predictor:

$$\mathrm{HR}_{x1} = e^{\beta 1} \tag{9.19}$$

An important assumption regarding the Cox-proportional hazard model is that the attributable hazard ratio of the predictors is constant through time. For example, if a high nuclear grade doubles the risk of death in the first year after diagnosis of a cancer, it also doubles the risk of death in the second year after the diagnosis.

For each predictor, the statistical software will report a p-value which will determine if the predictor is a statistically significant contributor to the hazard ratio or not. The p-value is calculated using the Wald test (see Chap. 7) using the estimated Beta-coefficient and its standard error (in simple terms, the 95% confidence interval of the Beta-coefficient should exclude 0 for the associated p-value to be statistically significant) [5, 6].

Example 9.4
Twenty patients with pancreatic adenocarcinoma were followed up for 1 year to determine the effect of clinical stage on survival. The results are shown in Table 9.7. A Cox-proportional model was fitted. Interpret the results.

Here is the Cox-regression model:

$$\mathrm{LogHR} = 4.194 \times \mathrm{Stage} \tag{9.20}$$

The p-value for the Beta-coefficient is 0.008. This means that stage is a significant contributor to the hazard ratio. In fact, we can calculate the relative risk (hazard ratio) of stage:

$$\mathrm{HR}_{\mathrm{Stage}} = e^{4.194} = 66.287 \tag{9.21}$$

This means that, at any time, if a patient has a tumor which is one clinical stage greater than another patient, then that patient is 66.287 times more likely to die compared to the patient with lower stage.

Table 9.7 Table of results for Example 9.4

Patient number	Follow-up time	Event indicator	Cancer stage
1	12	Censored	I
2	10	Censored	I
3	8	Censored	I
4	7	Censored	II
5	10	Censored	I
6	12	Censored	I
7	6	Censored	II
8	2	Dead	IV
9	4	Dead	III
10	6	Dead	II
11	4	Dead	III
12	12	Dead	I
13	8	Dead	II
14	10	Dead	I
15	3	Dead	IV
16	5	Dead	III
17	9	Dead	II
18	4	Dead	III
19	1	Dead	IV
20	1	Dead	IV

Summary

In this chapter, we reviewed survival statistics; these concepts form the fundamental concepts underlying many decision-making tools and algorithms in pathology. Many of the features that a surgical pathologist is required to report are because they have been shown to have a significant effect on the survival of the patients. Understanding of these concepts is required for most pathologists to interpret and analyze the literature regarding prognostication of survival of patients with specific diseases. Furthermore, some pathologists may actively participate in the design and conduct of prognostic studies which necessitate an understanding of survival statistics. The statistical tests introduced in this chapter are the main tests used in survival analysis and include Kaplan-Meier estimator, Log-rank test, and Cox-proportional hazards model.

References

1. Kleinbaum DG, Klein M. Survival analysis: a self-learning text. Springer Science & Business Media: USA; 2006.
2. Singh R, Mukhopadhyay K. Survival analysis in clinical trials: Basics and must know areas. Perspect Clin Res. 2011;2(4):145.

3. Lindsey JC, Ryan LM. Methods for interval-censored data. Stat Med. 1998;17(2):219–38.
4. Cox C, Chu H, Schneider MF, Muñoz A. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. Stat Med. 2007;26(23):4352–74.
5. Rodrıguez G. Parametric survival models. Technical report. Princeton: Princeton University; 2010.
6. Kestenbaum B. Epidemiology and biostatistics: an introduction to clinical research. Springer Science & Business Media: USA; 2009.

Validation of New Tests

10

Introduction

Pathology and laboratory medicine is an active and dynamic field where new tests and modalities are discovered and invented on a regular basis. The field of diagnostic medicine has a high rate of innovation and the laboratories need to adapt and implement new tests to address clinical needs and remain relevant.

Adding new tests usually starts with needs assessment; the process of needs assessment should be a continuous periodical review of the clinical needs with participation of the lab director, supervisors, and clinicians where needs are assessed based on the input from the stakeholders and new evidence obtained from literature review [1].

The next step in this process is to review all the tests (or instruments) that can address the needs highlighted in the need assessment process. This review will involve critical appraisal of the manufacture's claims and supporting evidence for the test. Technical and clinical specifications and parameters of the tests should be reviewed and compared with the needs highlighted.

The next step is to define the performance standards of the test and determine if they are compatible with the lab's resources. Some of these parameters and performance standards are related to financial cost which requires cost and cost-effectiveness analysis. However, some of the performance standards relate to the physical space, labor, and infrastructure needed for the test. Other concerns that should be considered at this stage include preanalytic considerations and turnaround time.

After a decision is made for adding a new test, the next step is to validate the test before full implementation. Validation ensures that the test performs as expected and satisfies the requirements of the lab, manufacturer, and regulatory bodies. The process of adding a new test is summarized in Fig. 10.1 [2–4].

In this chapter, we will focus on the process of test validation.

A. Momeni et al., *Introduction to Statistical Methods in Pathology*,
DOI 10.1007/978-3-319-60543-2_10

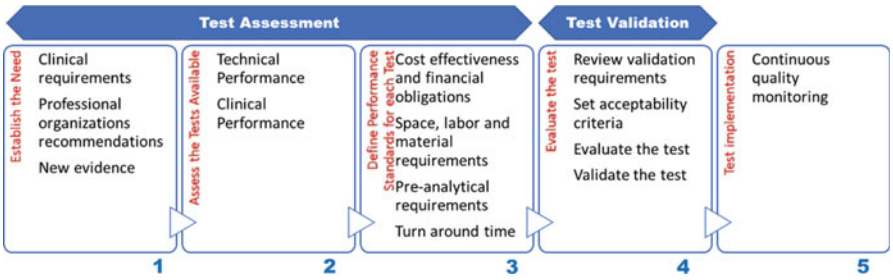


Fig. 10.1 Process of adding new tests

Test Validation

Regulatory bodies such as the Clinical Laboratory Improvement Amendments (CLIA) and College of American Pathologists (CAP) require laboratories to validate each new test that they add to their repertoire. Some states regulatory bodies (e.g., New York state) may have additional regulatory requirements before a test can be added to a laboratory's list of approved tests. Many of the new tests will already be technically approved by federal or state oversight bodies like the Food and Drug Administration; however, their adoption at a laboratory requires validation.

We introduced the concept of test assessment in Chap. 2. Every test needs to undergo assessment, this assessment will answer questions such as clinical usefulness, precision, accuracy, and cost-effectiveness. If after test assessment, a decision is made to add the test to the laboratory's test menu, then the test should be validated before the laboratory can bill for the test and report the results to the patients. Most regulatory bodies have a set of common or similar guidelines for test validation which we will describe in this chapter. It must be noted that sometimes instead of test assessment, a new test is added because of the recommendations of professional organizations (such as CAP), even in these situations, the added test needs validation before full adaptation.

The term "method evaluation" is used to describe the process of validating a new test. This process requires three elements: test assessment, validation, and verification. These elements are not necessarily sequential: part of the process of validation requires technical test assessment and verification is satisfied through documentation of the validation process.

- Test assessment refers to determination of analytical and clinical performance characteristics and was discussed extensively in Chap. 2. Here we will provide a brief review of test assessment.
- Validation requires that the test is shown, through objective measures, to fulfill the requirements of its specific intended use. Not only the tests need to satisfy

these requirements, but they need to do this consistently. Validation is required for all laboratory developed tests and modified FDA-approved tests.
– Verification is the objective evidence that the validation requirements are satisfied. Verification is required for all FDA-approved tests.

In the United States, tests generally fall into three regulatory categories: non-FDA-approved tests, FDA-approved non-waived tests, and FDA-approved waived tests. The first two categories require that the laboratory fulfills all the three elements mentioned above.

The method evaluation needs three steps: define performance goals, assess error, and finally compare the error with the goals. Here, we will walk you through each of these steps [5, 6].

Defining Analytical Goals

The purpose of a test is to diagnose a condition or assess an analyte with acceptable accuracy and precision. The term "acceptable" is used because in measurement there will always be a degree of inaccuracy and imprecision or in other words there will always be a degree of error. In defining analytical goals, we must determine what degree of error is acceptable to us, i.e., what level of precision and accuracy do we require. Setting the goals can be alternatively stated as setting the "acceptability criteria."

Laboratories can set their own requirements and goals; however, they must at least ensure that the level of error for a test is compatible with patient care (i.e., the error is small enough to have minimal impact on clinical decision making), is consistent with manufacturer's specifications, and is within the allowable error set by regulatory bodies (usually CLIA).

The acceptability criteria should state clear goals for levels of accuracy and precision. As we will discuss later, accuracy and precision goals are defined differently for qualitative and quantitative tests. For quantitative tests, the criteria should also include reportable range and reference intervals. These latter parameters are usually set based on the manufacturer's guidelines and are later verified by data collected from the population served by the laboratory.

Acceptability Criteria for Qualitative Tests

Qualitative tests usually return a binary response (e.g., detected vs. not detected or positive vs. negative), and rarely have more than two categorical responses. In laboratory medicine, semi-quantitative tests are also considered as qualitative tests for validation purposes. These semi-quantitative tests measure a quantitative value but report a categorical result based on set cutoffs (e.g., when levels of HBS antigen are measured but the result is reported as positive or negative based on a set cutoff).

Table 10.1 2×2 contingency table for qualitative tests

	Condition positive	Condition negative
Test positive	True positive (TP)	False positive (FP)
Test negative	False negative (FN)	True negative (TN)

The first goal in the criteria is accuracy which is concerned with how true is the result of the test when compared to the condition status of the test subjects. Accuracy is related to systematic error (bias). Accuracy for qualitative tests can be defined by going back to the 2×2 contingency table (see Table 10.1); accuracy is the ratio of all true results to all results:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{10.1}$$

The second goal in acceptability criteria of qualitative tests is precision which is also known as reproducibility. Simply stated, precision means that if the test is repeated multiple times the results should remain constant. Precision is related to random error. Precision has three components: within-run precision, between-run precision, and total precision. It is usually required that precision goals are set for all three components.

Reportable range is the range of values that the test (instrument) can detect directly before concentration or dilution. Reportable range for qualitative criteria depends on method criteria for a positive result. If the test is semi-quantitative and requires a cutoff for calling a positive result, then the cutoff and validation of cutoff become validation goals (as required by CAP). However, this depends on whether there is an actual quantitative output from the instrument or not.

Analytical sensitivity (detection limit) and specificity are other goals for qualitative test validation. While CLIA does not require laboratories to validate and verify sensitivity and specificity, CAP requires sensitivity validation for all tests and specificity validation for some tests (e.g., modified FDA-approved tests and laboratory developed tests).

Analytical sensitivity and specificity are different from diagnostic sensitivity and specificity; analytical sensitivity refers to the lowest concentration of analyte which the test can reliably detect as positive, while diagnostic sensitivity is the proportion of the test subjects with the target condition whose test result is positive (TP/TP + FN).

Analytical specificity refers to the ability of the test to only detect the analyte for which it was designed. Diagnostic specificity on the other hand is the proportion of individuals without the target condition who test negative (TN/TN + FP).

For non-FDA-approved tests, laboratories are required to establish the diagnostic specificity and sensitivity, while for FDA-approved tests, the requirement is only for analytical sensitivity and specificity as the process of FDA approval has already established the diagnostic sensitivity and specificity.

The final goal of the acceptability criteria for qualitative tests is "interference." The lab must either through interference studies or by review of literature

(or manufacturer's guidelines) determine the common interfering substances that can lead to inaccurate test results.

Acceptability Criteria for Quantitative Tests

Quantitative tests usually return a value within a range. The acceptability criteria for quantitative tests includes accuracy, precision, reportable range, and reference interval.

Accuracy for quantitative tests measures how close the measured analyte is to the true value of that analyte. Establishing accuracy requires method comparison experiments where either the test results are compared with a gold standard method or the test performance is measured for calibrators or reference material with known values of the analyte.

The accuracy for quantitative tests is achieved if linear correlation can be established between the test results and the true values of the analyte.

Precision for quantitative tests has two components: "repeatability" and "reproducibility." Repeatability is the degree of within-run agreement, while reproducibility is the degree of between-run agreement. Precision is established using F-test (analysis of variance), where the ratio of the measured variance to the expected variance should not be statistically significant (explained in detail later). For FDA-approved tests, reproducibility (long-term precision) is more critical, while for non-FDA-approved tests, the sequence of establishing precision requires testing for repeatability (short-term precision) followed by reproducibility.

Reportable range, as with qualitative tests, is the range of values that the test (instrument) can detect directly before concentration or dilution. For FDA-approved tests, the lab needs to verify that analyte levels at near critical levels of the manufacturer's reportable range can be measured and reported correctly (i.e., measuring sample with near low end and near high end values). For non-FDA-approved tests, establishment of reportable ranges requires a linearity experiment with serial dilutions (explained later).

Reference range or normal values are the expected values of the test on normal (non-affected) individuals. The reference range should either be established for the laboratory's target population or verified in cases where the target population is like the manufacturer's (or literature) sample population.

For non-FDA-approved test, analytic specificity and sensitivity should also be established.

In the next section, we will explain the experiments needed for validation or verification of a new test.

Validation Experiments

After goals are defined, a series of experiments must be undertaken to establish if the new test satisfies the requirements set by the acceptability criteria. These experiments must be well planned, documented, and reported. These experiments are not one-time experiments and should be repeated in predefined intervals as well

as whenever a critical component of the test (e.g., critical reagents, instruments, etc.) is changed. It is recommended that the laboratory have a written plan for performance of validation experiments.

Important considerations before validation include reagents, sample size, instrument, etc. It is important that experiment setup is exactly as the setup under which the test will be implemented, i.e., reagents and instruments should be the same with real patient samples (in some circumstances control samples are also acceptable).

One general concern for validation experiments is when running side-by-side experiments (e.g., method comparison experiment using a gold standard), the tests should be run on the same sample within a short timeframe or otherwise steps should be taken (e.g., refrigeration) to ensure that the sample quality does not change between the two tests.

We will explain the following validation experiments: accuracy and precision experiments for qualitative tests, method comparison experiments, F-test for precision, linearity experiments, total allowable error, and detection limit experiments.

The first step for validation experiments, however, is to determine the sample size needed [7–9].

Sample Size Calculations

For calculation of sample size, you have the choice of either following the established guidelines set by regulatory bodies or to calculate the sample size using sample size formulas. Table 10.2 shows the commonly agreed upon minimum sample sizes for different validation experiments.

Sample size calculations for quantitative tests. The following general sample formula can be used:

$$n = \frac{\left(Z_\alpha + Z_\beta\right)^2 S^2}{\Delta^2} \tag{10.2}$$

where n is the sample size; Z_α is the Z-score for the rate of acceptable type I error (false negative rate) which is usually set for a type I error rate of 0.05 (with corresponding two-sided Z-score of 1.96); Z_β is the corresponding Z-score for the rate of acceptable type II error (false positive rate) which is usually set at 0.01, 0.05, or 0.10 (with corresponding Z-scores of 2.326, 1.645, and 1.282 respectively); S^2 is the variance of data (determined using population data, or manufacturer's data); and Δ is the minimum clinically significant difference in the test results (e.g., for potassium concentration differences of 0.1 mEq/L might be considered clinically significant, but for sodium the 0.5 mEq/L might be considered clinically significant).

Another formula that can be used for both quantitative and qualitative tests is based on set levels of confidence and reliability. Confidence (accuracy) is the difference between 1 and type I error rate. Reliability is the degree of precision. For this formula, the failure rate must be decided as well, i.e., how many incorrect

Table 10.2 Sample size guidelines for test validation

Validation experiment	FDA-approved test	Laboratory developed test or modified FDA-approved test
Qualitative tests		
Accuracy experiments	40 specimens (CLIA requirement: at least 20)	40 specimens (CLIA requirement: at least 20)
Precision experiments	Minimum of 2 negative samples and 2 positive samples	Minimum of 2 negative samples and 2 positive samples
Reportable range	At least 3–5 low and 3–5 high positive samples	At least 3–5 low and 3–5 high positive samples
Reference range	At least 20 known normal samples	At least 120 reference samples
Quantitative tests		
Method comparison experiments	At least 20–40 samples	At least 40 samples
F-test for precision	At least 2–3 samples near each clinically important level	At least 2–3 samples near each clinically important level
Reportable range	At least 4–5 samples (low end, mid point, high end)	At least 5 dilution levels
Reference range	At least 20 known normal samples	At least 120 reference samples

results are we allowing for our validation process. For a failure rate of 0, the equation can be stated as:

$$n = \frac{\ln(1 - \text{confidence})}{\ln(\text{reliability})} \tag{10.3}$$

Usually the confidence level is set at 0.95 and reliability at 0.90 or 0.80 with zero failure rate which translates to a sample size of 29 and 14, respectively.

For failure rates other than 0, the results follow a binomial distribution (see Chap. 3). The calculation of the sample size is based on the following equation:

$$1 - \text{Confidence} = \sum_{i=1}^{f} \binom{n}{i} (1 - \text{Reliability})^{i} \text{Reliability}^{n-i} \tag{10.4}$$

where f is the failure rate and n is the sample size. Many statistical software programs have tools that will calculate the sample size using the above equation with given confidence and reliability levels.

It is important to note that the sample size does not necessarily mean the number of subjects or specimens. For example, in precision experiments, if a sample size is calculated using the above equation, the number signifies the number of experiments needed rather than the number of specimens.

Table 10.3 Results of validation experiment for Example 10.1

	Day 1	Day 2	Day 3	Day 4	Day 5	Total
True positive	10	9	9	10	10	48
False positive	0	2	1	2	0	5
True negative	10	8	9	8	10	45
False negative	0	1	1	0	0	2

Accuracy Experiment for Qualitative Tests

Accuracy experiment for qualitative tests is a method comparison method where the test is compared with either the gold standard or the test is used on known samples. The test is compared on 20 (or more) samples for five consecutive days and the accuracy is calculated using Eq. (10.1). If the accuracy levels obtained meets the acceptability criteria, the experiment is ended. However, if discrepancies are found, the experiment is extended for another 5 days.

Example 10.1
Q: A test is validated using 10 known positive samples and 10 known negative samples. The following results are obtained over the course of 5 days (Table 10.3). The acceptability criteria call for an accuracy of 95%. Determine if the test can be validated with the results obtained.

A: Using Eq. (10.1), we will obtain the following results:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \frac{48 + 45}{48 + 45 + 5 + 2} = \frac{93}{100} = 0.93 \qquad (10.5)$$

The validation results show that the test has failed to fulfill the accuracy criterion. In this situation, an extension of the experiment for another 5 days is warranted. If after the extension, the test still has not fulfilled accuracy criterion, then the validation has failed.

Precision Experiment for Qualitative Tests

For precision experiments, a minimum of two positive and two negative samples should be tested in triplicates for five consecutive days. Agreement of test results is measured as within-run agreement and between-run agreement. Total precision is also reported which is calculated from the sum of all between-run agreement results.

Example 10.2
Q: The precision of a new test with positive and negative results is being tested. Two known positive and negative samples are used for the testing. The samples are run in triplicates for five consecutive days with the Table 10.4 showing the results. Calculate the within-run, between-run agreements, and total precision.

Table 10.4 Results of validation experiment for Example 10.2

Sample	Day 1			Day 2			Day 3			Day 4			Day 5		
Positive	+	+	−	+	+	+	+	+	+	+	−	+	+	+	+
Positive	+	+	+	+	+	+	+	−	+	+	+	+	+	+	+
Negative	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−
Negative	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−

Table 10.5 Within-run agreement for Example 10.2

	Day 1	Day 2	Day 3	Day 4	Day 5
Within-run agreement	11/ 12 = 92%	12/ 12 = 100%	10/ 12 = 83%	11/ 12 = 92%	11/ 12 = 92%

A: The within-run agreement is the ratio of number of results in agreement within a run to the total number of results within that run. Table 10.5 shows the within-run agreement for different days of the validation experiment.

The between-run agreement is the ratio of number of results in agreement for a sample to the total number of results for that sample. Thus, the between-run agreements for samples 1–4 are 13/15 (86%), 14/15 (93%), 14/15 (93%), and 14/15 (93%), respectively. The total precision based on this experiment is 55/60 (91%).

For quantitative tests the precision is determined by a similar method: a replication study is performed where an analyte concentration is measured multiple times. However, the amount of pure error is now quantified using standard deviation.

Method Comparison Experiments for Quantitative Tests

The general approach for method comparison for quantitative tests involves either running 20–40 samples with known values or concurrently running 20–40 samples on the validated test and a gold standard test. The method comparison is recommended to be run over a period of at least five consecutive days. The results are then compared by running a linear correlation test. These results can also be inspected visually using a comparison plot which is essentially a linear correlation plot (see Chap. 4). The comparison plot can show linearity, outliers, and the range of results (Fig. 10.2a).

If the results have high linearity with a one-to-one agreement, then an alternative approach is to use a "difference plot" which shows the difference of the test versus the comparison on the Y-axis and the results of comparative method on the X-axis. The difference points should scatter around 0 on the Y-axis (Fig. 10.2b).

Any significant differences or outliers should prompt an investigation of the cause. Usually the first step is to repeat the measurement for that sample.

The linear correlation allows us to check for systematic error. Linear correlation will return a regression equation which will include the intercept (or constant (C))

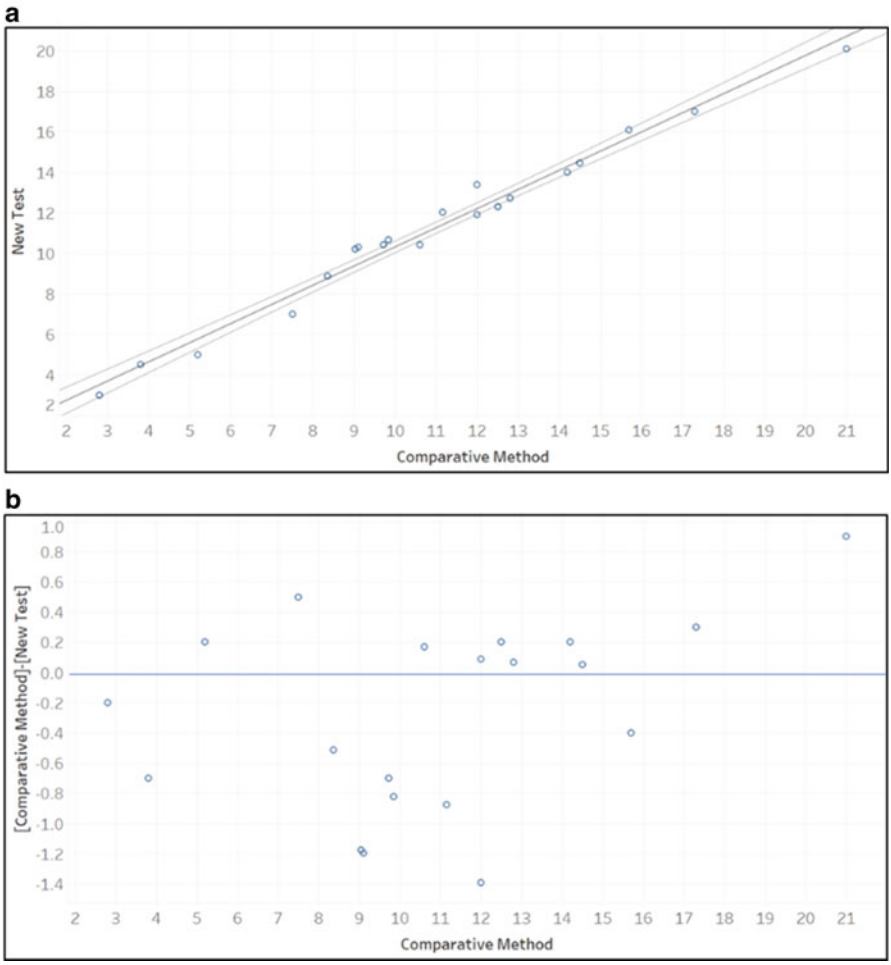


Fig. 10.2 Comparison plot of a new test versus the comparative method (**a**) with a fitted line. The difference plot is shown in the right panel (**b**)

and slope (B). The linear regression equation is usually calculated using the "best fit line" method (see Chap. 4). The best fit line approach finds the slope and intercept of a line where the sum of squared differences of the points from the line is minimum:

$$\min Q(C,B) \quad \text{for } Q(C,B) = \sum_{i=1}^{n} (y_i - C - Bx_i)^2 \tag{10.6}$$

The regression equation can be stated as follows:

$$\text{Comparative method result} = (B \times \text{New test result}) + \text{Constant} \tag{10.7}$$

The constant is indicative of constant systematic error, i.e., if the slope is 1, then the new test is different from comparative method result by a constant amount throughout its range of results. The slope in the equation shows proportional systematic error, i.e., the difference between the comparative test value and the new test value differs throughout the range.

Running a linear correlation on the results will also provide you with the standard deviation of the points around the fitted line, the confidence interval of the slope, the "correlation coefficient" (also known as "Pearson's r coefficient"), and a p-value.

A significant p-value is needed to say that there is linear correlation. In other words, a significant p-value is the first thing you should look at in the method comparison experiment which will tell you if the new test is useful for measuring the target analyte. Thus, a significant p-value is what you need for validation.

The next step is to look at the correlation coefficient to determine if there is any systematic error. Pearson's r coefficient shows how well the compared results change together and can have values of between -1 and 1.

Pearson's r statistics can be stated as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{10.8}$$

where n is the size of the sample, x is the test variable, and y is the comparison variable.

A perfect linear relationship between the two results will result in a Pearson's r coefficient of 1. A perfect correlation coefficient ($r \geq 0.99$) means that there is minimum systematic error. However, if $r < 0.975$, then a systematic error exists; in such situations, you need to run a t-test and a F-test to determine the source of the bias.

Of note, the correlation coefficient tends to be affected by random error as well. If the range of the analyte measures is narrow, the effects of random error on the correlation coefficient will be larger (thus, Pearson's r coefficient tends to be smaller for such analytes (e.g., electrolytes)). For analytes with a wider range of values, the effects of random error will be smaller and thus they tend to have higher correlation coefficients.

Example 10.3

A method comparison experiment is run comparing a new test with a gold standard test. The results of the experiment are provided in Table 10.6.

Table 10.6 Results of the method comparison experiment for Example 10.3

New test	Method comparison
14.45017	14.5
13.39091	12
12.73093	12.8
12.03	11.15634
12.3	12.5
10.43	10.59877
16.1	15.7
11.91149	12
5	5.2
7	7.5
3	2.8
20.1	21
10.66921	9.847511
14	14.2
10.42656	9.726387
17	17.3
10.3011	9.10431
10.21318	9.03637
4.5	3.8
8.87772	8.364023

The linear regression equation for the above results is:

$$\text{Comparative Method} = -0.676 + 1.04\ \text{New Test} \tag{10.9}$$

The calculated p-value for the regression equation is <0.001 which means that the test is highly correlated with the comparison method. The Pearson's r coefficient is 0.99 showing that there is no significant systematic error.

T-Test for Method Comparison Experiments

The t-test (see Chap. 6) should be run to determine if the mean of the two sets of results is the same or in other words t-test can determine if there is any systematic error (bias) in the mean of the two sets of values. In most method comparison experiments, it is best to use a paired t-test, since the same sample is being compared using two different methods (and thus a degree of similarity of the means expected and running an unpaired t-test will fail to detect the bias). If the t-test returns a nonsignificant value, then there is no systematic error.

If the t-test returns a significant p-value, then it shows that there is a significant bias (systematic error) in the mean of the two sets of values. If the t-test is significant, then you should go back to the linear regression equation to determine whether the source of the bias is the constant or the slope. Constant error is easily remedied by adding the constant to the new test results. However, for proportional errors, "recovery experiments" are needed.

Recovery experiments allow us to estimate the proportional systematic error. In these experiments a patient's specimen is divided into two equal aliquots and the concentration of the target analyte is measured in both. Then, a standard solution of the target analyte with known concentration is added to one aliquot (aliquot A), and an equal amount of diluent (e.g., water) is added to the other aliquot (aliquot B). The target analyte is measured again in the two aliquots. The expectation is that the difference between the two aliquots be the same as the amount of target analyte added to one of the tubes. The recovery percent is calculated as:

$$\text{Recovery}\% = \frac{(\text{Analyte amount in aliquot A}) - (\text{Analyte amount in aliquout B})}{\text{Amount of analyte added to aliquote A}} \times 100 \tag{10.10}$$

Note that the terms in the formula are "amount" rather than "concentration." However, most instruments and tests measure "concentration"; thus, before recovery percent is calculated, you need to calculate the amount of analyte using the measured concentrations and sample volumes.

These experiments should be run at least in duplicates. The sample size is dependent on the type of systematic error suspected and may vary from a few to 20 patients.

The difference of recovery percent from 100 is the proportional error percent. This measure should be smaller than the total allowable error set in the acceptability criteria of the lab and the CLIA requirements.

Example 10.4

A method comparison experiment is run comparing a new test with a gold standard test. The results of the experiment are provided in Table 10.7.

The linear correlation returned a Pearson's *r* coefficient of 0.971. A paired t-test was performed which returned a t-score of −2.827 with 19 degrees of freedom which translates to a significant p-value of 0.011. The results indicate that some form of systematic error (bias) exists. A recovery experiment was performed.

A sample with 10 mg/L of the test analyte was split into two aliquots of 10 ml (in tubes A and B). To tube A, 10 ml solution with a concentration of 100 mg/L of the target analyte was added. To tube B, 10 ml of distilled water was added. The concentration of the target analyte was measured again with results for tubes A and B being 40 mg/L and 5 mg/L, respectively. Calculate the proportional error and determine if the proportional error is less than the total allowable error of 1 mg/L at the middle of the range (10 mg/L is the middle of the range for this analyte).

First, let us calculate the amount of target analyte in the aliquots before the other solutions were added:

Each tube had 10 ml of a 10 mg/L solution. This can be written as $(10 \times {}^{10}/_{1000})$; thus, the amount of analyte in each aliquot is 0.1 mg. In the same manner, the 10 ml of the 100 mg/L solution has 1 mg of target analyte. The tube A sample after adding

Table 10.7 Results for Example 10.4

New test (mg/L)	Method comparison (mg/L)
14.45	14.5
13.39	15
12.73	14
12.03	13
12.3	12.5
10.43	11
16.1	15.7
11.91	12
5	7
7	8
3	2.8
20.1	23
10.67	9.85
14	14.2
10.43	13
17	17.3
10.3	9.1
10.21	11
4.5	6
8.88	9

the solution has 20 ml of 40 mg/L solution which translates to 0.8 mg of target analyte. The tube B sample after adding water has 20 ml of 5 mg/L solution which means that tube B has 0.1 mg of the target analyte (which equals the amount before the experiment since only water was added).

Now we can calculate the recovery percent:

$$\text{Recovery}\% = \frac{0.8 - 0.1}{1} \times 100 = 70\% \tag{10.11}$$

So, the proportional error percent is 30%. The total allowable error at 10 mg/L is 1 mg/L for this analyte which translates to a 10% allowable error; however, the proportional error percent is 30% which is much larger than the allowable error. This means that we have failed to validate the test as we have exceeded the allowable error.

F-Test for Precision

"F-test" or analysis of variance (see Chap. 6) compares the variance of the test method with the comparative method. In simple terms, F-test shows whether the variation observed in the test values is different from the variations observed for the comparative value. If no random error exists, you would expect the variations of the two sets of results to be similar, i.e., any variation observed in the test result is

caused by actual variation of the sample value rather than due to error. F-test for two variances is a simpler form of the ANOVA equation introduced in Chap. 6 and can be stated as:

$$F = \frac{S_1^2}{S_2^2} \quad \text{where } S_1^2 > S_2^2 \tag{10.12}$$

with S^2 being the variance of the values. The critical values can be looked up in a F-table (Appendix D) with (N−1,N−1) degrees of freedom where N is the sample size.

If the p-value shows no significance, then we can state that the random error in the test is not more than the random error of the comparison method, and conversely, a significant p-value signifies the existence of significant random error in addition to the random error of the comparison method.

An important measure here is the standard deviation of the test values which is the square root of the variance. The standard deviation (SD) is used as an indicator of random (pure) error in calculation of the total allowable error.

Example 10.5
Going back to Table 10.7, let us calculate the F-statistic for the method comparison experiment. The variance of the new test results is 17.982 and the variance of the comparative method results is 19.714. Now, we can calculate the F statistics:

$$F = \frac{S_1^2}{S_2^2} = \frac{19.714}{17.982} = 1.096 \tag{10.13}$$

Looking at the F-table for (19,19) degrees of freedom for the right tail of a two-tailed significance level of 0.05, we can see that the critical value is approximately 2.16. Since the calculated F-statistic is less than the critical, then we can say that the result is insignificant and no additional random error is present.

Linearity Experiments for Reportable Range

As part of the validation process, either the reportable range of a test should be established or the reportable range claimed by the manufacturer should be verified. This is done through "linearity experiments" also known as "analytical measurement range experiments." For linearity experiments, at least 5 samples with known concentrations of the target analyte are needed. These samples should have at least one sample near the low end of detection (detection limit), one sample near the high end of detection, one sample near the concentration of clinical interest (e.g., population average for electrolytes), and one or more samples with concentrations between the high and low end. Ideally, the sample concentrations of the target analyte are equally spaced. The measurements of these samples should preferably be made in triplicates or more.

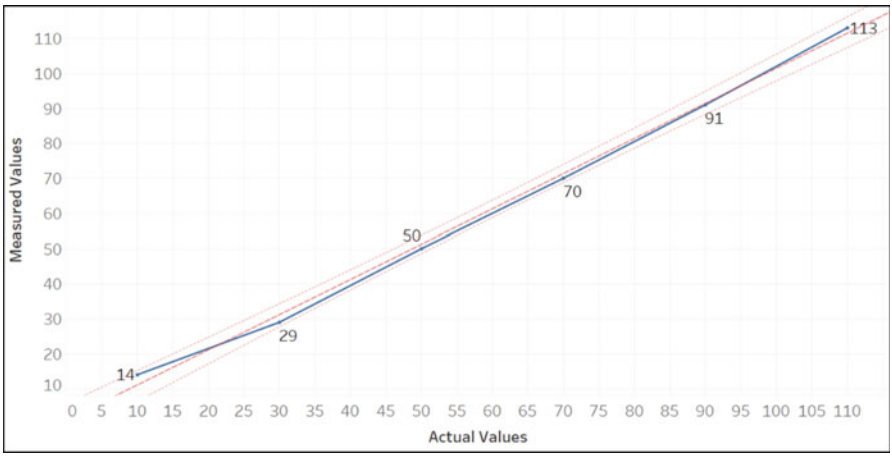


Fig. 10.3 Connected points line and the best fit line of a linearity experiment. The best fit line (*red dotted line*) has the most overlap with connected points line in the 30–90 range with deviations more visible at 10 and 110

After repeated measurements, the mean measured value for each of the samples should be plotted along with the actual concentrations on a scatterplot with Y-axis being the measured values scale and the X-axis being the actual concentration scale. Each consecutive pair of points is then connected with a line and a best fit line is then drawn for the points in the scatterplot. This allows for visual inspection of the connected points line with the best fit line: ideally, the best fit line and the connected points line should be the same, and the reportable range of the test would be areas where the fitted line and the connected line have the most overlap (Fig. 10.3).

Visual inspection, however, has limitations and ideally a least squares linear regression line should be calculated for the data (as in method comparison, see Chap. 4). After fitting the regression line, the "lack-of-fit error" should be calculated; this is essentially the sum of squared difference between the measured value for each actual value with the value predicted by the regression line for that point.

Since we have performed the experiment in triplicates, for each measurement point (which is the mean of triplicate measurements), we can calculate the sum of random error for the points which is the sum of squares of all the deviations of measured points from their average.

The sum of lack-of-fit error and random error is the total error of the linearity experiment.

To determine linearity, a lack-of-fit F-test (G test) is run; this test is essentially a variation of the F-test in which the ratio of the lack-of-fit error to random error is calculated:

$$G = \frac{\text{Sum of Squares of Lack-of-fit error}}{\text{Sum of Squares of Random error}} \tag{10.14}$$

The G is then multiplied by the ratio of degrees of freedom of the pure error and lack-of-fit error. The degrees of freedom of pure error equals to total number of measurements minus the number of actual values. The degrees of freedom of lack-of-fit error equals the total number of actual values minus 2:

$$\text{DF ratio} = \frac{\text{DF of Random error}}{\text{Df of Lack-of-fit error}} = \frac{n-c}{c-2} \tag{10.15}$$

where n is the total number of measurements and c is the number of actual values.

The F-score can be constructed as:

$$F = G \times \text{DF ratio} \tag{10.16}$$

The critical values for the F-test can be looked up in a F-table with $(n-c, c-2)$ degrees of freedom. If the p-value is not significant, then there is linearity. Conversely, a significant p-value rejects linearity. In these situations, if the test has been performed for verification, then we have failed to verify the reportable range of the manufacturer.

For laboratory developed tests, there are two solutions: either new samples should be added (for instances, where few samples were tested) which are further from the low end and high end. Or in cases where sufficient points were measured, the most extreme pair of points is removed and the F-test is calculated again (this process can be continued while there are more than five points remaining until a reportable range can be calculated for the test).

A big shortcoming of the least squares method is that outlier exert considerable influence on the fitted line and consequently it is best to visually inspect the data before the regression line is fitted and address the significant outliers (usually by repeating the experiment for that sample). F-test is also highly affected by precision (random error) and can sometimes assume linearity when the points form a nonlinear correlation.

For these reasons, CAP has proposed using the polynomial method. In polynomial approach, the first step is to check for nonlinearity (either quadratic or cubic). If no nonlinear correlation is identified, then the measurement points are assumed to be linear and are called "Linear 1." If a significant nonlinearity is identified, then this nonlinearity is checked to determine if it is clinically significant (by testing the data against clinically relevant allowable error). If the nonlinearity is not clinically significant, then a value of "Linear 2" is returned, which essentially means that the data is treated as if it was linear. If there is nonlinearity that is also clinically significant, then the polynomial method returns a value of "nonlinear" which means that the validation has failed for the reportable range. Explanation of the calculations for the polynomial method is beyond the scope of this current book.

Example 10.6

A linearity experiment has been performed for a new test. The results are reported in Table 10.8. Determine if the reportable range of the test can be validated.

Table 10.8 Results for linearity experiment in Example 10.6

Actual values	Repeated measure 1	Repeated measure 2	Repeated measure 3	Average for the point	Fitted line value	Sum of squares of pure error	Sum of squares of lack-of-fit error
10	12	11.75	13	12.25	10.46	0.875	3.200
30	28.7	29.5	30.7	29.63	30.60	2.026	0.943
50	50	49	51	50.00	50.74	2	0.559
70	70	70	70.1	70.03	70.89	0.006	0.735
90	90.2	90.3	89.3	89.93	91.03	0.606	1.212
110	113	112.2	114	113.07	111.17	1.626	3.567
Total						7.141	10.218

To calculate the G value, we need to divide the sum of lack-of-fit errors of the points by the sum of pure (random) error:

$$G = \frac{\text{Sum of Squares of Lack-of-fit error}}{\text{Sum of Squares of Random error}} = \frac{10.218}{7.14} = 1.43 \tag{10.17}$$

We have 18 total measurement and 6 actual values; thus, we can calculate the DF ratio:

$$\text{DF ratio} = \frac{n-c}{c-2} = \frac{12}{4} = 3 \tag{10.18}$$

Consequently, the F statistic will be:

$$F = G \times \text{DF ratio} = 1.43 \times 3 = 4.29 \tag{10.19}$$

The critical value of F for (12,4) degrees of freedom is 5.91. Because $4.29 < 5.91$, we have failed to reject the null hypothesis, thus showing that test is linear throughout its reportable range.

Allowable Total Error

"Total error" or "total analytical error" (TAE) is the sum of systematic error and random error. It has been shown that total error is a more accurate measure of diagnostic error than bias (systematic error) alone. The systematic error will be calculated from a method comparison study and the random error is calculated from a replication study (which can be a part of the method comparison study). Total analytical error is then defined as:

$$\begin{aligned} &\text{TAE} = \text{Bias} + 2\text{SD} \text{ for two-tailed estimates or} \\ &\text{TAE} = \text{Bias} + 1.65\text{SD for one-tailed estimates} \end{aligned} \tag{10.20}$$

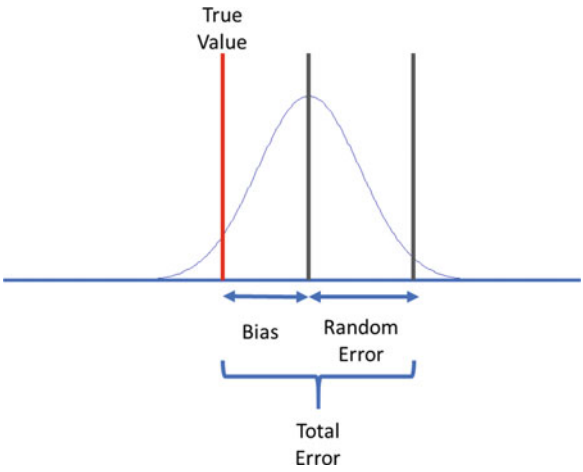


Fig. 10.4 Total error is a combination of bias and random error

Simply stated, the measured value can be different from the true value not only by the amount of systematic error but also by random error (random error can alleviate or aggravate the bias) (Fig. 10.4).

For verification purposes the total error should be assessed for 20 patients; however, for laboratory developed tests, the current requirement is to check at least 120 patients (or samples) or preferably 120 patients for each decision-level concentration.

"Allowable total error" (ATE) is the amount of total error that is acceptable for the intended use of the test. CLIA and other regulatory bodies have set requirements for ATE of different analytes, and these requirements are the baseline ATE that laboratories need to follow. However, they may choose to adopt more stringent ATE based on their clinical setting. This is called the clinical allowable error.

In Six Sigma concepts, there are four levels of ATE: bias +2SD, bias +4SD, bias +5SD, and bias +6SD. Each Sigma can be defined as ((Level of ATE – bias)/SD). The 6S or Six Sigma is the tolerance limit of the test (this translates to running two levels of controls per analytic run (see next chapter)).

After the laboratory sets an ATE for a test, the next step is to draw a "method decision chart." The method decision chart has the allowable bias percent on the Y-axis with a scale from 0 to ATE and the allowable random error (imprecision) percent (described as percentage of SD or CV) on the X-axis with a scale from 0 to 0.5 ATE. The next step is to draw Sigma lines, which are drawn by connecting the y-intercept at ATE and the x-intercept at (corresponding level ATE/SD (or CV)). For example, Fig. 10.5 shows a method decision chart for a test with ATE of 10% and SD of 1.

The next step would be to chart the operating point (total error) of the test on the chart (using the measured bias and imprecision). Depending on where on the chart the test is performing, you can make decisions on the quality of the test. The regions on the chart correspond to different levels of performance; from right to left

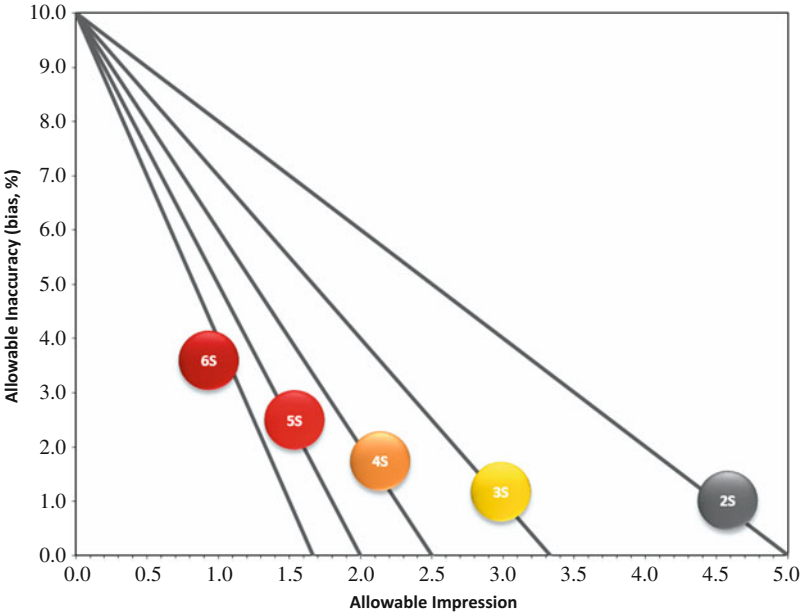


Fig. 10.5 Method decision chart for a test with ATE of 10% and SD of 1. Lines represent different levels of Sigma

these performance levels are unacceptable, poor, marginal, good, excellent, and world class, each corresponding to different levels of Sigma (2–6S and finally better than 6S).

The test should perform at a marginal performance level or better to be allowed for regular operational use. At marginal performance level, the implementation of the test requires 4–8 quality control samples per analytical run and stringent quality monitoring strategy. A test with good performance needs 2–4 quality control samples per analytical run whose measured values should fall within 2.5 standard deviations of their actual value. Excellent performance requires 2 quality control samples per run with an acceptability range of 2.5–3 standard deviations. Finally, world class performance only needs 1 or 2 quality control samples per run with an acceptability range of 3–3.5 standard deviations.

Detection Limit Experiments

"Detection limit experiments" are needed for determination of analytical sensitivity. These experiments are based on the results of the test on a blank and a spiked sample. The blank sample should have a zero concentration of the target analyte. The spiked sample should have a low concentration of the target analyte (corresponding to manufacturer's claim for detection limit); usually several spiked

samples are required with progressively higher concentrations of the target analyte. The basis of the experiment itself is a replication study where both the blank sample and the spiked sample are measured repeated times to establish a mean concentration as well as standard deviation. The recommended number of repetitions is 20 times for verification purposes and 60 times for validation purposes.

The first estimate determined is called the "limit of blank" (LoB) which is the highest concentration that is likely to be observed for a blank sample with a one-sided confidence interval of 95%. This corresponds to mean blank concentration plus 1.65 standard deviations of the blank measurement.

"Limit of detection" (LoD) is the next estimate of the experiment; this limit is lowest amount of target analyte that can be detected with a probability of 95%. This corresponds to limit of blank plus 1.65 standard deviations of the spiked measurement. While LoD shows the analytical sensitivity, it does not show how accurate are the measurements at that concentration. In fact, usually measurements at LoD are unreliable and should not be part of the reported operational range.

Thus, the next estimate needed which will set the operational limit of the test should be measured. This estimate is called "limit of quantification" (LoQ) and refers to the lowest concentration of the target analyte that can be detected with acceptable accuracy and precision. Limit of quantification needs multiple spiked samples to be measured; LoQ will be mean spiked concentration where the total error (bias plus 2 standard deviation) is less than the allowable total error set for that test by the laboratory or regulatory bodies.

A similar concept to LoQ is the "functional sensitivity" which is the lowest concentration at which the coefficient of variation (CV) is 20%. However, as CV is the ratio of standard deviation to mean, then functional sensitivity only represents the precision of the test.

For verification purposes, the laboratory needs to establish that the test meets the specifications of the test set by the manufacturer [8, 10–13].

Notes on Validation of Immunohistochemical Tests

In anatomic pathology, one of the critical tests that needs validation and verification is immunohistochemical staining. Immunohistochemical stains are often used to guide important diagnostic and/or prognostic decisions (e.g., HER2/neu status in breast carcinomas), as such issues of accuracy and precision are significant. Furthermore, sensitivity and specificity are also relevant issues that need to be established before a diagnostic decision is based on immunohistochemical stains. To establish sensitivity and specificity, a review and critical appraisal of literature is needed (see Chap. 12). Here, we will briefly discuss issues regarding validation of immunohistochemical (IHC) stains.

The steps in validation of IHC stains include assay optimization, establishment of analytical sensitivity and specificity, and concordance studies. Assay optimization is a critical step in IHC validation and aims to develop an IHC protocol that addresses the issues of antigen retrieval and detection. Based on the manufacturer's

guidelines, the assay needs to be optimized to ensure that the quality and pattern of IHC stain is comparable with the manufacturer's specification. While automation has greatly helped with assay optimization, there is still need for changes and tweaks to IHC protocols to ensure optimal IHC results.

Concordance studies are the crucial step in IHC assay validation. There are different approaches to concordance studies: the simplest form of concordance study is to compare the results with the expected results based on morphology or tissue origin. The *Human Protein Atlas* project (www.proteinatlas.org) is a comprehensive resource that can be used for choosing appropriate tissue controls for IHC validation.

Another method is to use a gold standard or previously validated test and test the same tissue using both the new test and the validated test. The results can then be compared. For IHC this occasionally means validation using non-IHC methods such as fluorescent in situ hybridization (FISH), flow cytometry, molecular studies, or even clinical outcomes. An alternative to this approach is to test the tissue in another laboratory that has already validated the new IHC test.

Ideally, IHC validation studies need to use pertinent positive and negative appropriate for the intended clinical use. Laboratories need to test at least ten positive and ten negative tissues for non-predictive IHC markers. In cases where the degree and pattern of positivity is part of the clinical decision making, the concordance study needs to include samples with different expression levels or patterns as well. For predictive IHC markers, the initial validation needs 20 samples for each staining pattern. As the sample size increases, the confidence interval for the overall concordance narrows; for 1 discordant result, the confidence interval of concordance for 10 and 20 samples is 57–100% and 75–100%, respectively.

The intended use of the IHC assay is important in the sample size calculations: IHC assays that will act as stand-alone clinical decision-making tools need more stringent validation with a larger sample size. However, when IHC is used as part of a panel, the requirements for validation are less stringent reducing the number of samples needed for validation.

It is suggested that each time the protocol changes for an IHC assay or one of the components of the test is changed, the laboratory needs to confirm assay performance using two samples for each staining pattern.

Current guidelines recommend an overall concordance (agreement) of 90% between the new test and the comparison method for validation of the new test. If overall concordance is less than 90%, a McNemar test should be run to establish whether the discordant results are statistically significant. If no statistical significance is found and the overall concordance is sufficiently high (e.g., $>80\%$), then we can still accept the validation study results.

Unfortunately, IHC validation has a subjective analytical component, i.e., interpretation of results is dependent on the pathologist. To minimize this, often it is recommended that the results are interpreted by more than one pathologist and consensus results are used in validation. However, major discrepancies and disagreement between the raters should be addressed.

Currently, there are no recommendation on measuring inter-observer and inter-method agreement (usually it is based on overall agreement with no statistical comparison). However, we purpose that an agreement statistical test is applied: Kappa's D for binary results (e.g., positive/negative), Spearman's rho for ordinal variables (e.g., scores 1–3 for HER2/neu staining), and Pearson's *r* for continuous variables (e.g., number of tumor infiltrating lymphocytes).

Recent advances in computerized image analysis have meant that currently there are many commercial and open-source software available that can allow for objective evaluation of the IHC assay. This can help in validation studies by reducing the human error bias [14].

Summary

Adding a new test requires a vigorous validation and verification process. Even if regulatory bodies approve a test, this does not mean that test will have a similar performance in the laboratory setting as the original validation setting. The process of validation/verification requires that goals (acceptability criteria) are set and experiments are run to test whether the test meets the goals that are set. These goals and experiments need to address issues such as accuracy, precision, sensitivity, and specificity. Many of these criteria should be periodically checked to ensure that the test performance is still at the level required by acceptability criteria. We will address some of the statistical concepts in pathology and laboratory medicine in the next chapter.

References

1. McPherson RA, Pincus MR. Henry's clinical diagnosis and management by laboratory methods. Elsevier Health Sciences: USA; 2016.
2. Van den Bruel A, Cleemput I, Aertgeerts B, Ramaekers D, Buntinx F. The evaluation of diagnostic tests: evidence on technical and diagnostic accuracy, impact on patient outcome and cost-effectiveness is needed. J Clin Epidemiol. 2007;60(11):1116–22.
3. Knottnerus JA, van Weel C, Muris JW. Evaluation of diagnostic procedures. Br Med J. 2002;324(7335):477.
4. Ball JR, Micheel CM, editors. Evaluation of biomarkers and surrogate endpoints in chronic disease. National Academies Press: USA; 2010.
5. Taverniers I, De Loose M, Van Bockstaele E. Trends in quality in the analytical laboratory. II. Analytical method validation and quality assurance. TrAC Trends Anal Chem. 2004;23 (8):535–52.
6. Dufour DR. Laboratory general checklist: how to validate a new test. College of American Pathologists USA; 2008.
7. Westgard JO, Lott JA. Precision and accuracy: concepts and assessment by method evaluation testing. CRC Crit Rev Clin Lab Sci. 1981;13(4):283–330.
8. Chan CC, Lee YC, Lam H, Zhang XM, editors. Analytical method validation and instrument performance verification. Hoboken: Wiley; 2004.
9. Mansfield E, O'Leary TJ, Gutman SI. Food and Drug Administration regulation of in vitro diagnostic devices. J Mol Diagn. 2005;7(1):2–7.

10. Hens K, Berth M, Armbruster D, Westgard S. Sigma metrics used to assess analytical quality of clinical chemistry assays: importance of the allowable total error (TEa) target. Clin Chem Lab Med. 2014;52(7):973–80.
11. Mattocks CJ, Morris MA, Matthijs G, Swinnen E, Corveleyn A, Dequeker E, Müller CR, Pratt V, Wallace A. A standardized framework for the validation and verification of clinical molecular genetic tests. Eur J Hum Genet. 2010;18(12):1276–88.
12. Aziz N, Zhao Q, Bry L, Driscoll DK, Funke B, Gibson JS, Grody WW, Hegde MR, Hoeltge GA, Leonard DG, Merker JD. College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. Arch Pathol Lab Med. 2014;139(4):481–93.
13. Jhang JS, Chang CC, Fink DJ, Kroll MH. Evaluation of linearity in the clinical laboratory. Arch Pathol Lab Med. 2004;128(1):44–8.
14. Patrick LF, Linda AB, Lisa AF, Alsabeh R, Regan SF, Jeffrey DG, Thomas SH, Karabakhtsian RG, Patti AL, Marolt MJ, Steven SS. Principles of analytic validation of immunohistochemical assays. Arch Pathol Lab Med. 2014;138(11):1432–43.

Statistical Concepts in Laboratory Quality Control

11

Introduction

In the previous chapter, we discussed test validation. However, once a test is validated, continuous monitoring of quality metrics is needed to ensure that the test performance is within the limits set by validation experiments as well as the requirements by the lab and regulatory bodies. This requires periodic quality control experiments and occasional corrective actions to address possible problems. As with validation, thorough documentation of the process is needed. The process of quality control has at its roots many statistical underpinnings and assumptions; in fact, the process of quality control is also called "statistical process control." The goal of quality control is to minimize variability and maximize accuracy and precision; this requires measurements of quality metrics and interpretation and analysis of these quality metrics by statistical methods.

Laboratory quality control has two strategies: internal quality control and external quality control. Internal quality control refers to quality control procedures that are performed on a routine (per run or daily) basis within the laboratory. The main aim of internal quality control is to check for precision. The external quality control is performed periodically (e.g., through proficiency testing) and compares the performance of the laboratory with an external quality control (either other laboratories or a reference center).

Quality control experiments require control materials or samples. These samples are close to the sample matrix of the patients and contain the target analyte of a test; the concentration of the analyte is near the clinical decision limits, and this usually means that quality control samples have different levels of target analyte (e.g., low, normal, high). Quality control samples are either standardized samples produced by the test manufacturer, or they can be developed in-house.

It must be noted that most statistical tests used in quality control assume that almost all testing results follow a normal (Gaussian) or near-normal distribution. Understanding of the concepts of this chapter therefore requires a basic understanding of statistical terms including mean, standard deviation, coefficient of variation,

A. Momeni et al., *Introduction to Statistical Methods in Pathology*,
DOI 10.1007/978-3-319-60543-2_11

and normal distribution (and Z-scores) which we have previously discussed in Chaps. 2 and 3.

Detailed discussion of procedures and methods in quality management requires a textbook of its own. For reference purposes, an excellent text in quality control and statistical methods in laboratory medicine is Dasgupta et al. [1]. In this chapter, we will discuss some of the statistical concepts fundamental to the process of laboratory quality control [1–3].

Control Limits

In simple terms, "control limits" are the upper and lower limits of allowed control values, i.e., the results of the control samples can fluctuate between these limits. Thus, control limits have an upper and lower control limit (UCL and LCL, respectively) and a center value (CL) around which the results fluctuate. For each analytical run (or each day), quality control samples must be tested before patient samples, and the laboratory must ensure that the control results are within the control limits; any results outside of quality control limits require corrective actions (e.g., repeat measurement).

The calculation of the control limits is done using a replication study; this requires that a control sample (one control sample per control level) is repeatedly tested (20–30 times, the exact sample size can be calculated using Eq. 10.3 in Chap. 10 for zero outliers or Eq. 10.4 in Chap. 10 for one or more outliers). Then, the mean and standard deviation of the control sample levels are established. The mean plus or minus three standard deviations ($\mu \pm 3\sigma$) forms the "trial limits." If any of the replication study results is outside of the trial limits, then that result is rejected, and the mean and standard variation are calculated again using the remaining values, and a new "trial limit" is set. This process is repeated until no results fall outside the trial limits. The final trial limits are then set as the UCL and LCL, and the final mean is set as the CL.

The basis of control limits is that 99.7% of random fluctuations of the control sample value fall within the control limit. Thus, if a measurement falls outside of the limits, a very rare event has occurred (with a probability of less than 3/1000) warranting an intervention.

Control limits only address random error. For systematic error measurement, a comparison method is needed to identify systematic error. Any systematic error found needs to be corrected using a recovery experiment and calibration (see Chap. 4). In fact, periodic check for accuracy is needed (e.g., through proficiency testing or running calibrators) to allow for identification and rectification of bias in the test.

Levey-Jennings Charts

A "Levey-Jennings chart" allows us to visually inspect the quality control process of the laboratory. Inspection of this chart helps with identification of random and systematic errors without the need for advanced mathematical computations. The basis for this chart is to show the fluctuations of control samples around their mean, and, in fact, these charts are a visual graph of quality control samples over time. The X-axis of the chart is the time (or runs) of the test, and usually 10–20 data points (corresponding to the most recent runs) are plotted on the X-axis. The Y-axis shows the value of the target analyte with CL, ±1 standard deviation (SD) through ±3 SD marked by horizontal lines. The measured values of the quality control sample for each run are then plotted on the chart (Fig. 11.1).

If the quality control value points fluctuate around the average (i.e., no two consecutive points fall on the same side of the CL), then there is only random error in the chart. On the other hand, if two or more points fall on the same side of CL then it may show a systematic error or bias (as the number of consecutive points on the same side of CL increases, the probability that a bias has occurred will increase).

Evaluation of Quality Control Results and Levy-Jennings Charts Thus far, we have discussed the fundamentals of Levy-Jennings chart evaluation. Since these charts are plots of the results of the assays of controls for a given analyte, the question arises as to how to evaluate the validity of the results of these controls themselves. In general, all results should lie within ±2 standard deviations of the mean (cl) value that has been determined for the control. However, there are more

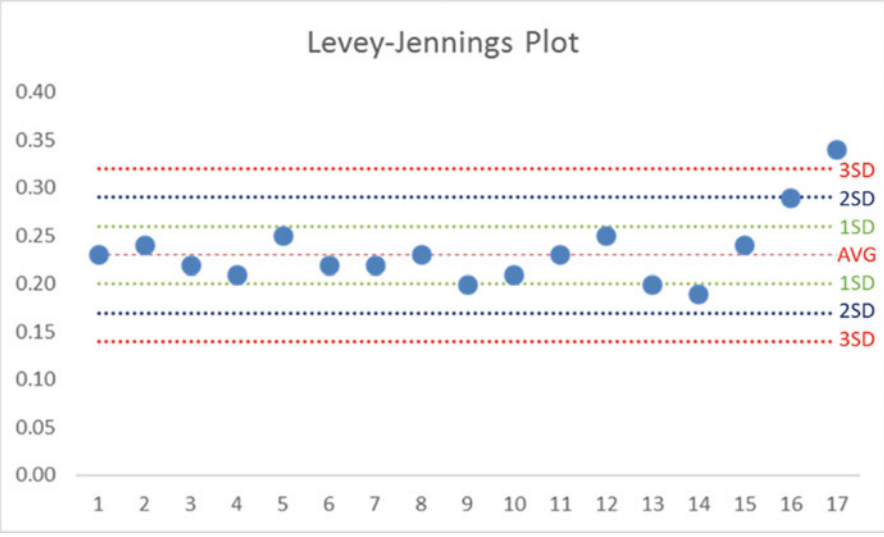


Fig. 11.1 Levey-Jennings plot of an analyte with CL of 0.23 and standard deviation of 0.03

specific rules as to how to perform these evaluations. In addition, there are rules for determining the validity of the results of Levy-Jennings plots. Both sets of rules are called the Westgard rules as we now discuss.

Westgard Rules

In most clinical laboratories, controls for each analyte are assayed once per eight-hour shift so that three controls for each analyte level are analyzed every 24 h. The normal number of controls for each analyte is three; high, normal or intermediate, and low. More recently, it has become customary to use two rather than three controls. Depending on the number of controls used for each analyte, a set of rules, known as the Westgard rules, established by Dr. James O. Westgard of the Department of Pathology at the University of Wisconsin, has been adopted universally as the criteria for accepting or rejecting quality control results.

The rules are as follows:

1. If three controls are used for a given analyte, if two control values lie within ± 2 standard deviations of the mean (cl) value, and the third control is found to have a value that is >2 standard deviations from the mean but is <3 standard deviations from the mean, the results are acceptable.
2. If three controls are used for a given analyte, if two control values lie within ± 2 standard deviations of the mean (cl) value, and the third control is found to have a value that is >3 standard deviations from the mean, the results are unacceptable.
3. If three controls are used for a given analyte, if two control values lie outside of, i.e., are greater than, 2 standard deviations of the mean (cl) value, and the third control is found to have a value that is within 2 standard deviations from the mean, the results are unacceptable. This rule applies even if the two outlying results are within 3 standard deviations of the mean.
4. If two controls are used for any given analyte and if either or both controls are found to lie outside of 2 standard deviations from the mean (cl), the results are unacceptable. This rule applies even if one or both outlying results are within 3 standard deviations of the mean.

These rules are the basic ones that govern all quantitative quality control. There are several further rules relating to trends in quality control as revealed by Levy-Jennings plots:

1_{3S} Rule

If one control value falls beyond the 3 standard deviations limit. This is a criterion used for random error detection. This means that the test has failed QC and corrective actions are needed (Fig. 11.2).

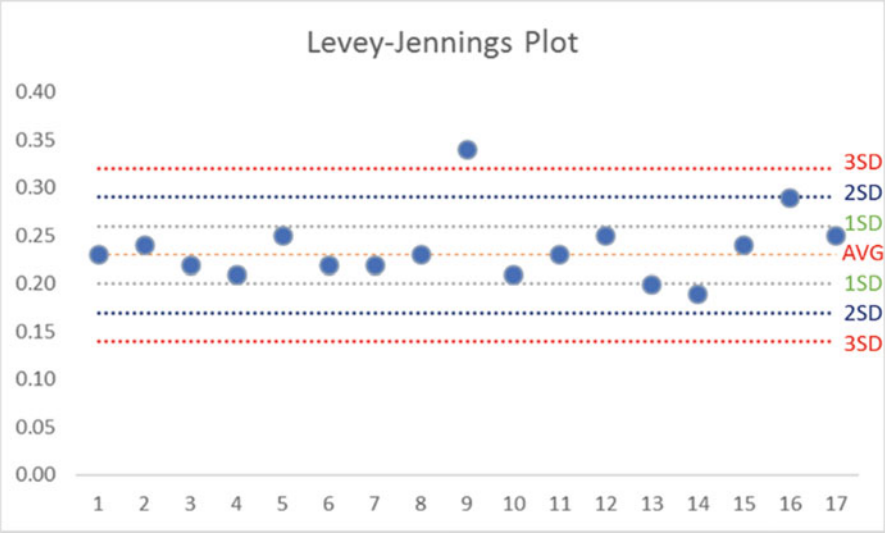


Fig. 11.2 The 1_{3S} rule. Sample 9 has fallen outside the 3 standard deviations limit

2_{2S} Rule
If two consecutive control values fall between the 3 standard deviations and 2 standard deviation limits. This is a criterion used for systematic error detection. This means that the test has failed QC and corrective actions are needed (Fig. 11.3).

R_{4S} Rule
If two consecutive control values are more than 4 standard deviations apart. This is a criterion used for random error detection. This means that the test has failed QC, and corrective actions are needed (Fig. 11.4). Alternatively, if high and low control samples are run and if the sum of deviations of the two samples is more than 4, then the QC has failed.

4_{1S} Rule
If four consecutive control values fall on the same side of the CL and are at least one standard deviation away from CL then QC has failed. This is a criterion used for systematic error detection (Fig. 11.5).

10_x Rule
If 10 consecutive control values fall on the same side of the CL then QC has failed. This is a criterion used for systematic error detection (Fig. 11.6).

Corrective Actions for Results That Are Out of Range If quality control results for an analyte are rejected, patient results for this analyte cannot be reported. It is therefore necessary to investigate the reason for the out-of-range result(s) for the control(s). The first action is simply to repeat the assay on the control in question to

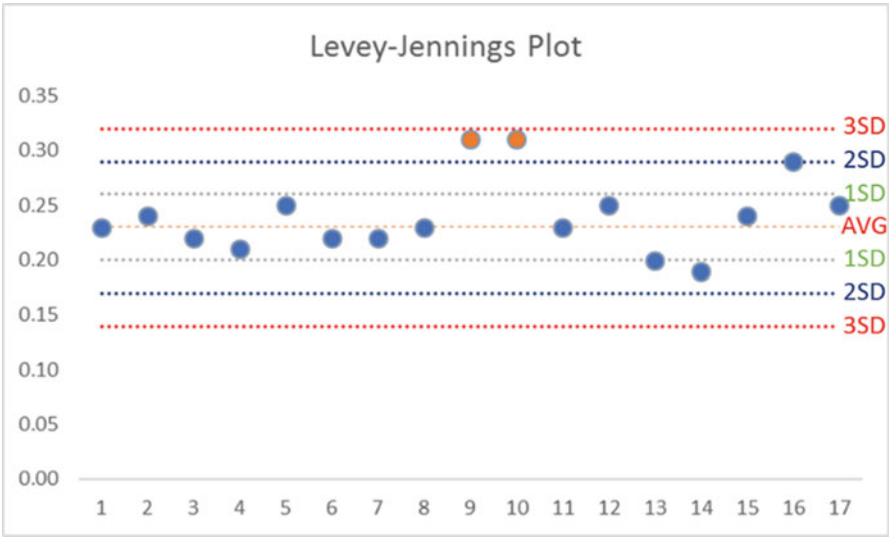


Fig. 11.3 The 2_{2S} rule. Samples 9 and 10 have values that are more than 2 standard deviations higher than the CL

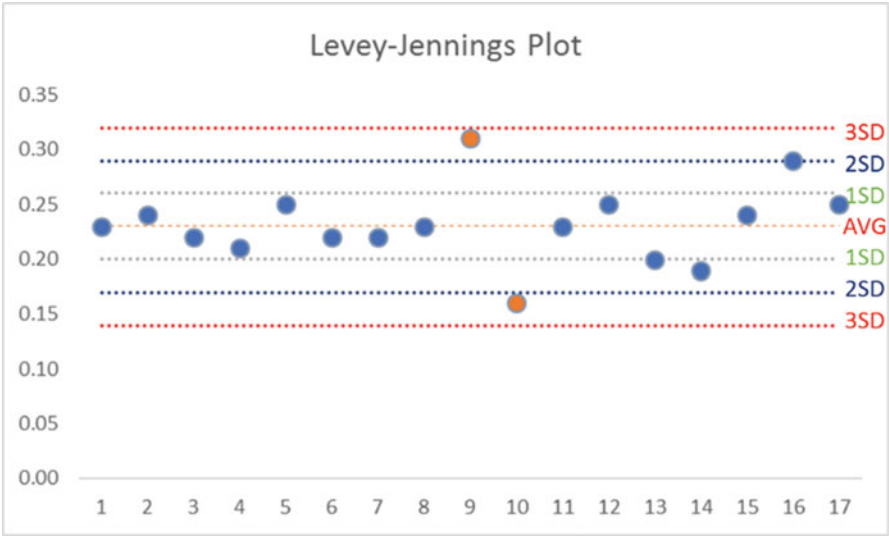


Fig. 11.4 The R_{4S} rule. Samples 9 and 10 are more than 4 standard deviations apart

test whether the error was a random error. If the result is now in range, the control result is acceptable, and the out-of-range result is considered to be a random error. This repeat value is recorded on the Levy-Jennings chart. If the repeated result is still out of range, the assay can be repeated one more time to determine if the result

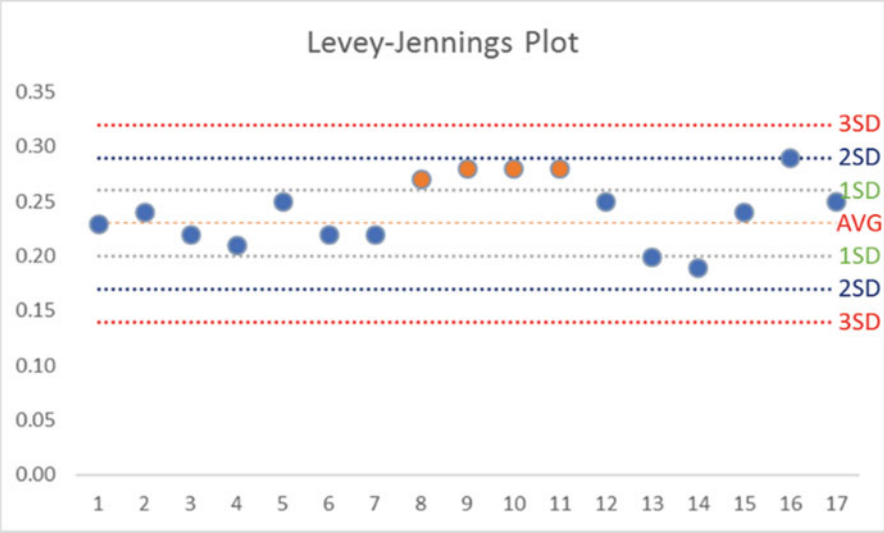


Fig. 11.5 The 4_{1S} rule. Samples 7 through 11 have values that are more than 1 standard deviation away from the CL

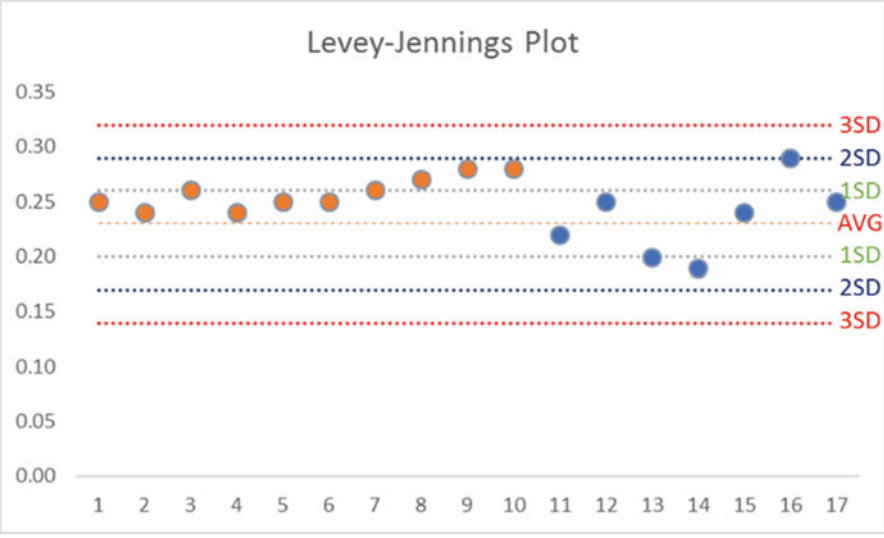


Fig. 11.6 The 10x rule. Sample 1 through 10 have values that fall on one side of the CL

becomes within range. If the second repeat assay still gives an unacceptable result, the next action is usually to recalibrate the assay on the analyzer. *Once recalibration is performed, all controls must be re-assayed.*

If, after recalibration, one or more controls are out of range, it is best to contact the manufacturer of the analyzer and have the analyzer, reagents, and controls checked. Some operators choose to evaluate new assay reagents to determine if the "old" assay reagents have deteriorated and/or to evaluate new controls since the "old" controls may have themselves deteriorated. Performing the latter steps is time-consuming and may not be appropriate for a busy laboratory service. If either the reagent or the control is changed, validation procedures must be implemented. For example, if the reagent is changed, a calibration must be performed. Then, the controls must be assayed. It is advisable to assay the controls 20 times to obtain a new mean and standard deviation with the new reagent. These should be tested using the Student's t-test (Chap. 6) against the former mean and standard deviation with the "old" reagent to determine if the means and standard deviation are the same [2, 4–7].

Average of Normals

"Average of Normals" (AoN) is another internal quality control method that uses patient results as controls for the quality control process. The basic principle behind this approach is that normal individual results are expected to be near the population normal value of the target analyte. Thus, if the target analyte of multiple normal patients is measured, we would expect the results to fluctuate around that population average. The AoN method can only be used to detect systematic error.

One of the methods used in AoN is called the Hoffman and Waid method. In this method, the mean value of normal samples is compared to a mean reference value. While mean population values for many analytes are known, it is recommended that the laboratory calculates the mean value using its own patient population. Usually as part of the validation process, experiments are performed to establish the reference range of the analyte (see Chaps. 2 and 10). In these experiments, the target analyte is measured in a sample of normal patients, and a reference range is established. We can also extract the mean (μ) and standard deviation (SD) of the reference value from these experiments [8].

The next step is to calculate the standard error (SE) of the normal results which is given by

$$\mathrm{SE} = \frac{\mathrm{SD}}{\sqrt{N}} \tag{11.1}$$

with N being the sample size.

Now we can calculate the 95% confidence (95% CI) interval of the normal results:

$$95\%\mathrm{CI} = \mu \pm 1.96 \times \mathrm{SE} \tag{11.2}$$

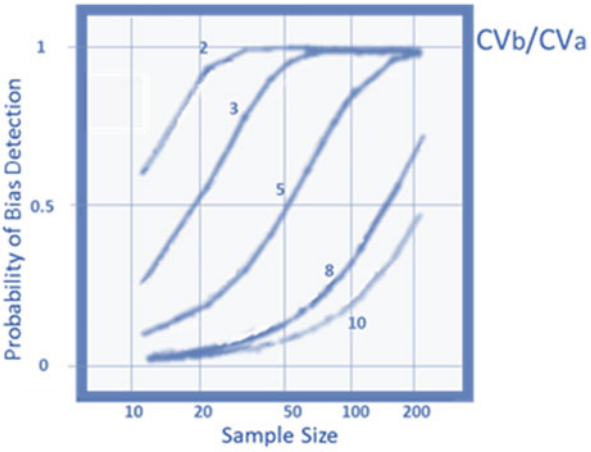


Fig. 11.7 Cembrowski nomogram correlating the CV_b/CV_a with the sample size and probability of bias detection

This confidence interval represents the upper and lower limits of normal for the target analyte.

Every day (or every workshift) a sample of normal results that are reported as part of the routine laboratory operation is extracted (recommended sample size is at least ten patients). These patients will form the control sample of the AoN method, and the mean value of the target analyte in these patients is calculated. If the average of normals is beyond the limits of normal of the target analyte, then we have detected a systematic error (bias) in the results requiring corrective action.

Size calculations are important in AoN method. As the size of the control sample increases, the effectiveness of this method increases. The size is determined by the ratio of the biological variance of the target analyte (CV_b) versus the variance of the method (CV_a), i.e., (CV_b/CV_a). The Cembrowski nomogram shows the correlation of this ratio with sample size and probability of bias detection (Fig. 11.7).

Another approach is to use a t-test (see Chap. 6) to determine if there is any statistically significant difference between the daily normal average and the average value obtained from the mean reference value. Using this approach, obtaining a p-value less than 0.05 signifies the presence of a substantial bias. However, running a t-test requires that the size of the two samples (sample used for calculation of daily average and sample used for calculation of the mean reference) be equal [9–11].

Delta Check

"Delta check" refers to the comparison of a patient's current result to a previous result from the patient. It is one of the quality metrics that can be used in the laboratory and allows for identification of random error. The concept behind delta check is that, if the two tests are performed over relatively brief periods (2–5 days), then the difference in the values of a patient should be minimal (unless a major physiologic/pathologic event has occurred). Delta check is of importance in

checking for possible preanalytical errors as well and can possibly point out sample mislabeling if a particularly discordant result is found.

Delta check or intraindividual variability can be expressed as the absolute difference of two values, or alternatively as the ratio of the two values. Delta can also be expressed as the percent of change from previous ratio.

$$\Delta\% = \frac{|\text{ Current value} - \text{Previous value }|}{\text{Previous value}} \times 100 \tag{11.3}$$

Δ should be less than the "delta check limit." These limits set the boundaries for allowable random variation of the test value in a patient; identification of a delta value beyond this limit either shows a significant error or a major physiologic change (sometimes warranting result flagging or notification of clinicians).

Delta check limits or "reference change value" can be derived from the analytical and biological variation of the target analyte. Each measured analyte has a degree of acceptable variation within an individual (CV_I) (values for the within-individual variation can be found at www.westgard.com/biodatabase1.htm) and an analytical variation (CV_A) which is the coefficient of variation of the test as determined in the lab (calculated at verification or validation). The reference change value can be given by

$$\text{Reference Change Value} = 1.414 \times \ Z \ \times \sqrt{\text{CV}_\text{A}{}^2 + \text{CV}_\text{I}{}^2} \tag{11.4}$$

Z is the corresponding Z-score for the degree of significance of the difference. The degree of significance is either set at 95% or 99% with corresponding two-tailed Z-scores of 1.96 and 2.58, respectively [12, 13].

Example 11.1

Q: Creatinine in a laboratory has been found to have a control mean of 1.5 mg/dL and a standard deviation of 0.3. A patient, who had a creatinine level of 1 mg/dL last week, was tested again today and was found to have a creatinine level of 2 mg/dL. Does the delta check in this patient show a significant change at a level of significance of 95% (within-individual variability of creatinine is 5.95)?

A: First let us calculate the coefficient of variation of creatinine in our lab:

$$\text{CV}_\text{A} = \frac{\sigma}{\mu} = \frac{0.3}{1.5} = 0.2 \tag{11.5}$$

Now we can calculate the reference change value:

$$\text{Reference Change Value} = 1.414 \times \ 1.96 \ \times \sqrt{0.2^2 + 5.95^2} = 16.49\% \tag{11.6}$$

The next step is to calculate the delta percentage:

$$\Delta\% = \frac{2-1}{1} \times 100 = 200\% \tag{11.7}$$

Since the delta percent (200%) is considerably bigger than the reference change value (16.5%) then we can say that the change in the result is significant.

Delta checks are not suitable for all analytes; electrolytes and glucose benefit less from delta checks for error detection, while enzymes generally benefit from delta checks. Part of the rationale for this is explained below; however, for a more detailed explanation, you can read Sampson et al. [2].

Part of the ability of delta checks to identify errors for a test stems from a concept called the "index of individuality." This index is the ratio of within-subject variability (CV_I) of a test to between-subject variability (CV_G) of the test. If this index is low (usually <0.6), it shows that there is variation between individuals that is more than the variation of test in a subject. In such tests the target analyte usually has narrow variation for each person, yet the reference interval (between-subject variation) is usually wide. Thus, changes in the analyte level may not push the patient out of the reference range, yet these changes (e.g., a patient with alkaline phosphatase level (normal range 44–147 IU/L) of 40 IU/L now has an alkaline phosphatase level of 80 IU/L) can still be greater than the expected within-subject variability. In these situations, it is advisable to use delta checks to see if a real change has occurred. The next step will be for the lab to understand whether a significant delta check signifies error or it is due to a clinical status change in the patient.

Tests with high index of individuality are less likely to need delta checks for changes to be noticed since they usually represent narrow reference intervals and significant changes in the analyte level usually lead to an abnormal result flag for that analyte [14].

Moving Patient Averages

"Moving patient averages" can also be used to detect systematic error. Moving average is a series of averages of different patient values; these subsets partially overlap as the shifting forward is smaller than the size of the subset. For example, an average is calculated on the first ten patients, then the next average is calculated by skipping the first patient and calculating the average for patients 2 through 11 and so on.

One of the methods for moving average calculations is called the exponentially weighted moving average ($\bar{X}_{M,i}$), i.e., the average of each batch is weighted down by previous averages. This can be stated as

$$\bar{X}_{\mathrm{M},i} = r\bar{X}_i + (1-r)\bar{X}_{\mathrm{M},i-1} \tag{11.8}$$

where $\bar{X}_{\mathrm{M},i}$ is the current moving average, r is the weight for current values (with possible values of $0 < r < 1$, usually it is set between 0.05 and 0.25 with

recommended value of 0.1), and $\bar{X}_{\mathrm{M},i-1}$ is the previous moving average. $\bar{X}_{\mathrm{M},0}$ is the target value for the analyte (mean of reference samples).

The next step is to compare the exponentially weighted average with control limit for that batch. The control limit is given by

Control limits of exponential moving average

$$= \bar{X}_{\mathrm{M},0} \pm \mathrm{L}\sigma\sqrt{\left|\frac{r}{2-r}\left[1-(1-r)^{2i}\right]\right|} \tag{11.9}$$

Where L is a constant set based on the confidence level (for 95% CI, L equals 2), and σ is the standard deviation of the current batch.

The exponential moving average should be within the control limits of the exponential moving average. Otherwise a drift or shift in values has occurred (systematic error) requiring corrective action.

An alternative to the exponentially weighted moving average is the "Bull's algorithm." In Bull's algorithm, the moving average ($\bar{X}_b$) is based on subsets of 20 samples with 19 representing patient values and one representing the previous moving average. However, the weights assigned to these values are different.

The general formula for Bull's moving average can be written as

$$\bar{X}_{b,i} = (2-r)\bar{X}_{b,i-1} + rD \tag{11.10}$$

where $\bar{X}_{\mathrm{b},i}$ is the current moving average, r is the weight for current values (with possible values of $0 < r \leq 1$, usually set to 1), $\bar{X}_{\mathrm{b},i-1}$ is the previous moving average, and D is calculated from the value of current measurements in the batch. Bull's algorithm is usually used in hematology analyzers, and different companies use different equations for calculation of D. Here we will provide a simple equation for Bull's moving average that assumes a value of 1 for the weight.

$$\bar{X}_{b,i} = \bar{X}_{b,i-1} + \left(\frac{\sum_{j=1}^{\mathrm{N}}\sqrt{X_j - X_{b,i-1}}}{N}\right)^2 \tag{11.11}$$

where N is the number of results in the batch.

The control limits of Bull's moving average are set as $\bar{X}_{b,0} \pm 3\%\bar{X}_{b,0}$ with $\bar{X}_{b,0}$ being the target value for that analyte.

The advantage of Bull's algorithm is that it's not just based on normal values but actually all measurements in a run participate in the calculation of the moving average with outliers' effect usually filtered out by the formula used for calculation of D [15–18].

Statistical Concepts for External Quality Control

Part of the quality control process requires periodic comparison of laboratory performance with external quality measures. This often means comparison of a test result of a sample with a known accurately determined concentration of an analyte or comparison of the test results with the results for the same sample from other laboratories. In the latter case, the comparison is made between the test result in the lab and a consensus mean and standard deviation.

There are two main indices that determine the performance of the laboratory in comparison with other laboratories: "standard deviation index" (SDI) and "coefficient of variation ratio" (CVR).

Standard deviation index is a measure of accuracy and compares the result acquired by the laboratory with the mean results of the peer group. The SDI is given by

$$\text{SDI} = \frac{\text{Laboratory test result} - \text{mean result for peer group}}{\text{Standard deviation of peer group}} \tag{11.12}$$

The interpretation of the SDI is like Z-scores with 95% confidence interval corresponding to an SDI of approximately ± 2. In simple terms, the performance of the laboratory should be within 95% confidence interval of the mean result obtained by the peer group. Any values beyond ± 2 show a significant discrepancy and require investigation and possible corrective actions. However, there are caveats to this approach that are discussed in Chap. 4.

Regulatory agencies use four main rules based on SDI for laboratories for significant systematic error:

1. One SDI value exceeding ± 3.
2. The addition of the two SDI scores for high- and low-level analyte is greater than 4 (this is a warning and does not mean that the lab has failed the proficiency test).
3. Two SDI values from five consecutive proficiency tests are greater than 1 (this is a warning and does not mean that the lab has failed the proficiency test).
4. The average of five successive SDI values is greater than 1.5 (this is a warning and does not mean that the lab has failed the proficiency test).

The coefficient of variation ratio (CVR) on the other hand is a measure that better reflects precision and is the ratio of CV obtained by the lab to the consensus CV.

$$\text{CVR} = \frac{\text{Laboratory CV}}{\text{Consensus CV}} \tag{11.13}$$

CVR >1.5 signifies a need for investigation of the cause of imprecision with values greater than 2 requiring corrective action.

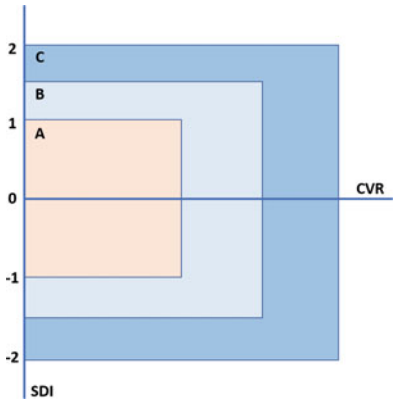


Fig. 11.8 SDI/CVR chart shows the SDI on the X-axis and CVR on the Y-axis

SDI and CVR can be showed together in an SDI/CVR chart (a chart of total error) which plots SDI on the Y-axis and CVR on the X-axis. The plot has three regions (a, b, and c) with ideal performance expected to be in region a. Performance in region b is a warning sign and requires investigation. Performance in region c means that the lab has a total error greater than that allowed by the external quality control (i.e., the lab has failed the proficiency test) and requires corrective actions (Fig. 11.8) [4].

Summary

In this chapter, we reviewed the statistical concept that is used in laboratory quality management. Quality management is an ongoing effort that ensures that the laboratory is performing satisfactorily and that test results are accurate and precise. This requires frequent monitoring of quality metrics and application of statistical tools to identify possible performance problems.

References

1. Dasgupta A, Wahed A. Clinical chemistry, immunology and laboratory quality control: a comprehensive review for board preparation, certification and clinical practice. Academic Press: USA; 2013.
2. Sampson ML, Rehak NN, Sokoll LJ, Ruddel ME, Gerhardt GA, Remaley AT. Time adjusted sensitivity analysis: a new statistical test for the optimization of delta check rules. J Clin Ligand Assay. 2007;30(1–2):44–54.
3. McPherson RA, Pincus MR. Henry's clinical diagnosis and management by laboratory methods. Elsevier Health Sciences: USA; 2016.
4. Westgard JO. Six sigma quality design and control. Madison: Westgard QC, Incorporated; 2001.

5. Karkalousos P, Evangelopoulos A. Quality control in clinical laboratories. INTECH Open Access Publisher: USA; 2011.
6. Westgard JO, Westgard SA. The quality of laboratory testing today. Am J Clin Pathol. 2006;125(3):343–54.
7. Westgard JO. Internal quality control: planning and implementation strategies. Ann Clin Biochem. 2003;40(6):593–611.
8. Green GA, Carey RN, Westgard JO, Carten T, Shablesky L, Achord D, Page E, Van Le A. Quality control for qualitative assays: quantitative QC procedure designed to assure analytical quality required for an ELISA of hepatitis B surface antigen. Clin Chem. 1997;43(9):1618–21.
9. Cembrowski GS, Chandler EP, Westgard JO. Assessment of "average of normals" quality control procedures and guidelines for implementation. Am J Clin Pathol. 1984;81(4):492–9.
10. Douville P, Cembrowski GS, Strauss JF. Evaluation of the average of patients: application to endocrine assays. Clin Chim Acta. 1987;167(2):173–85.
11. Wheeler LA, Sheiner LB. A clinical evaluation of various delta check methods. Clin Chem. 1981;27(1):5–9. Chicago
12. Ricós C, Alvarez V, Cava F, Garcia-Lario JV, Hernandez A, Jimenez CV, Minchinela J, Perich C, Simon M. Current databases on biological variation: pros, cons and progress. Scand J Clin Lab Invest. 1999;59(7):491–500.
13. Bull BS, Elashoff RM, Heilbron DC, Couperus J. A study of various estimators for the derivation of quality control procedures from patient erythrocyte indices. Am J Clin Pathol. 1974;61(4):473–81.
14. Aslan D, Kuralay F, Tanyalsin T, Topraksu M. Use of averages of patient data for quality control. Accred Qual Assur J Qual Comp Reliab Chem Meas. 1999;4(9):431–3.
15. Lucas JM, Saccucci MS. Exponentially weighted moving average control schemes: properties and enhancements. Technometrics. 1990;32(1):1–12. Chicago
16. Hunter JS. The exponentially weighted moving average. J Qual Technol. 1986;18(4):203–10. Chicago

Critical Appraisal of Diagnostic Studies

12

Introduction

It has been shown that diagnosis utilizes approximately 5% of healthcare costs, yet 60% of the clinical decision-making process is dependent on the diagnosis. Reduction of misdiagnosis is now being recognized as a major goal in patient safety efforts. Approximately 40,000–80,000 hospital deaths per year in the United States are attributed to misdiagnosis; many of these deaths are preventable deaths that can be avoided if a correct and timely diagnosis is made. Pathology and laboratory medicine in their major roles as a source of diagnosis or a major contributor to the diagnostic process require a push toward more accurate and precise testing and diagnosis, and this requires integrating the best available evidence into the everyday practice of pathology.

Practice of medicine has evolved into a concept known as "evidence-based medicine" (EBM) where the decisions of the physicians are made based on a combination of best available evidence, experience, and patient needs and values; in the field of pathology and diagnostic medicine, this implies that the pathologist can formulate a relevant question regarding their practice, find the evidence that answers the question, appraise the evidence, and apply it to their practice.

A significant part of the evidence-based practice of medicine is determining what best evidence is; very often, the practicing pathologist is faced with a flurry of new diagnostic studies that make claims that can alter the current practice of pathology. However, before the findings of the diagnostic studies are implemented, there needs to be a thorough review of the evidence with possible sources of bias and inaccuracy identified and addressed. This process is known as critical appraisal of evidence. Critical appraisal requires that the quality of the evidence is assessed especially regarding claims of benefit, effectiveness, and applicability. It must be noted that the process of critical appraisal is often intertwined with a systematic review of evidence, where findings of one study should be assessed and interpreted alongside findings from similar studies. Such review of evidence is often reported as systematic reviews and/or meta-analysis [1. 2].

A. Momeni et al., *Introduction to Statistical Methods in Pathology*,
DOI 10.1007/978-3-319-60543-2_12

In this chapter, we will focus on critical appraisal as it applies to diagnostic studies and tests and introduce the concepts of systematic review and meta-analysis in the context of diagnostic studies.

Levels of Evidence

Evidence can come in many forms and formats from letters to the editor to systematic reviews. Each of these formats has associated criteria for publication and dissemination; based on this, the different forms and formats of evidence can be categorized into hierarchical quality levels. As we move up through the hierarchy, due to increasingly stringent conditions for conduct of the study as well as reporting the study, the quality of the evidence tends to increase and the possibility of bias decreases. This, in turn, means that evidence from a higher level is more reliable and can be adopted more easily into the practice of medicine.

Unfortunately, much of the evidence in pathology, especially anatomic pathology, is of lower quality and leads to recommendations that lack strength: many of the papers are from cross-sectional case-control studies or from limited cohorts with inadequate statistical power. Thus, every pathologist needs to know how to assess the quality of the evidence and identify sources of bias in the literature.

Based on the recommendation from "Center for Evidence-Based Medicine" (CEBM, www.CEBM.net), in diagnostic medicine, the highest level of evidence (level 1a) is a systematic review of homogeneous level 1 diagnostic studies; the term homogeneity means that the findings of all diagnostic studies were consistent with each other with no conflicting finding reported. Another type of study/evidence that also has the highest quality level is a clinical decision rule based on multiple level 1 studies from different clinical centers.

"Systematic reviews" are studies that employ strategies to pool the evidence relating to a subject to limit bias. These studies require evidence assembly, critical appraisal, and synthesis of all relevant studies. These studies follow strict rules for assessing evidence and including them in the final evidence synthesis. Systematic reviews have the potential to provide high-quality recommendations if the rules were strictly followed. Meta-analysis is a quantitative summation of results provided in a systematic review.

Level 1 "clinical decision rules" (CDRs), also known as clinical decision support, are evidence-based, externally validated clinical decision guidelines that address a specific question. Clinical decision rules are often collated and produced using expert panels based on the best available evidence. In pathology and laboratory medicine, clinical decision rules often define diagnostic criteria for diagnosis of a disease condition.

We will elaborate the concepts of systematic review and clinical decision rules later in this chapter.

The next evidence level (1b) includes clinical decision rules that are tested and validated within one institute. This level also includes evidence from validating cohort studies with satisfactory reference standards (controls).

Validation cohorts are studies which are conducted to confirm the findings of a derivation cohort. This often involves checking to see if the findings from one cohort can be replicated in another patient sample (ideally conducted in a different institute by different researchers). For example, a derivation cohort has found out that maternal level of protein X is a predictor of fetal abnormalities. A validation cohort will then be if protein X is measured in another group of pregnant women, and they are followed to see if there will be any fetal anomaly. A good validation cohort requires good reference standards. We discussed the statistical concepts relevant to external validation in Chap. 7, and we will further explain the conduct and design of cohort studies in Chap. 13.

The third high-quality evidence level (1c) includes studies with results showing "absolute specificity rule in (absolute SpPin) and/or sensitivity rule out (Absolute SnNout)." As you may recall from Chap. 2, as the specificity of a test for a disease increases, the probability that a positive test result rules in the disease increases. An absolute SpPin result means that a study has found a test to have a very high specificity that a positive test equals having the disease: this often requires specificity in excess of 95%. An absolute SnNout requires a test to have a sensitivity of more than 95% so that a negative test result effectively rules out the disease.

The second level of evidence has two sublevels: 2a and 2b. Level 2a evidence is a systematic review with homogeneity which summarizes level 2 (or a mix of level 2 and level 1) evidence. Level 2b evidence includes exploratory cohorts and clinical decision rules that are only validated internally. Exploratory cohorts are cohort studies that follow two (or more) groups of individuals with and without an exposure (or test result) and document the incidence of a target outcome. For example, a study that follows patients with normal and high PSA levels and documents the incidence of prostate cancer in the two groups is an exploratory cohort. For exploratory cohorts to be included as a level 2b evidence, they need to have good reference standards.

Third level of evidence also has two sublevels, 3a and 3b, with 3a being a systematic review of 3b studies. Level 3b evidence includes a nonconsecutive cohort study or a cohort study without satisfactory reference standards. The term "nonconsecutive cohort" refers to a study where all eligible patients are not included in the study.

Level 4 evidence is evidence obtained from case-control studies or from cohort studies with non-independent or poor reference standards.

Finally, level 5 evidence consists of expert opinion or inferences from bench results, physiology, etc.

Levels of evidence as it pertains to diagnostic studies are summarized in Table 12.1.

Table 12.1 Levels of diagnostic evidence

Level	Type of study
1a	Systematic review with homogeneity or clinical decision rules based on level 1 studies
1b	Validation cohort study (with good controls) or CDR validated within one institute
1c	Accuracy studies with very high sensitivity or specificity
2a	Systematic review of level 2 (or a mix of level 1 and 2) studies
2b	Exploratory cohorts with good controls and internally validated CDRs
3a	Systematic review that includes level 3 studies
3b	Non-consecutive cohort
4	Case-control studies, cohorts with poor reference standards
5	Expert opinion

Evidence-Based Recommendations

The end goal for any review of evidence is to formulate recommendations that can guide the everyday practice of the pathologist. Evidence-based recommendations are made using the best available evidence and information. Furthermore, recommendations consider factors such as relative importance of outcomes, baseline risks of outcomes, magnitude of relative risk (or odds ratio), absolute magnitude of effect, precision of the findings, and associated costs. These are summarized in Table 12.2. We have explained many of these factors in previous chapters, and consequently in this chapter we will focus more on quality metrics of the evidence.

The quality levels of the evidence are of utmost importance since the strength of recommendations derived from the evidence is dependent on the quality level of evidence. Generally, the recommendations have four strength grades: A, B, C, and D. Grade A recommendations are guidelines or recommendations that are based on solid evidence and have been shown to be consistently valid in different populations or practice settings. Grade A recommendations are generally based on level 1 evidence. In these recommendations, the benefits clearly outweigh the risks (or vice versa), and these recommendations can be applied in most clinical settings. Pathologists should follow a grade A recommendation unless strong evidence or rationale for an alternative approach exists.

Grade B recommendations draw from consistent level 2 or 3 evidence (i.e., no heterogeneous or contradictory studies) or are extrapolations from level 1 evidence. Extrapolation refers to situations where the findings of a study are applied to a different setting with clinically important distinctions, for example, if the recommendation of a level 1 study regarding the use of a test for inpatients is used for an outpatient setting. Grade B recommendations should be followed in the right clinical context in combination with the clinical judgment of the pathologists.

Grade C recommendations are based on level 4 evidence or are extrapolations of level 2 or 3 evidence.

Grade D recommendations are based on level 5 evidence or are derived from evidence with significant inconsistency.

Table 12.2 Factors to be considered in making an evidence-based recommendation

Factor	Approach to evaluation of evidence
Quality of evidence	Strong recommendations require high-quality evidence. Possible sources of bias should be assessed with compensations in quality level of evidence for the degree of bias. Questions about accuracy and validity of tests should be answered before considering other factors
Relative importance of the outcomes	The evidence should be clinically significant as well as being relevant to the practice setting of the pathologist. For example, a cutting-edge high-cost test may not be appropriate for a laboratory providing service to a small population
Baseline risks of outcomes	Knowing the baseline risks or, in diagnostic terms, the pretest probability is an important factor in the decision-making process. Also, factors such as the burden associated with the disease should also be considered. For example, thalassemia screening is a low-yield test, but since thalassemia major is a considerable health burden, then screening is still justified
Magnitude of relative risk	In diagnostic medicine, this translates to likelihood ratio; the larger the likelihood ratio of a test, the stronger is a recommendation based on that test. For example, cardiac troponins have a very high positive likelihood ratio: if they are positive, it is highly likely that the patient has myocardial infraction
Absolute magnitude of effect	Sometimes absolute effect is more important than relative risk. For example, why should we adopt a new high-cost test to differentiate two cancer subtypes if the current treatment and prognosis for them are the same?
Precision	Precision in terms of both repeatability and replicability is important in diagnostic tests: If a test is repeated multiple times, will the results remain the same? If a test is run in a different laboratory or is used for a different population, will the results remain the same? Imprecision decreases the strength of a recommendation
Cost	Cost usually has an inverse relationship to strength of recommendation, with high cost issues of generalizability and accessibility arising. However, cost alone is not always a good metric for financial burden of a test, and one should also examine the cost-effectiveness of a test

In general, clinical decision making should rely on grade A recommendations. In situations where grade A recommendations do not exist, grade B recommendations can be used. Decision making should not be based on grade C recommendations as they are likely to have significant bias and usually cannot be the sole basis for decision making. Grade D recommendations should be avoided and alternative recommendations or approaches sought [3].

Critical Appraisal of Diagnostic Studies

Evidence-based recommendations are not solely made based on the level of evidence. To make these recommendations, the actual quality of evidence should be assessed first. Level 1a evidence can still suffer from major bias if the methodology or conclusions have bias (the nature of bias in systematic reviews is different, and it will be explained in a separate section). Thus, after determining the level of the evidence, the evidence should be analyzed for possible errors, inconsistencies, or biases (we will discuss the possible error forms in the next section). Many EBM organizations have checklists or questionnaires that allow a quick assessment of the evidence (e.g., British Medical Journal EBM Toolbox available at http://clinicalevidence.bmj.com/x/set/static/ebm/toolbox/665061.html).

Generally, these questionnaires aim to answer three questions: Are the results accurate? Does the test have acceptable discrimination power? (i.e., Is the test capable of distinguishing affected individuals from unaffected individuals?) Are the results applicable to your practice setting?

Accuracy of results means that the diagnostic study is valid, i.e., it is measuring what it is supposed to measure, and it is measuring it correctly. Studies that establish the accuracy of a test are unimaginatively called "diagnostic accuracy tests": In these tests a series of patients are tested using the target (index) test as well as a reference standard, and a blinded comparison is made between the results of two tests. Alternatively, for conditions that do not have a reference standard, a cohort with case and control groups is studied, for example, individuals who tested positive for a test and individuals who tested negative are followed up for a period to determine the development of the target condition.

Thus, checking for validity requires answering questions such as:

- Was an appropriate spectrum of patients studied? For example, a molecular test was found to rule out thyroid malignancy in cytology specimens. If the diagnostic accuracy study was performed on a group of patients with very low pretest probability, then the study has spectrum bias.
- Was everyone tested using the reference standard? Sometimes the reference standard is used as a confirmatory test only in cases where the index test was positive; for example, in study of cervical Pap smears (index test), biopsies (reference test) were only performed if the Pap smear was positive. This is sometimes due to the possible harm or cost associated with the reference test. Selective testing of patients with the reference standard can lead to a significant overestimation of the accuracy of the diagnostic test (this is known as verification bias).
- Was there an objective and blinded comparison between the results of the two tests? Blinded comparison means that interpretation of the results should be blinded to the results of the other test or to the existence of the target condition in the patient. For example, in case-control studies, we choose a known group of affected patients and a known group of healthy individuals and compare the test performance in the two groups versus a reference standard. This introduces

Table 12.3 Sources of bias in diagnostic studies

Bias	Definition
Spectrum bias	Spectrum bias is present when a study uses a highly selective sample. The spectrum of patients in a study thus may not be reflective of the true clinical setting
Verification bias	Verification bias occurs when only a selected number of patients tested with the index test get the reference test. This usually tends to overestimate the effectiveness of the index test. Another type of verification bias, known as differential verification, occurs when some patients are verified with one test, and some are verified with another test. Yet another type of verification bias, known as incorporation reference bias, occurs when the index test is part of the reference test or contributes to the reference test
Observer bias	This bias occurs when the comparisons are not blinded, objective, or independent and usually leads to an overestimation of the effect. Studies have shown observer bias to be one of the main contributors to overall bias

significant bias. The index test and the reference test should be performed in the same group of individuals blinded to their disease status.

– Were the results confirmed in a second study? In other words, was there external validation of the results? The findings from each exploratory cohort need to be confirmed using a validation cohort.

If a study fails to satisfy any one of these questions, then an assessment should be made of whether the associated bias is large enough to make the results of the study invalid. Even if the results remain valid, some people consider downgrading the level of evidence if any bias exists (even inconsequential bias). We have summarized the sources of bias in diagnostic studies in Table 12.3.

In fact, part of the evaluation of a diagnostic study involves finding the answers to the above mentioned questions. Readers should always get a clear understanding of the aim of the study, the spectrum of patients, the index test performed, the reference standard performed, and whether blinding was performed. If any of these elements is lacking or the relevant bias question cannot be answered, then there should be a suspicion of bias in the data presented.

Another important factor in evaluating diagnostic tests is to determine what type of results is being reported for the study. We have explored the types of answers relevant to study of accuracy in Chap. 2. The aim of some studies is to establish sensitivity and specificity of a test, i.e., to determine the test accuracy. Other studies, however, aim to measure the performance of the test in a population; these measures are called predictive values (including likelihood ratio) and usually are not generalizable to other populations. Distinguishing between these accuracy goals is very important and can (or should) influence the decision of a pathologist to adopt a test.

Finally, considering all the other factors mentioned earlier in this section, the pathologists should decide whether a test is applicable to their clinical practice setting [4].

Systematic Reviews

Systematic reviews are summary evidence that are developed using systematic methods to identify, select, critically appraise, and collate primary studies relevant to a specific question. Systematic reviews can also incorporate a quantitative collation of the results known as meta-analysis where the findings and results of the studies are combined to provide a single statistical measure. For example, likelihood ratio of a test from multiple studies can be combined to provide a single likelihood ratio.

The overall aim of a systematic review is to minimize bias and summarize the evidence in order to facilitate development of practice guidelines or recommendations. Different guidelines exist regarding the conduct of a systematic review; we recommend "Cochrane Handbook for Diagnostic Test Accuracy Reviews" accessible at http://methods.cochrane.org/sdt/handbook-dta-reviews.

Regardless of the guideline followed, the general steps for conducting a systematic review remain the same. The process of writing a systematic review requires six steps (Table 12.4):

The process starts by forming a research question and setting the goals for a systematic review. The questions should have clinical relevance and relate to a diagnostic challenge. Usually the questions and objectives are formatted based on the "PICO framework". This framework breaks the question into four main components:

1. Patients (P): What is the target population? For example, the target population can be women older than 20 years requiring cervical cancer screening.
2. Index test (I): What is the test being evaluated? For example, HPV DNA testing.
3. Comparator (C): What is the reference standard? For example, cervical Pap smear.
4. Outcome (O): What is the outcome? For example, sensitivity.

Thus, following the PICO framework and using the examples above, we can define the research question: "Sensitivity of HPV DNA testing in comparison with cervical Pap smear for screening of cervical lesions."

A more complete framework also includes study design in defining the question (hence PICOS). This will determine what types of study will be included in the systematic review process. As part of the PICO framework, for each question,

Table 12.4 Steps in conducting a systemic review

Step 1: Formulating the question and establishing the eligibility criteria
Step 2: *C*onducting systematic search and evidence selection
Step 3: Evidence quality assessment and critical appraisal
Step 4: Extracting data from individual studies
Step 5: Analysis and synthesis of data
Step 6: Preparing the report

subtopics may also be considered; for example, for HPV DNA testing, the technology and the method or analyzer used can also be considered.

Usually it is recommended that after the initial formulation of the research question, a limited search is conducted and the research question revised based on the findings from the limited search for evidence.

Defining the question is followed by developing the inclusion/exclusion criteria. These criteria determine the evidence to be included in the review based on relevance of the studies to the research question as well as the quality of evidence. The inclusion criteria should be set so that all relevant evidence is included in the review process and is usually derived from the PICO framework. The exclusion criteria, on the other hand, should be set to exclude studies with unsatisfactory quality or significant bias.

Step two will be to conduct a systematic search: this should be an exhaustive process that searches for evidence in multiple databases in order to identify as many studies relating to the research question. All systematic reviews require thorough documentation of this step. The searching process starts with defining a "search phrase" and "search terms": this can be aided by thesaurus-based databases (e.g., MEDLINE/PubMed). The search phrase is a logical statement formed by search terms with Boolean operators (e.g., "AND," "OR," "NOT"). Remember that adjustment of search phrase for different databases is sometimes needed.

This is followed by selection of databases for conducting the search; it is recommended that more than one database is used for the search. The next phase is to conduct the search and extract the evidence that meet the inclusion criteria.

It is important to know that systematic search is a multiphase process: the initial search may yield a limited number of studies, and thus expansion of the search is often needed. Different methods are recommended for search expansion. One is called "snowballing" and refers to searching the reference list or bibliography of the studies obtained in the initial search to identify possible studies that were left out. It is also important to search for "gray" literature: these are study results that may not have gone through the vigorous peer review process and are presented as conference abstracts or reports. This so-called gray evidence forms an important part of the search as they may represent "publication bias" (explained later). The researcher may also consider searching in alternative languages.

The extracted evidence should then be screened to select the relevant studies; this process should be conducted by at least two researchers with disagreements either resolved through consensus or by involving a third reviewer. There are different guidelines for the selection process. We suggest using the guidelines presented in Chap. 7 of the Cochrane handbook (see above). The screening process starts by review of titles and abstracts which can allow a quick review of the evidence and exclusion of studies that do not meet the research question or eligibility criteria. For some studies, you may be required to review the full text of the evidence to make this determination. A record of excluded evidence with reasons for exclusion should be kept. The results of the selection process should be summarized in a "PRISMA-P flowchart" (Fig. 12.1).

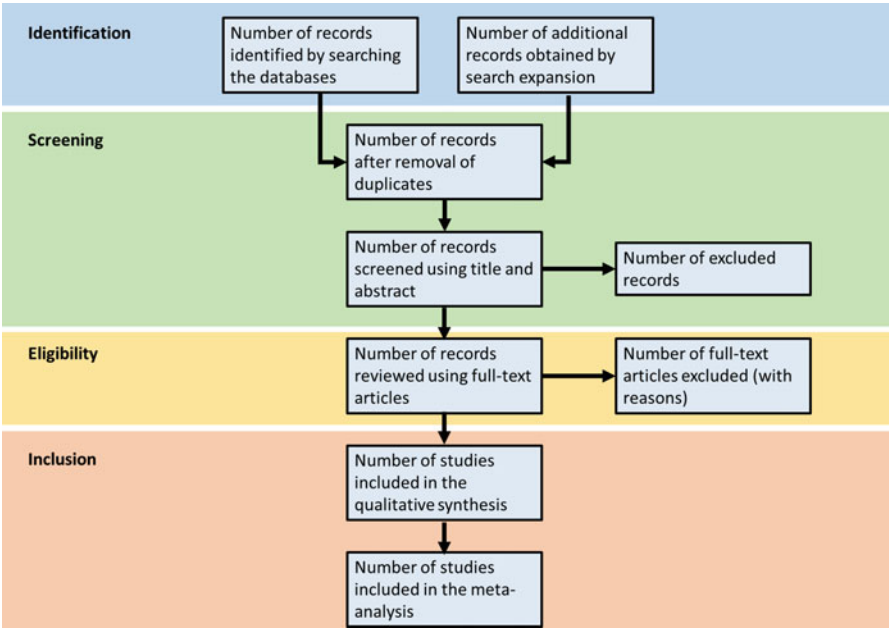


Fig. 12.1 The PRISMA flowchart

The next step is to critically appraise the evidence and identify bias and variation in the selected evidence. Variation refers to the differences in the design and conduct of the studies. Sometimes the variation may be significant enough to render the findings of a study irrelevant to the research question of the systematic review. Bias refers to any systematic error in the data due to flawed study design or conduct that may render the diagnostic accuracy reported in the study unreliable and inaccurate. We explained some of the major biases in diagnostic accuracy tests in the previous section; if a major bias has occurred, then that study should be excluded from the systematic review process or alternatively the effect of that article on the outcome can be explored as part of the meta-analysis. For critical appraisal of diagnostic accuracy studies, we suggest using the revised "Quality Assessment of Diagnostic Accuracy Studies" (QUADAS-2) tool available at http://www.bristol.ac.uk/social-community-medicine/projects/quadas/quadas-2.

The QUADAS-2 tool appraises the evidence with regard to risk of bias and applicability. For each of these concerns, different domains are checked: For bias, patient selection, index test, reference standard, and flow and timing are assessed. For applicability, patient selection, index test, and reference standard are assessed. For each study, each one of these domains is ranked as high risk, low risk, or unclear risk. This tool should be modified based on the research question and requirements of the systematic review.

After completion of selection and appraisal, the information required to answer the research question should be extracted from the studies by two (or more)

Table 12.5 2 × 2 contingency table for diagnostic accuracy studies

	Index test	
	Positive	Negative
Comparator positive	True positive (TP)	False negative (FN)
Comparator negative	False positive (FP)	True negative (TN)

independent researchers. The degree of agreement between the researchers at this stage is important and is usually reported as a Kappa coefficient (see Chap. 5) in the final results. There is a set of standard information that should be extracted from all studies; however, for diagnostic accuracy studies, the 2 × 2 contingency table information should also be extracted (Table 12.5).

Other information that should be recorded are cutoff values and the rationale for the chosen cutoff value. In studies where the information needed for the 2 × 2 contingency table is not available, then the researchers are required to calculate the information from the other available information in the report (e.g., if the number of patients with the disease and healthy patients are known and sensitivity and specificity are provided, then TP, TN, FP, and FN can be calculated).

The results can then be summarized and used for qualitative synthesis where individual results are provided but no pooling of data occurs. This can be helpful to draw generalized conclusions about the research question. However, while it is often preferred if the data are pooled and a quantitative synthesis is performed, this is contingent on high degree of homogeneity between the studies. Unfortunately, due to the "threshold effect" (explained later) and the fact that many diagnostic studies are low-level evidence, the risk of bias in these studies is high, and thus performing a meta-analysis can sometimes lead to erroneous conclusions drawn from the data. We will discuss the statistical concepts related to meta-analysis in the next section [4, 5].

Meta-analysis

Meta-analysis involves a quantitative summation of the findings of a systematic review. Meta-analysis requires the extraction of quantitative results from all the studies included in the systematic review. For diagnostic meta-analysis, information such as sensitivity, specificity, positive likelihood ratio and negative likelihood ratio, and overall diagnostic accuracy should be extracted (see Chap. 2).

The calculated descriptive statistics of the individual studies can be used to draw a descriptive forest plot and a summary receiver operating characteristics (SROC) plot and curve.

Forest Plot

Forest plots show the calculated descriptive statistics (e.g., sensitivity) of all individual studies in one plot. The Y-axis shows the individual primary studies. The studies are ordered based on their sample size (usually in decreasing order with

studies with smaller sample size being higher on the Y-axis). The first line of the Y-axis is usually used to show the summary of the descriptive measure. The X-axis shows the calculated descriptive statistics along with its 95% confidence interval.

There are different approaches to calculating the summary measure. The simplest approach is called "separate pooling with fixed effect" where each accuracy measure is pooled separately.

In this approach, the summary measure is calculated using a concept known as "inverse variance weighting." The rationale behind this concept is that the noisier the results from one study are (i.e., their variations are more or in other words their 95% confidence interval is larger), the less they should contribute to the calculation of the 95% confidence interval of the summary measure. Based on this rationale, for calculation of the mean summary measure, each calculated descriptive statistics is weighted down by its variance. Thus:

$$\omega_i = \frac{1}{{\sigma_i}^2} \tag{12.1}$$

where ω_i is the weight of a study and ${\sigma_i}^2$ is the variance of the calculated descriptive statistics. For sensitivity and specificity, the variance can often be calculated based on sample size and observed sensitivity (or specificity):

$${\sigma_i}^2 = \frac{X_i(1 - X_i)}{m_i} \tag{12.2}$$

where m_i is the sample size of the study and X_i is the descriptive statistic (e.g., sensitivity) reported as a proportion (e.g., 0.95).

Consequently, the weighted mean ($\widehat{\mu}$) can be calculated using:

$$\widehat{\mu} = \frac{\sum_{i=1}^{n} \omega_i X_i}{\sum_{i=1}^{n} \omega_i} \tag{12.3}$$

where n is the number of studies in meta-analysis and X_i is the calculated descriptive statistics for each study.

The variance of the weighted mean is given by:

$$\mathrm{Var}(\widehat{\mu}) = \frac{1}{\sum_{i=1}^{n} \omega_i} \tag{12.4}$$

Example 12.1

Q: A systematic review was performed to determine the sensitivity and specificity of test A for a disease. The systematic review identified ten studies that met the eligibility criteria. The accuracy data from these ten studies were extracted and are summarized in Table 12.6. What is the summary sensitivity and specificity for this study? Derive the forest plot for the sensitivity from the systematic review.

Table 12.6 Summary of results from Example 12.1

Study number	Sample size	Sensitivity	Lower CI sensitivity	Upper CI sensitivity	Variance sensitivity	Weight sensitivity	Specificity	Lower CI specificity	Upper CI specificity	Variance specificity	Weight specificity
1	300	0.85	0.809594	0.890406	0.000425	2352.941	0.65	0.596026	0.703974	0.000758	1318.681
2	150	0.9	0.85199	0.94801	0.0006	1666.667	0.6	0.5216	0.6784	0.0016	625
3	120	0.8	0.728431	0.871569	0.001333	750	0.7	0.618007	0.781993	0.00175	571.4286
4	100	0.95	0.907283	0.992717	0.000475	2105.263	0.65	0.556514	0.743486	0.002275	439.5604
5	100	0.9	0.8412	0.9588	0.0009	1111.111	0.6	0.50398	0.69602	0.0024	416.6667
6	70	0.8	0.706294	0.893706	0.002286	437.5	0.6	0.485234	0.714766	0.003429	291.6667
7	65	0.85	0.763193	0.936807	0.001962	509.8039	0.75	0.644731	0.855269	0.002885	346.6667
8	20	0.7	0.49916	0.90084	0.0105	95.2381	0.55	0.331964	0.768036	0.012375	80.80808
9	20	0.98	0.918642	1	0.00098	1020.408	0.6	0.385293	0.814707	0.012	83.33333
10	20	0.75	0.560224	0.939776	0.009375	106.6667	0.8	0.624692	0.975308	0.008	125

A: Using the Eq. (12.3) and based on the calculated sensitivity and specificity of the studies and their respective weights (1/variance), we can calculate the summary sensitivity and specificity:

$$\widehat{Summary\ Sensitivity} = \frac{\sum_{i=1}^{10} \omega_{i(sen)} Sensitivity_i}{\sum_{i=1}^{10} \omega_{i(sen)}} = 0.8891 \tag{12.5}$$

$$\widehat{Summary\ Specificity} = \frac{\sum_{i=1}^{10} \omega_{i(spe)} Specificity_i}{\sum_{i=1}^{10} \omega_{i(spe)}} = 0.6507 \tag{12.6}$$

Thus, the meta-analysis shows that the pooled (summary) sensitivity and specificity of test A for diagnosis of the target disease is 0.89 and 0.65, respectively. The forest plot for sensitivity is shown in Fig. 12.2.

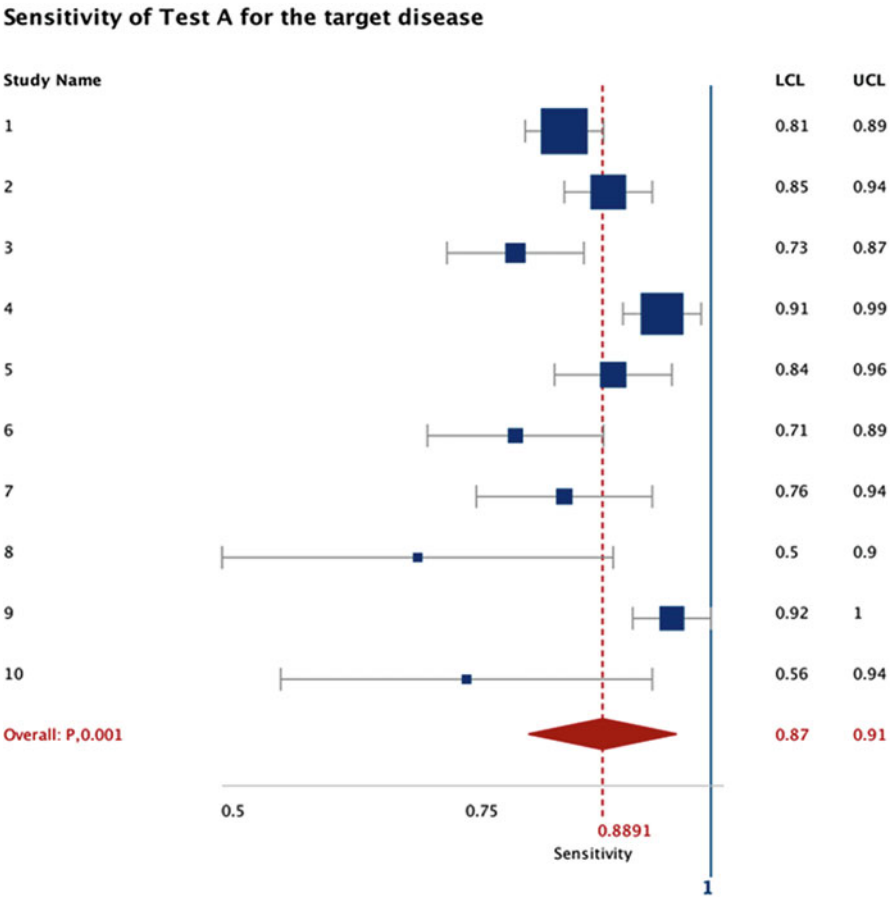


Fig. 12.2 Forest plot of sensitivity for Example 12.1. The *blue squares* are the sensitivity of each diagnostic test. The size of the squares reflects the sample size of the study. The 95% confidence intervals are also shown. The *red rhombus* reflects the summary sensitivity

While separate pooling is easy to perform and is the most commonly reported pooled statistic measure in meta-analysis, it has a major flaw: separate pooling ignores the fact that sensitivity and specificity are related, i.e., there is always a trade-off between sensitivity and specificity. This is part of the phenomenon known as threshold effect.

Many dichotomous tests (i.e., with positive and negative results) are actually quantitative measures that were dichotomized based on a numerical threshold. This threshold is usually defined using a ROC curve (Chap. 2). Changing the threshold value then can lead to changes to the sensitivity and specificity of the test: if the threshold is lowered, the sensitivity increases and specificity decreases. An opposite effect is observed when the threshold is increased. When studies in a systematic review use different threshold values, it can be said that a "threshold effect" exists. In fact, part of the heterogeneity and lack of correlation between different studies may be due to the threshold effect.

For this reason, SROC especially a variant known as hierarchical SROC (HSROC) is preferred for calculation of the summary measures. It is recommended that separate pooling approach is used for diagnostic odds ratio instead of sensitivity and specificity.

Summary Receiver Operating Characteristics Plot

When the possibility of threshold effect exists, it is better to use summary receiver operating characteristics plot and curve to summarize the data of the systematic review. The SROC allows us to visually inspect for threshold effect, and it can also help us to visualize the overall correlation of studies and the trade-off between sensitivity and specificity.

As in regular ROC plot, the X-axis represents the specificity (1 - false positive rate) and the Y-axis represents sensitivity (true positive rate). The actual SROC curve can be obtained using different statistical models with Moses-Littenberg approach and hierarchical approach being more common.

In the Moses-Littenberg approach, the logits of sensitivity and false positive rate of each study are calculated:

$$\begin{aligned}\text{Difference of logits } (D) &= \text{logit}(\text{sensitivity}) - \text{logit}(\text{false positive rate}) \\ &= \ln \frac{\text{True Positive Rate}}{1 - \text{True Positive Rate}} - \ln \frac{\text{False Positive Rate}}{1 - \text{False Positive Rate}}\end{aligned} \tag{12.7}$$

$$\begin{aligned}\text{Sum of logits } (S) &= \text{logit}(\text{sensitivity}) + \text{logit}(\text{false positive rate}) \\ &= \ln \frac{\text{True Positive Rate}}{1 - \text{True Positive Rate}} + \ln \frac{\text{False Positive Rate}}{1 - \text{False Positive Rate}}\end{aligned} \tag{12.8}$$

Sum of logits tends to increase as the overall proportion of positive test results increase; due to this the sum of logits can be used as a proxy for the test threshold.

The next step is to fit a linear regression line to these values using ordinary least-squares approach (see Chap. 4) (with difference of logits being the dependent and the sum of logits being the predictor):

$$D = \beta_0 + \beta_1 S + \varepsilon \tag{12.9}$$

The next step is to calculate the sensitivity for different levels of specificity (i.e., calculate the expected sensitivity):

$$\text{Expected Sensitivity} = \frac{1}{1 + \frac{1}{e^{\beta_0/(1-\beta_1)}} \left(\frac{\text{False Positive Rate}}{1-\text{False Positive Rate}}\right)^{\frac{1+\beta_1}{1-\beta_1}}} \tag{12.10}$$

Different false positive rates (e.g., 0.1, 0.2, 0.3, etc.) are fed into the formula to calculate the corresponding expected sensitivity. The next step is to draw the curve using the expected sensitivities and their corresponding false positive rates.

The problem with this approach is that it tends to underestimate the accuracy. For this reason, hierarchical SROC models are preferred. The explanation of these models is beyond the scope of this book, but those interested can read the following article: "Wang et al. Hierarchical models for ROC curve summary measures: Design and analysis of multi-reader, multi-modality studies of medical tests. Statistics in medicine. 2008 Jan 30;27(2):243–56.[1]."

Testing for Heterogeneity

If the results between different studies vary, we should always ask ourselves "why do the results vary?" Is the variation purely due to random variation? Is it due to the threshold effect? Or is there some bias or some unexplained variation present? Thus, as part of the meta-analysis, it is important to perform a "test for heterogeneity" to understand whether there is variation and whether it is random or nonrandom.

Testing for heterogeneity is either done using the "Cochran's Q statistics" or the "Higgins I^2 statistics." We have previously explained the principles of Q statistics in Chaps. 5 and 9. The Q statistics is a chi-squared measure that assesses whether the proportions of true positive, true negative, false positive, and false negative are the same across multiple groups (here multiple studies). The critical levels for Q statistics are determined based on the alpha level and degrees of freedom (which are number of studies minus 1). Generally, if the p-value for Q statistics is less than 0.1, then there is significant heterogeneity between the studies.

The Higgins I^2 statistics is calculated using the Q statistics:

$$\frac{I^2 = Q - DF}{Q \times 100\%} \tag{12.11}$$

where Q is calculated from the Cochran Q test and DF is the number of studies minus one. If the I^2 measure is more than 50%, then substantial heterogeneity exists with values of between 75% and 100% showing considerable heterogeneity.

To judge the heterogeneity, both the I^2 and the p-value from the Q test are needed: the p-value shows whether there is nonrandom heterogeneity, and the Higgins I^2 shows the degree of heterogeneity [6–12].

Publication Bias

Unfortunately, it has been shown that studies that have a positive finding are far more likely to be published than studies that have a neutral or negative finding. This has caused a significant bias in the literature. Thus, an important part of systematic review is to assess "publication bias."

The visual assessment of publication bias is done using a "funnel plot." In this study the diagnostic odds ratio of the study is shown on the X-axis, and the standard error (or precision) is shown on the Y-axis. Furthermore, "Egger's test" or the "Begg test" are used to assess whether there is significant publication bias by assessing the asymmetry of the plot.

Egger's test fits a linear regression line to a normalized odds ratio (odds ratio divided by its standard error) against the precision (inverse of the standard error) using ordinary linear regression:

$$\text{Standard Normal Deviate} = \frac{\text{Odds Ratio}}{\text{Standard Error of Odds Ratio}} \tag{12.12}$$

$$\text{Precision} = \frac{1}{\text{Standard Error of Odds Ratio}} \tag{12.13}$$

$$\text{Standard Normal Deviate} = \beta_0 + \beta_1 \,\text{Precision} \tag{12.14}$$

When fitting a line, both the intercept (β_0) and the slope (β_1) will have a confidence interval associated with them.

The interpretation of the Egger's test is based on the intercept of the fitted line. If there is no publication bias, then the intercept of the fitted line (β_0) will be equal to zero (the 95% confidence interval of the β_0 includes 0). To determine whether the intercept equals zero or not, we can run a one-sample t-test. The value of the t-test can be calculated by dividing the intercept by its standard error (with the degrees of freedom being the number of studies minus 2). If the p-value is smaller than 0.1, then we can deduce that a significant publication bias exists [9].

Example 12.2

Q: Table 12.7 shows the results of a systematic review. Derive the funnel plot, and based on the Egger's test determine whether a significant publication bias exists.

A: The funnel plot for the example is shown in Fig. 12.3.

Table 12.7 Results of the systematic review for Example 12.2

Study name	Diagnostic odds ratio	Standard error
Study 1	2.20	0.25
Study 2	1.80	0.21
Study 3	1.90	0.27
Study 4	2.05	0.14
Study 5	0.05	0.20
Study 6	−0.60	0.21
Study 7	2.00	0.22
Study 8	1.80	0.21
Study 9	0.40	0.22
Study 10	2.10	0.16
Study 11	−0.40	0.21
Study 12	−0.50	0.20

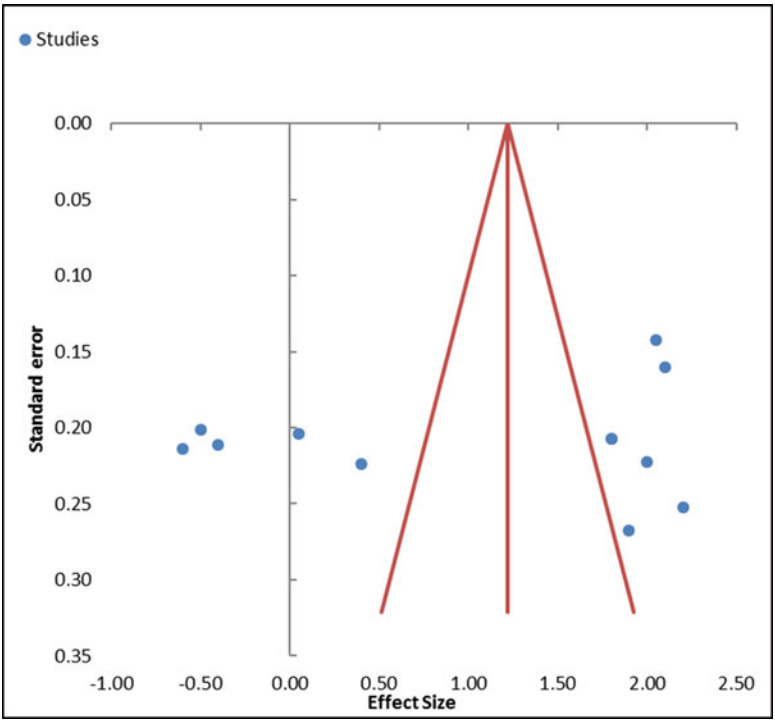


Fig. 12.3 Funnel plot for Example 12.2

The fitted Egger's regression line has a slope of 2.95 and intercept of −9.08. The standard error of the intercept is 9.606 with the 95% confidence interval being [-28.292 to 10.124]. Since the confidence interval includes 0, then we can say that no significant publication bias exists. The calculated p-value will be 0.344 again confirming that no significant publication bias exists.

Summary

In this chapter, we discussed the process of critical appraisal of evidence and briefly covered the topic of systematic review and meta-analysis. It is very important that pathologists don't accept the studies at their face value; many diagnostic studies have a degree of bias that can significantly distort their outcomes. As a result, every pathologist needs to know how to evaluate the study for quality and possible sources of bias. Furthermore, as part of the decision-making process, the pathologist should be able to interpret systematic reviews and meta-analyses; these studies can potentially be more informative than primary studies.

References

1. Wang F, et al. Hierarchical models for ROC curve summary measures: Design and analysis of multi-reader, multi-modality studies of medical tests. Stat Med. 2008;27(2):243–56.
2. Mallett S, Halligan S, Thompson M, Collins GS, Altman DG. Interpreting diagnostic accuracy studies for patient care. BMJ. 2012;345(jul021):e3999.
3. Kim KW, Lee J, Choi SH, Huh J, Park SH. Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: a practical review for clinical researchers-part I. General guidance and tips. Korean J Radiol. 2015;16(6):1175.
4. Whiting PF. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529.
5. Leeflang MMG. Systematic reviews of diagnostic test accuracy. Ann Intern Med. 2008;149 (12):889.
6. van Rhee H, Suurmond R, Hak T. User manual for Meta-Essentials: Workbooks for meta-analyses (Version 1.0).
7. Ghersi D, Berlin J, Askie L. Cochrane prospective meta-analysis Methods Group. COCHRANE METHODS 2011;35.
8. Campbell JM, Klugar M, Ding S, Carmody DP, Hakonsen SJ, Jadotte YT, White S, Munn Z. Diagnostic test accuracy. Int J Evid Based Healthc. 2015;13(3):154–62
9. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315(7109):629–34.
10. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009;6(7):e1000097.
11. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart LA. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev. 2015;4(1):1.
12. Lee J, Kim KW, Choi SH, Huh J, Park SH. Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: a practical review for clinical researchers-Part II. statistical methods of meta-analysis. Korean J Radiol. 2015;16(6):1188.

Designing Diagnostic Studies

13

Introduction

Research and investigation are important components of the practice of pathology. Pathologists are mainly involved in two types of clinical research: diagnostic or prognostic. Diagnostic research is conducted to improve diagnostic procedures and tests with the aim of improving diagnostic accuracy. Prognostic research mainly aims to identify and quantify factors that dictate the prognosis in patients [1].

Research conducted in pathology follows two general design strategies; the research is either descriptive where usually a series of patients or cases are chosen or pathologic or laboratory characteristics are measured in these patients. Alternatively, the research can be analytic where the relationship between two factors (e.g., two tests) is quantified. Analytic research can either be experimental or observational. Observational analytic research usually involves comparing a test or diagnostic procedure between a case group and a control group. Analytic research in diagnostic accuracy studies involves comparing an index test with a comparator (the so-called gold standard) to establish the accuracy of the index test. Observational analytic research in diagnostic medicine is performed as cohorts or case-control studies. Analytical studies can also be experimental; the individuals are randomized to two groups with one receiving an intervention (or test) and the other receiving an alternative intervention. This randomized trial design in diagnostic medicine is usually limited to studies of effectiveness (including cost-effectiveness).

In this chapter, we will explain diagnostic research design.

Diagnostic Research Design

Diagnostic research like all other forms of research requires a sound design to limit the bias and ensure that the study can achieve its intended goal. The design should be appropriate for the objective of the study. The research design should address

A. Momeni et al., *Introduction to Statistical Methods in Pathology*,
DOI 10.1007/978-3-319-60543-2_13

Table 13.1 Levels of diagnostic research

Level of research	Type of research	Study objectives
1	Technical accuracy and feasibility research	Technical validity, precision, cost, proof of concept
2	Diagnostic accuracy research	Sensitivity, specificity, predictive values, likelihood ratio
3	Diagnostic decision research	Changes in diagnosis, misdiagnosis, clinician decision-making impact
4	Therapeutic choice research	Changes in treatment choices or treatment practice by clinicians
5	Patient outcome research	Changes in patient-related outcomes (e.g., survival, quality of life, remission, disease control, etc.)
6	Population impact research	Cost-effectiveness, disease burden, public health measures

questions such as rationale, objectives, methodology and design, sample size, data collection, bias minimization strategies, data analysis, cost, and ethical considerations. The design process should be documented in a "research protocol."

There are several important decisions that should be made before the process of research design is begun. The first decision is about the objective and level of the diagnostic study. The research should clarify whether the research goal is to perform test research or diagnostic research; test research refers to the research that is performed to determine the technical accuracy and precision of a test. Test research is essentially a validation study where the technical characteristics of a test are determined. Diagnostic research, on the other hand, aims to determine the applicability of the test for diagnosis of a condition and determine elements of diagnostic accuracy (e.g., diagnostic sensitivity and specificity). Levels of diagnostic research are shown in Table 13.1.

For diagnostic research, it is important to know where, in the clinical diagnostic pathway, the test will be utilized (see Chap. 2). The test will either replace an existing test (replacement) in the pathway or it will be a screening tool upstream of the pathway (triage test), or it will be a new test in the pathway that can either subclassify a diagnostic condition or lead to a more accurate diagnosis (add-on test) (Fig. 13.1).

We are required to answer three broad questions for every new test in clinical pathways: What is the diagnostic accuracy of the test? How will patient outcomes be affected by using this test? How cost-effective is the test? Different study designs answer these questions for different locations of the pathway, for example, diagnostic accuracy testing for a triage test (emphasis on sensitivity) is different from diagnostic accuracy testing for an add-on test (emphasis on specificity).

In diagnostic accuracy studies, the aim is to show that either the new test has a better sensitivity or specificity compared to the existing test or that it has comparable sensitivity and specificity, but it costs less (or is safer).

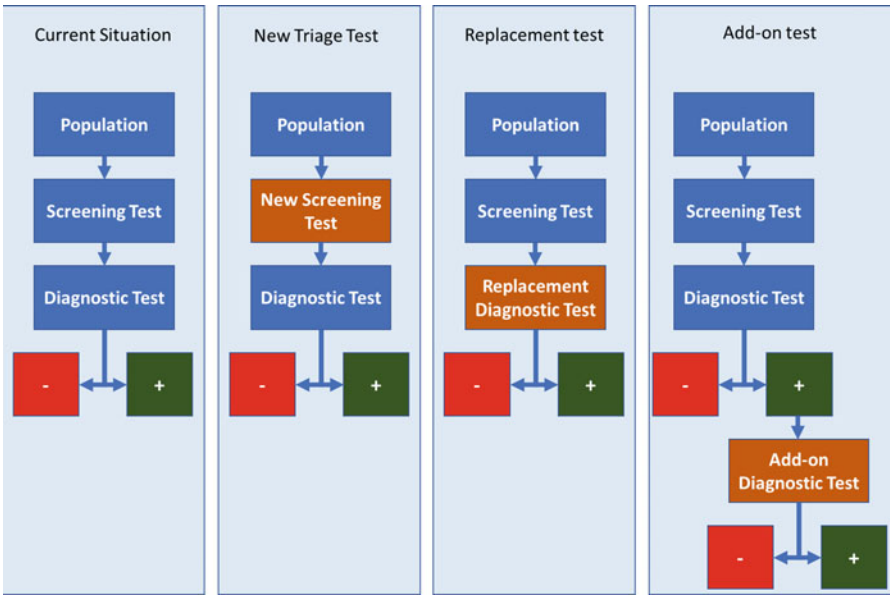


Fig. 13.1 Intended purpose of a test determines the objective of the diagnostic research

It is important to note that improved diagnostic accuracy does not necessarily impact patient outcomes, or, when it impacts patient outcomes, it may cause more harm than good. Often when the relation between the diagnostic accuracy and patient outcomes is not clear, there is a need to conduct a randomized trial to determine the impact of the new test on patient outcomes.

In randomized trial design, a random group from the target population is tested with the new test (case group), and a random group is tested with the reference standard test (control group). The patients will then receive treatment based on the outcomes of their respective tests, and the patient outcomes are compared between the two groups.

If previous randomized trials have shown clear benefit in treating patients affected by a condition (i.e., evidence for improved patient outcomes with treatment exists), then, for a test that diagnoses that condition, testing diagnostic accuracy suffices, and there is no need to determine patient outcomes or directly assess cost-effectiveness. In other words, there is no need to conduct a randomized trial to determine the benefit of the test for patients especially if the new test has clear positive attributes (such as safety, cost, etc.). In these situations, a diagnostic accuracy study, where the index test is compared with the reference test to determine test accuracy, is conducted. For example, cytology evaluation of fine needle aspiration of a pancreatic lesion is far safer (and more cost-effective) than an intraoperative biopsy and frozen section diagnosis, and, since, based on randomized trials, clear treatment guidelines exist for neoplastic lesions identified in the

pancreas, then determining the diagnostic accuracy of fine needle aspiration suffices for establishing its role in the diagnostic decision-making pathway [2].

Phases in Clinical Diagnostic Studies

Following technical research that is conducted to develop a new test and define its technical characteristics in a controlled laboratory environment, the next step is to study the test in real patients and assess diagnostic accuracy, applicability, safety, and clinical outcome of the new test. This is achieved through clinical diagnostic studies.

Clinical trials for new interventions have four phases. The first phase trials have low complexity and cost, and, as the researchers move to the next phases of trial, the complexity and cost increase. The reason for a phase-by-phase approach to clinical trials is to limit the risk to the pharmaceutical companies and establish safety and utility characteristics of the intervention before moving on to a large-scale trial which requires considerable investment and time.

While trial phases are much more common for interventions and drugs, a similar concept also exists for pathology and laboratory medicine. In commercializing a new diagnostic test, a good strategy would be to perform the necessary research in phases in order to limit the possible risks and add structure to the process of diagnostic research.

The phase I diagnostic research is performed to establish the normal range of results in a healthy group of individuals. The sample size for phase I research should be randomly drawn and be large enough to account for possible confounders or interactions such as age, gender, race, etc. Phase I studies are usually cross-sectional observational studies with random sampling of a normal population.

Phase II studies are either case-control or cohort studies and aim to establish the diagnostic accuracy of a test. In this phase, comparisons are made between a group of individuals with the target condition and a group of healthy individuals. Phase II studies aim to determine accuracy measures such as sensitivity, specificity, and predictive values. They are also used to determine cutoffs (using perhaps ROC curves) for quantitative tests to distinguish diseased and healthy states.

Phase IIA studies establish the diagnostic accuracy of the test by comparing the diseased and healthy individuals. In phase IIA studies, the disease status of the participants is known (i.e., case-control study), or a reference standard is used to establish the disease status of the participants. Comparisons are made either between the healthy and diseased group, or they are made between the results of the index test and the comparator.

Phase IIB studies determine whether there is a correlation between a quantitative test result and severity of the disease. Many tests are only useful to dichotomize individuals to healthy and diseased states and as such have single cutoff values. But there are other tests that can predict the severity of the disease and either have no cutoff (i.e., direct interpretation of quantitative result is needed) or have multiple cutoffs corresponding to different severity levels of the disease. For example,

creatinine level is a quantitative measure that relates to the severity of the renal disease with higher values indicating lower glomerular filtration rate. Phase IIB studies aim to establish and quantify the correlation of test result and severity and ideally fit a regression model that can predict the severity based on the test.

Phase IIC studies usually are prospective studies (cohorts) that are conducted to establish the predictive value of the test in a group of individuals with unknown disease status. These studies consist of exploratory cohorts where the test is performed in a randomly selected group of individuals whose disease status is unknown (or the researchers are blinded to it) and then either following the patients to determine the presence or incidence of the disease or performing a reference standard test to establish the disease status. The findings of an exploratory cohort need to be confirmed by running a validation cohort where the test is repeated in another group of individuals.

Phase III studies aim to evaluate the clinical impact of a new test, namely, the benefit and harm to the patient. Phase III studies are randomized trials where the individuals are randomized to receive either the index test or the comparator, and the treatment decision depends on the results of the tests. The patient outcomes are then compared between the two groups. In cases where a clear clinical benefit exists for better diagnosis (or a less costly or a safer test), sometimes phase III studies are not needed. Randomized diagnostic trials are difficult to design, implement, and analyze. These trials are often very costly and may face many complications. For example, sample size calculations require adjustment for discordance rates because the only results that matter are those that arise due to discordance between the index test and comparator.

Most stand-alone tests that are to be marketed as platforms for diagnosis of a disease or condition are required by regulatory bodies to have completed phase III diagnostic trials.

Phase IV studies are performed to evaluate the actual impact of introducing a test in clinical practice; these studies are based on evidence emerging from the different practice settings using the new test and are usually conducted following a systematic review of phase II and phase III studies. Phase IV studies are also concerned with changes in the testing conditions and sample matrix, to evaluate factors such as storage and handling of specimens on the test results.

Table 13.2 shows the different phase of diagnostic research. For the sake of comparison, we include, alongside, the phases for therapeutic (also termed treatment or intervention) research. Both are required steps by the US Federal Drug Administration (FDA).

In the next sections of this chapter, we will focus on phase II diagnostic accuracy studies.

Table 13.2 Phases of clinical diagnostic research compared with intervention research

Phase	Diagnostic research	Therapeutic (intervention) research
Phase I	Studies on normal population to determine normal range of the test	Safety evaluation of the intervention/drug in a small group (including determination of safe dosage range)
Phase II	Studies to establish the diagnostic accuracy of the test, usually using a comparator test	Evaluation of effectiveness and determination of therapeutic dosage by extending phase I trials to a larger group of individuals
Phase III	Clinical impact studies; randomized trials to determine the clinical effect of the test	Randomized clinical trials with large sample size to evaluate the efficacy and side effects and compare the intervention/drug with other treatment options
Phase IV	Follow-up studies to determine the actual clinical effect of the test in different practice settings	Long-term follow-up studies performed after release of a drug to monitor for long-term adverse effects and benefits

Diagnostic Accuracy Studies

Perhaps the most important step in diagnostic research is to determine the diagnostic accuracy of the test; diagnostic accuracy is the major determinant of the adoption of a new test. Furthermore, pathologists in academic or community practice settings are more likely to take part in diagnostic accuracy studies than other phases of clinical diagnostic research. The aim of these studies is to determine how well a test can discriminate between diseased and healthy states. This requires the evaluation of the index test versus a reference standard (comparator). The results of the test can be dichotomous requiring analysis using a 2×2 contingency table and extraction of diagnostic accuracy metrics such as sensitivity, specificity, positive predictive value, negative predictive value, and likelihood ratio, or alternatively, the results of the two tests can be evaluated using a quantitative scale requiring correlation and regression analysis.

The common element in these types of study is that the index test and comparator should be compared for all patients with selective performance of the comparator leading to verification bias. Diagnostic accuracy studies are either conducted on known patients (case-control design) or unknown patients (cohort design). The case-control design is relatively simple to conduct, yet it suffers from bias: in fact, it has been reported that case-control design can overestimate the diagnostic accuracy by two to three folds [3–6].

Before we introduce some of the designs used in diagnostic accuracy testing, we must explain "index test" and "reference standard."

Index Test

Index test is the target approach or methodology that is being studied in a diagnostic accuracy study. Index test is not necessarily a single diagnostic test: It can also be a combination of diagnostic tests. For example, in anatomic pathology, a panel of immunohistochemical antibodies can be an index test.

A diagnostic accuracy study does not necessarily aim to evaluate all accuracy aspects of the index test; based on the clinical application of the test, the diagnostic study can focus on one or more accuracy measures. For example, an index test that is going to be used for triage/screening will require an evaluation of the test's sensitivity.

The index test results should also be interpretable in light of its probable clinical application. For example, if it is to be used as a rule-in/rule-out test for a disease, then its results should be dichotomized (see Chap. 2). As tests are often quantitative measures, pilot studies (usually observational cross-sectional studies) are needed to determine the cut-off values for the test.

Reference Standards

"Ground truth" refers to the true disease status of an individual. For many diseases, it is often very difficult to establish the ground truth with a 100% certainty because either no test exists that can have 100% accuracy or the perfect test is not practical to perform. Thus, proxy measures and benchmarks are often used to establish the disease status of individuals.

Reference standard is a benchmark that is available under reasonable conditions; this means that the reference standard is not necessarily the perfect test, but it is the closest test to the ground truth that can be practically performed. For example, cardiac troponins can be considered as the reference standard for myocardial infarction because the ground truth can only be truly revealed by histopathologic examination of the heart. Commonly, the reference standard is a well-established testing methodology that has been thoroughly tested, and its accuracy and reliability have been confirmed. While ideally a reference standard should have a very high sensitivity and specificity, in choosing a reference standard, we should consider what aspect of accuracy we are interested in: If the index test is to be used for triage/screening, then the reference standard should have a very high sensitivity. For add-on tests, specificity of the reference standard is often more important.

Reference standard and index test should be independent with no residual measurement effects, i.e., measuring one test should not affect the results of the other test. For example, in comparison of digital rectal examination with serum PSA level, if the rectal examination precedes the PSA level determination, it will cause an increase in the serum PSA level. The issue of independence is particularly problematic when the index test is an improved version of the reference test.

The problem with reference standards is the existence of "reference standard bias": if the results of the reference standard test do not mirror the ground truth, then

comparing the index test with the reference standard introduces a bias in interpretation of the index test.

If the reference standard is perfect, then the naïve accuracy estimate which is the calculated sensitivity and specificity of the index test (based on the results of the diagnostic accuracy experiment) equals to the true accuracy of the test. However, if the reference test is imperfect, then the naïve estimates of accuracy are always underestimates of the true values. In fact, the index test may have a better accuracy than the reference standard, and we will fail to show it.

In dealing with imperfect reference standards, we can use several solutions. The simplest solution would be to calculate the naïve estimates of accuracy knowing and reporting the imperfection of the reference standard. Such qualification allows the readers of the report to know the possibility of reference standard bias.

Alternatively, adjustments of the naïve accuracy estimates can be made to account for the reference standard bias. This requires advanced statistical modeling and knowledge of parameters such as true sensitivity and specificity of the reference standard test.

Another option is to perform a randomized patient outcome study instead of the diagnostic accuracy tests and compare the patient outcomes using the index test versus using the reference standard. This is by far the best solution since it circumnavigates the issue of accuracy and focuses on patient outcome (which, in reality, is the end goal of all testing).

Yet another option would be to measure concordance (agreement) between the two tests instead of accuracy measures such as sensitivity and specificity. In this approach, the degree of agreement between the tests can be stated using statistical measures such as Cohen's Kappa coefficient (see Chap. 5). For continuous measures, however, usual measures of agreement such as ordinary least squares method are not applicable since both the reference standard and index test have variability and noise (in the least squares model, it is assumed that one of the values is the true value, i.e., the value of the reference method, and does not have any error, or in other words, it is fixed as discussed in Chap. 4). Thus, other statistical regression models should be used to account for variability in results of both the index test and the reference standard.

In ordinary least squares regression, the slope of the fitted line changes if we interchange the axes, i.e., if we plot the index method values on the *X*-axis and the reference method values on the *Y*-axis. In this case, the least squares best-fit correlation line is used to produce a regression line with the lowest overall error (S in Eq. 4.18 in Chap. 4).

Alternatively, another method, called major axis regression, can be used for each estimated regression function relating the values of one variable (e.g., reference test results) to the values of another variable (e.g., index test results). In this method, a loss function is defined and calculated. The loss function is the product of Y-distance and X-distance of the observations from the fitted line, and this product is minimized (unlike least square models where only one distance (usually the vertical distance) is minimized), i.e., the best-fit line is found by minimization of the sum of areas of the triangles formed between the observation value and the line.

For example, a linear regression model can be shown as:

$$Y = \beta_0 + \beta_1 X \tag{13.1}$$

For ordinary least squares model, the loss function (L) depends on residuals and is calculated using:

$$L = \sum_{i=1}^{n} (Y - (\beta_0 + \beta_1 X))^2 \tag{13.2}$$

(see Eq. 4.4 in Chap. 4). In these models, the best-fit line is found by minimizing the loss function. However, in major axis regression, the loss function for the linear model will be:

$$L = \sum_{i=1}^{n} \frac{(Y - (\beta_0 + \beta_1 X))^2}{1 + {\beta_1}^2} \tag{13.3}$$

Another approach would be to use "Bland-Altman plots" (also known as Tukey mean-difference plots). This method allows a visual inspection of the correlation of the two measures (index test results and reference standard results) throughout their range of measurement. This measure is different from agreement: It shows that as one variable changes, the other changes as well; however, it does not necessarily mean that the two variables are measuring the same thing. For example, in comparing PSA level with prostate volume (measured from MRI), there is high correlation between the measures. However, in measuring mean corpuscular hemoglobin concentration using colorimetric versus light scatter methodology, the high correlation between results also implies agreement (since they are both measuring the same thing).

In Bland-Altman graphs, the difference between the values of the index test and the reference standard is the Y-axis, and the average of the two values is the X-axis. A solid line is drawn at the mean of the difference between the two values with two dotted lines representing the upper and lower bounds of the 95% confidence interval. The points representing $S(X, Y) = \left(\frac{S1+S2}{2}, S1 - S2\right)$ are then plotted on the graph. For high correlation, all the points should fluctuate around the mean with variations from the mean being contained within the 95% confidence interval limit (Fig. 13.2) [7–9].

Examples of Diagnostic Accuracy Study Designs

In this section, we will introduce some of the designs used in diagnostic accuracy studies and briefly discuss the advantages and disadvantages of each one.

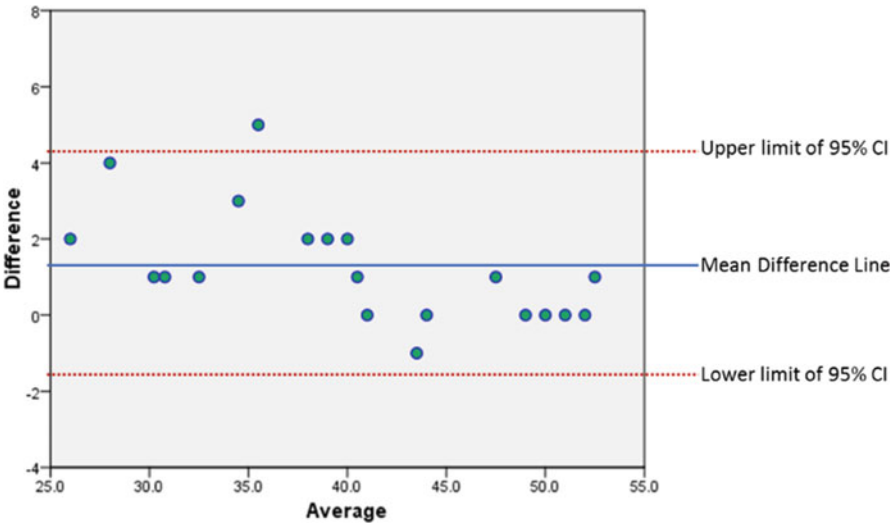


Fig. 13.2 Bland-Altman plot

Observational Studies

The simplest type of diagnostic studies is observational studies such as case series where a test or feature is measured in a series of patients/samples with known disease condition. These studies can be prospective or retrospective and cross-sectional. Observational studies are single-arm studies and have no control group. Thus, no inference can be made about the discrimination power of the test. However, these studies can be used as pilot studies before diagnostic accuracy studies to determine the potential of the test as a diagnostic tool. Findings that will support further evaluation of the test include consistent results in a considerable proportion of the cases.

Case series is one of the observational studies that are commonly used in pathology. This study design has serious associated bias. For example, since the status of the patients is known, the researcher might (intentionally or unintentionally) opt to choose cases where a favorable test result is more likely (selection bias). Furthermore, since many case series lack random sampling, they usually have spectrum bias which makes the comparison of the case series studies difficult.

In observational studies, attempts in comparing with other observational studies should be avoided. Also the researchers should avoid making causal or diagnostic inferences based on the results of a case series. The data from an observational study, at best, should be treated as a prevalence (for cross-sectional, retrospective designs) or an incidence (for prospective designs) measure.

Paired Comparative Accuracy Studies

These groups of studies have a study arm and a control arm; the test is measured in both groups and the results are compared. Case-control studies and cohorts are some of the diagnostic accuracy study designs that follow a paired comparative approach.

Case-Control Studies

Case-control studies are retrospective studies where the test or feature is measured in a group of patients with a known diagnosis of the target condition, and these results are compared with the results from a group of individuals without the target condition. The group of healthy individuals is chosen so that they match the baseline characteristics of the diseased group. Usually an attempt is made to match the groups for known interferences or confounders (e.g., age, gender, etc.). Ideally, the groups should be similar in every aspect but the presence of disease. However, in reality, controlling for every possible source of interference or confounding factor might prove impossible. Case-control studies, just as with observational studies, suffer from selection bias.

The outcome of choice in case-control studies is the odds ratio: These studies can provide us with general notion of the diagnostic odds ratio. The odds ratio (defined in Chap. 2) is a measure of association, and a high odds ratio essentially means that the test results are different for the two groups. This does not allow us to make a causal inference about the test, but it can still show the possible discrimination power of the test (subject to bias).

These studies are generally easy to conduct and interpret and are best suited for diseases with low prevalence. In rare diseases, it has been shown that the odds ratio is a good approximation of the relative risk and can be used for causal inference.

Cohort Study

In cohort studies, the status of the patient is unknown and only becomes known through follow-up. In these studies, the individuals undergo a test, and the incidence of the disease (or other outcomes) between the patients who tested positive and those who tested negative is compared. In diagnostic accuracy studies, this means that the individuals are tested with the index test (with positive test results considered as exposed group and the negative test results considered as unexposed group), and then they are tested with the reference standard to check for their disease status. To control for selection bias, cohort studies must follow a "consecutive sampling," i.e., all the patients who meet the eligibility criteria (inclusion/exclusion criteria) within the time frame of the study should be included in the study. The researchers should not interfere in case selection in any way as it will produce selection bias.

Furthermore, in cohort designs, the reference standard should always follow the index test to avoid possible selection bias based on the results. Also, all patients tested with the index test should also be tested with the reference standard (otherwise we are introducing verification bias into the results).

In cases where the aim is to use the index test to replace a current test, both the index test and the current test are measured in all cases, and all the results are compared with the reference standard results. In some cases, it is acceptable to compare the accuracy of the index test with the accuracy data of the current test extracted from a comprehensive systematic review; however, this approach should be avoided if possible.

Cohort studies are the standard design for diagnostic accuracy studies.

Randomized Comparative Accuracy Studies

Sometimes performing both the index test and the reference standard in the same patients is difficult. For example, if one of the tests is invasive, then performing both tests on the same patient can be unethical. Other situations which make paired comparative studies impractical or problematic include when the two tests interfere with each other or when the aim is to assess the clinical impact of the test. Thus, we have to use a randomized comparative design.

In randomized comparative designs, individuals are randomized either to the index test group or to the reference standard group. In each arm, all the clinical decisions are made based on the results of the test for the patients in that arm, and a clinical outcome related to that disease is measured and compared between the two arms. In this design, the diagnostic accuracies of the tests are not directly compared, but their impact on clinical outcomes is compared.

As we mentioned earlier, these designs are often expensive and difficult to conduct. However, they provide valuable information that is often needed for inclusion of a test in diagnostic pathways [9, 10].

Sample Size Calculations

One of the major concerns in design of diagnostic studies is to ensure that the sample size is adequate. Small sample size increases the possibility of type II error (i.e., reduces the power of the study) which means that the probability of finding a true effect decreases. On the other hand, as the sample size increases, the cost and complexity of the study increase. Thus, finding the optimal sample size is one of the priorities of clinical researchers.

For case-control studies where the true status of the individuals is known, we can use the following equation for sample size calculation when the aim is to evaluate either sensitivity or specificity:

$$n = \frac{Z_{\frac{\alpha}{2}}^2 \widehat{P}\left(1 - \widehat{P}\right)}{d^2} \tag{13.4}$$

where n is the sample size, $Z^2_{\frac{\alpha}{2}}$ is the squared value of the Z-score corresponding to our desired confidence interval of the sensitivity (or specificity) (usually, 95% confidence interval is desired; thus, α will be 0.05), $\widehat{P}$ is the expected sensitivity (or specificity) (based on prior studies or the expectations of the clinicians), and d^2 is the squared maximal margin of error estimate. The maximal margin of error estimate is usually set as 0.05 or 0.02. However, the margin of error can be calculated if the standard deviation of the sensitivity (or specificity) is known in the population:

$$d = Z_{\frac{\alpha}{2}} \times \text{Standard Deviation} \tag{13.5}$$

If the true disease status is not known in the cases (e.g., in cohort studies) but the prevalence of the disease in a given population is known, then the sample size calculated in Eq. (13.4) should be adjusted for the prevalence. If the study aims to examine sensitivity, the adjustment will be:

$$n_{\text{adjusted}} = \frac{n}{\text{prevalence}} \tag{13.6}$$

If the study aims to examine specificity, the adjustment will be:

$$n_{\text{adjusted}} = \frac{n}{1 - \text{prevalence}} \tag{13.7}$$

If the study aims to examine both specificity and sensitivity, then the larger of the adjusted sample sizes is used.

As you can see, for rare diseases where prevalence is low, the sample size increases considerably; thus, a case-control study is preferred (because the sample size remains the same irrespective of prevalence).

Other formulas can be used if the aim of the diagnostic accuracy test is to determine other accuracy measures (e.g., likelihood ratio) [11].

Reporting of Diagnostic Accuracy Studies

Reporting of diagnostic accuracy studies should follow a standard format so that the readers can extract the relevant information needed for evaluation of the diagnostic test. The standard reporting also allows the readers to identify the possible sources of bias in the study and obtain essential information about the design of the study. To this end, a standardized format was proposed by the EQUATOR network called the "Standards for Reporting of Diagnostic Accuracy Studies" (STARD). STARD provides a checklist that includes all the necessary information that should be reported in diagnostic accuracy studies. Some investigative journals require the researchers to submit the STARD checklist along with their manuscript to the journal. Based on STARD, the information that should be reported include

elements such as study design, eligibility criteria, enrollment flowchart, test methods (including detailed descriptions of the tests to allow replication), reference standard chosen and rationale for the choice, and statistical analysis methods. STARD also requires the researchers to include the 2×2 contingency table of the index test versus the reference standard in their manuscript. The complete STARD checklist can be found at www.stard-statement.org [12].

Summary

In this chapter, we introduced the concept of the diagnostic accuracy study and discussed the considerations of such studies. We also provided a brief introduction to different designs used for diagnostic accuracy studies.

References

1. di Ruffano LF, Hyde CJ, McCaffery KJ, Bossuyt PM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. BMJ. 2012;344:e686.
2. Thompson M, Van den Bruel A. Diagnostic Tests Toolkit. Chichester: Wiley; 2011.
3. Glasser SP. Research methodology for studies of diagnostic tests. In: Essentials of clinical research. Springer International Publishing: Netherlands; 2014. pp 313–26.
4. Bossuyt PM, Irwig L, Craig J, Glasziou P. Diagnosis: comparative accuracy: assessing new tests against existing diagnostic pathways. Br Med J. 2006;6:1089–92.
5. Meier K. Statistical guidance on reporting results from studies evaluating diagnostic tests. Comment. U.S. Department of Health and Human Services, Food and Drug Administration: USA; 2007.
6. Jinyuan LI, Wan TA, Guanqin CH, Yin LU, Changyong FE. Correlation and agreement: overview and clarification of competing concepts and measures. Shanghai Arc Psychiatry. 2016;28(2):115.
7. Cardoso JR, Pereira LM, Iversen MD, Ramos AL. What is gold standard and what is ground truth? Dental Press J Orthod. 2014;19(5):27–30.
8. Hawkins DM, Garrett JA, Stephenson B. Some issues in resolution of diagnostic tests using an imperfect gold standard. Stat Med. 2001;20(13):1987–2001.
9. Chang SM, Matchar DB, Smetana GW, Umscheid CA. Methods guide for medical test reviews. Agency for Healthcare Research and Quality: Rockville; 2012.
10. Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case–control and two-gate designs in diagnostic accuracy studies. Clin Chem. 2005;51(8):1335–41.
11. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. J Clin Epidemiol. 1991;44(8):763–70.
12. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, De Vet HC. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Clin Chem Lab Med. 2003;41(1):68–73.

Statistical Concepts in Modern Pathology Practice

14

Introduction

The main purpose of pathology is to diagnose diseases. Pathologists classify patients into affected and unaffected classes, and they also further subclassify the affected individuals based on additional information. Pathologists try to give patients accurate diagnoses based on the most probable disease. To do this, they gather information (from clinical to morphological to molecular), summarize the information, and compare the information against a list of possible entities. However, as technology has advanced, the amount of information available has exponentially increased as well. Current technology allows us to fully sequence the genome of a tumor and evaluate the transcriptome and even protein expression of the tumor. This exponential growth in information makes it next to impossible for humans to be able to use most of this information in a meaningful way. Thus, computer algorithms and statistical models have been developed that can deal with this so-called big data.

Clustering performed for understanding is an attempt at classification of the data. As pathologists, we have long used data to cluster diseases into different classes. For example, we use the morphologic characteristics of a tissue to classify it as benign or malignant and even further subclassify it into meaningful disease categories.

However, clustering can be performed to summarize the data as well; the data may be multidimensional or have many variables and components which make its understanding or analysis very difficult. In these situations, cluster analysis allows us to summarize the data for clusters and use these summary cluster prototypes for analysis. For example, direct comparison of RNA expression profiles of tumors may prove very difficult, but through clustering we can identify prototypes of the cancers in the data and compare their expression profile [1].

While in-depth explanation of the concepts behind clustering and classifying algorithms and models is beyond this current book, in this chapter we will briefly

A. Momeni et al., *Introduction to Statistical Methods in Pathology*,
DOI 10.1007/978-3-319-60543-2_14

introduce two of the most common clustering approaches used in classifying and clustering patients.

Clustering Algorithms

The aim of clustering is to use the information contained in the data to divide that data into clusters or groups. This clustering is performed either to find natural clusters that are present in the data or to summarize the data to facilitate its analysis and understanding. In general, the success of clustering (the ability to clearly distinguish clusters) depends on the homogeneity of the data within each cluster and difference between the clusters. A major issue with clustering is defining the cluster: what constitutes a cluster, what are its boundaries, and how many clusters are we looking for in a dataset? If clustering is based on previous classifications, the answer to these questions is simple. For example, in pathology a successful clustering might be a binary distinction between a benign and malignant state. However, the answers to these questions are not always straightforward. In recent years in pathology, we have observed an ongoing attempt at further subclassifying tumors into different entities: while this has sometimes been successful and useful, in other times it has led to increasing confusion because of the lack of distinction between the entities.

Classification algorithms can be "supervised" or "unsupervised". In supervised clustering, a classification model is developed using data with known class labels; this is very similar to the regression models we discussed in Chap. 7. Clustering algorithms are mainly unsupervised classifiers, in that they use unlabeled data and use its structure to divide it into clusters (classes). For example, if we have the data from a group of normal individuals and a group of patients and use a binary logistic regression to differentiate between the diseased and normal individuals, then we are using supervised classification. However, if we have the mutational data from patients with a tumor and use hierarchical clustering to identify different mutation patterns in the tumor, then we are using unsupervised classification.

Here we will discuss two clustering approaches: *K*-means clustering and hierarchical clustering.

K-Means Clustering

"*K*-means clustering" in simple terms clusters the data into "*K*" number of classes. To do this, the clustering algorithm defines a prototype (also known as a centroid) for each cluster: this prototype is the mean of a group of points and assigns points to a group based on their proximity to the centroid. The number of classes is defined by the users based on their clustering needs. For example, a pathologist looking to cluster patients into high risk and low risk will set the K as 2. The general concept for a *K*-means clustering is as follows:

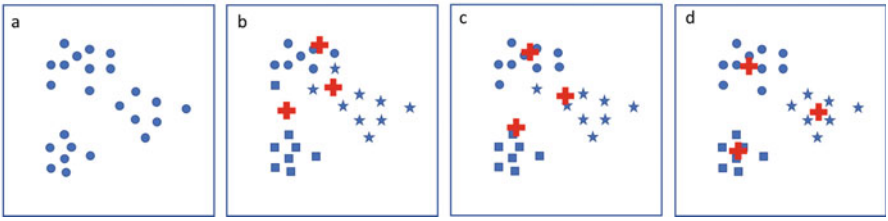


Fig. 14.1 *K*-means clustering is shown in this figure. In this case, we want to find three clusters in the data. The initial data points are shown in panel (**a**); the algorithm chooses three random initial centroids (*red crosses*) and assigns the points to the groups (**b**); the centroids are adjusted and the points are reassigned (**c**); the process continues until the centroids are fixed (**d**)

1. "*K*" initial centroids are chosen.
2. The points are assigned to the closest centroid, with all the points assigned to the centroid considered as a cluster.
3. The centroid (mean of the group) is updated based on the points assigned to cluster.
4. The points are again assigned to the closest centroid, with all the points assigned to the centroid considered as clusters.
5. The process is repeated until no points change clusters (in other words, the centroids don't change anymore).

We have shown this process in Fig. 14.1.

The assigning of the points to a cluster is based on their proximity to the centroid of that cluster. Different methods exist for defining the proximity. One approach is based on Euclidean (L^2) distance. This distance is the straight-line distance between two points in Euclidean space. Euclidean distance is easily determined in two dimensions. For example, if we are clustering the data based on only two continuous variables X and Y, then the distance between a centroid ($p(x_1,y_1)$) and a data point ($p(x_2,y_2)$) can be defined as:

$$\text{Euclidean distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{14.1}$$

A very common application of *K*-means clustering in pathology is in automatized hematology analyzers and flow cytometry where two parameters for each cell are measured: forward light scatter and side light scatter. Using *K*-means clustering, we can cluster the cells into our defined classes; the assignment will be based on the two-dimensional Euclidean distance of the points from centroids (in hematology analyzers the initial centroids are often predefined).

The Euclidean distance can be measured in N dimensions as well. For two N-dimensional points p and q, (with dimension p_1,q_1 through p_n,q_n) the Euclidean distance can be calculated by:

$$\text{Euclidean distance} = \sqrt{\left((q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots + (q_n - p_n)^2\right.} \quad (14.2)$$

The main component affecting the clusters in *K*-means clustering is where the initial centroids are placed. Thus, the choice of where the centroids are is often important; one approach is to choose multiple random initial centroids where the clustering is performed in different iterations, and for each iteration a different set of initial points is chosen, and the "sum of squared error" (SSE) is measured to determine the best initial starting centroids and the best cluster definitions.

For each point, the Euclidean distance of the point to the centroid of its cluster is measured, and the value is squared (in order to place more emphasis on outliers that are far from the cluster centroid). Then the squared Euclidean distances of all points are summed up: this sum is called the sum of squared error. Models with smaller sum of squared error are usually preferred since they show that the points are more concentrated around their centroids (less scatter).

$$\text{SSE} = \sum_{i=1}^{K} \sum_{x \in c_i} \left(\text{Euclidean distance}(c_i, x)^2 \right. \quad (14.3)$$

where K is the number of clusters, and c_i is the centroid for cluster i.

Based on this, it can be shown that the best centroid for a cluster is the mean of the points in that cluster:

$$c_i = \frac{1}{m_i} \sum_{x \in c_i} x \quad (14.4)$$

where m_i is the number of points in the cluster i.

Other approaches used for point assignment in *K*-means clustering include squared Euclidean distance, cosine proximity, Manhattan distance, and Bregman divergence. Each of these is suited for different applications. The explanation of these approaches is beyond the scope of this book.

Hierarchical Clustering

Another common clustering approach is "hierarchical clustering" where a nested approach to clustering is followed to build a hierarchy of clusters. We commonly encounter hierarchical clustering in medical taxonomy and in the way diseases are structured. For example, the diseases are first clustered by their site and then clustered by their pathologic mechanism (inflammatory, neoplastic, etc.), and the classification continues until we reach a single disease entity. The advantage of this approach to *K*-means clustering is that we don't need to define the number of clusters.

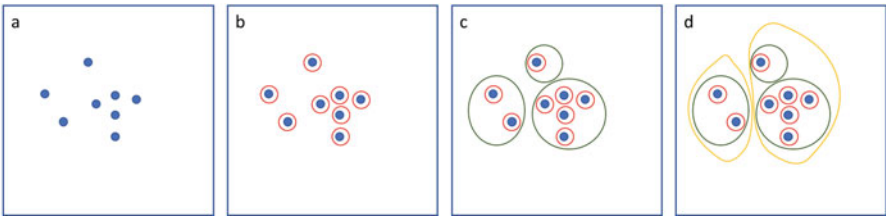


Fig. 14.2 A set of points (**a**) are clustered using agglomerative hierarchical clustering. The clustering begins with each point being a cluster (**b**), then the nearest clusters are merged leading to three clusters (**c**), and the nearest clusters are again merged (**d**) leaving only two clusters. The process continues until one cluster containing all the points is obtained

The hierarchical clusters can be built using a top-down divisive approach or an agglomerative bottom-up approach. Here, we will focus on the agglomerative approach where each point is considered as its own cluster and then clusters are merged as we move further up the hierarchy. The process continues until only one cluster remains (Fig. 14.2).

The merging of clusters is performed based on cluster proximity. At each level, the two closest clusters are identified using a proximity matrix and they are merged. The proximity matrix is then updated to account for the merged clusters, and the process is repeated until only one cluster remains.

There are different ways for defining the proximity between clusters. One approach is to use the minimum distance (e.g., Euclidean distance) between the points of clusters, and the two clusters with the smallest minimum distance are merged together, and the process is repeated until only one cluster remains. Alternatively, the maximum distance between the points of the clusters or the average distance can also be used. We can also use cluster centroids and measure proximity between the centroids of the clusters and merge the clusters with closest centroids.

The results of hierarchical clustering can be shown in a treelike graph known as a "dendrogram" (Fig. 14.3).

Example 14.1

Two-dimensional data for five points are provided in Table 14.1. Using a minimal Euclidean distance method, apply hierarchical clustering to these points.

The proximity matrix of the Euclidean distances between these points is shown in Table 14.2.

In the first step, we have to find the smallest minimum distance between the points. Looking at Table 14.2 we can see that Point 1 and Point 5 have a Euclidean distance of 3.6. Point 2 and Point 4 also have a Euclidean distance of 3.6. Thus, the first step is merging of Point 1 and Point 5 into one cluster and Point 2 and Point 4 into another cluster. At this stage, we are left with three clusters: Cluster 1 (Point 1 and Point 5), Cluster 2 (Point 3), and Cluster 4 (Point 2 and Point 4).

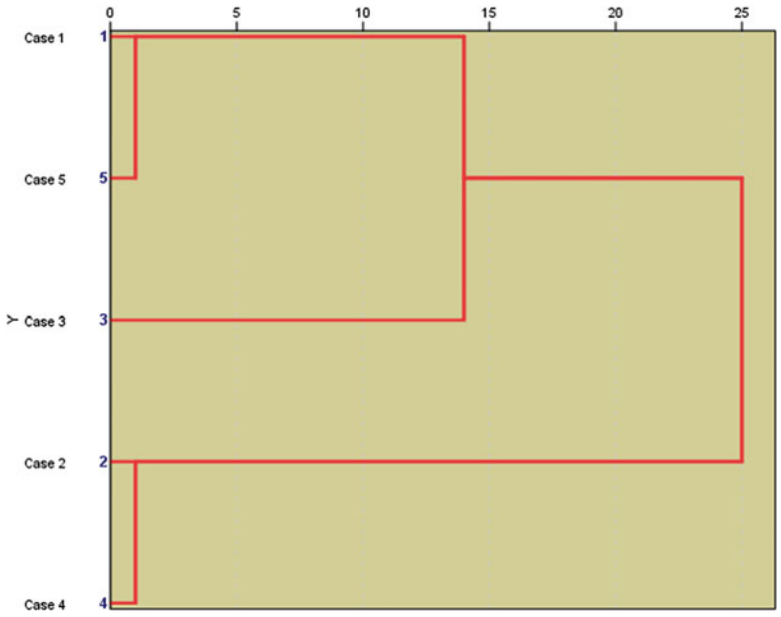


Fig. 14.3 Dendrogram for Example 14.1

Table 14.1 Two-dimensional data for five points

X	*Y*
4	9
6	4
2	3
8	1
1	7

Table 14.2 Proximity matrix for Example 14.1

	Point 1	Point 2	Point 3	Point 4	Point 5
Point 1	0	5.385165	6.324555	8.944272	3.605551
Point 2	5.385165	0	4.123106	3.605551	5.830952
Point 3	6.324555	4.123106	0	6.324555	4.123106
Point 4	8.944272	3.605551	6.324555	0	9.219544
Point 5	3.605551	5.830952	4.123106	9.219544	0

The next step is to merge these clusters. We can see that the minimum Euclidean distance between the points of the clusters is equal between Cluster 1 and Cluster 2 (where the distance between Point 3 and Point 2 is 4.12) and Cluster 2 and Cluster 3 (where the distance between Point 3 and Point 5 is 4.12). As we have a tie, we will randomly merge either Cluster 1 with Cluster 2 or Cluster 3 with Cluster 2.

Now we have reached a point where we have two clusters: Cluster 1 (which includes Point 1, Point 5, and Point 3) and Cluster 2 (Point 2 and Point 4). Now we can merge the two clusters to form our final inclusive cluster that includes all points.

Figure 14.3 is the dendrogram for this example.

The main implication of hierarchical approach is that as you move down the hierarchy, the homogeneity of the clusters increases, i.e., points belonging to the same bottom-level cluster are much more similar to each other than points in a top-level cluster.

The main application of hierarchical clustering is in molecular pathology where tumors are hierarchically clustered based on their mutational pattern or RNA expression pattern. This has been helpful in identifying different subgroups of tumors which have distinct mutational patterns. These results in turn have been quite useful in understanding the progression of tumors as well as paving the way for possible targeted therapies [2–4].

Summary

In this chapter, we discussed clustering and introduced two of the main clustering approaches employed in diagnostic medicine: *K*-means clustering and hierarchical clustering. These algorithms are often computationally intensive (especially when dealing with multidimensional data) requiring advanced statistical software and high computing capacity. For a complete discussion of these approaches as well as other clustering approaches, we recommend the interested readers to read *Statistical Modeling and Machine Learning for Molecular Biology* by Alan Moses, CRC Press 2017.

References

1. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. Cambridge: Cambridge university press; 2008.
2. Kononenko I. Inductive and Bayesian learning in medical diagnosis. Applied Artificial Intelligence an International Journal. 1993;7(4):317–37.
3. Nithya N, Duraiswamy K, Gomathy P. A survey on clustering techniques in medical diagnosis. International Journal of Computer Science Trends and Technology (IJCST). 2013;1(2):17–23.
4. Tan PN. Introduction to data mining. Pearson Education India: India; 2006.

Appendix A

Z-scores table: the values in the cells show the area under the curve to the left of *Z*

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5	0.504	0.508	0.512	0.516	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.591	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.648	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.67	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.695	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.719	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.758	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.791	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.834	0.8365	0.8389
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.877	0.879	0.881	0.883
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.898	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.937	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.975	0.9756	0.9761	0.9767
2	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.983	0.9834	0.9838	0.9842	0.9846	0.985	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.989
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.992	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.994	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.996	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.997	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.998	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

(continued)

A. Momeni et al., *Introduction to Statistical Methods in Pathology*,
DOI 10.1007/978-3-319-60543-2

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
3	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.999	0.999
3.1	0.999	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Appendix B

Chi-square distribution table: the values in the cell are chi-square critical values corresponding to the alpha levels and degrees of freedom

DF	$\alpha =$ 0.995	$\alpha =$ 0.990	$\alpha =$ 0.975	$\alpha =$ 0.950	$\alpha =$ 0.900	$\alpha =$ 0.100	$\alpha =$ 0.050	$\alpha =$ 0.025	$\alpha =$ 0.010	$\alpha =$ 0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993

(continued)

A. Momeni et al., *Introduction to Statistical Methods in Pathology*,
DOI 10.1007/978-3-319-60543-2

DF	$\alpha =$ 0.995	$\alpha =$ 0.990	$\alpha =$ 0.975	$\alpha =$ 0.950	$\alpha =$ 0.900	$\alpha =$ 0.100	$\alpha =$ 0.050	$\alpha =$ 0.025	$\alpha =$ 0.010	$\alpha =$ 0.005
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Appendix C

T-score distribution table: the values in the cell are t-score critical values corresponding to the alpha levels and degrees of freedom

	One-tailed α										
DF	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690

(continued)

A. Momeni et al., *Introduction to Statistical Methods in Pathology*,
DOI 10.1007/978-3-319-60543-2

	One-tailed α										
DF	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
Two-tailed α	1	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001

Appendix D

F-score distribution table: the values in the cell are f-score critical values corresponding to the degrees of freedom with alpha level of 0.05

DF	1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.39	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.97	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.10	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.97	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.56	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.33	3.47	3.07	2.84	2.69	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.38	2.32	2.28
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.26
25	4.24	3.39	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19

(continued)

A. Momeni et al., *Introduction to Statistical Methods in Pathology*,
DOI 10.1007/978-3-319-60543-2

DF	1	2	3	4	5	6	7	8	9	10
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.17
31	4.16	3.31	2.91	2.68	2.52	2.41	2.32	2.26	2.20	2.15
32	4.15	3.30	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14
33	4.14	3.29	2.89	2.66	2.50	2.39	2.30	2.24	2.18	2.13
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11
36	4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.15	2.11
37	4.11	3.25	2.86	2.63	2.47	2.36	2.27	2.20	2.15	2.10
38	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09
39	4.09	3.24	2.85	2.61	2.46	2.34	2.26	2.19	2.13	2.08
40	4.09	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
41	4.08	3.23	2.83	2.60	2.44	2.33	2.24	2.17	2.12	2.07
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.07
43	4.07	3.21	2.82	2.59	2.43	2.32	2.23	2.16	2.11	2.06
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.15	2.09	2.04
47	4.05	3.20	2.80	2.57	2.41	2.30	2.21	2.14	2.09	2.04
48	4.04	3.19	2.80	2.57	2.41	2.30	2.21	2.14	2.08	2.04
49	4.04	3.19	2.79	2.56	2.40	2.29	2.20	2.13	2.08	2.03
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03

F-score distribution table: the values in the cell are f-score critical values corresponding to the degrees of freedom with alpha level of 0.01

DF	1	2	3	4	5	6	7	8	9	10
1	4052.19	4999.52	5403.34	5624.62	5763.65	5858.97	5928.33	5981.10	6022.50	6055.85
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
6	13.75	10.93	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.52	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.90	3.81
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69

(continued)

DF	1	2	3	4	5	6	7	8	9	10
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	8.29	6.01	5.09	4.58	4.25	4.02	3.84	3.71	3.60	3.51
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
23	7.88	5.66	4.77	4.26	3.94	3.71	3.54	3.41	3.30	3.21
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
27	7.68	5.49	4.60	4.11	3.79	3.56	3.39	3.26	3.15	3.06
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
29	7.60	5.42	4.54	4.05	3.73	3.50	3.33	3.20	3.09	3.01
30	7.56	5.39	4.51	4.02	3.70	3.47	3.31	3.17	3.07	2.98
31	7.53	5.36	4.48	3.99	3.68	3.45	3.28	3.15	3.04	2.96
32	7.50	5.34	4.46	3.97	3.65	3.43	3.26	3.13	3.02	2.93
33	7.47	5.31	4.44	3.95	3.63	3.41	3.24	3.11	3.00	2.91
34	7.44	5.29	4.42	3.93	3.61	3.39	3.22	3.09	2.98	2.89
35	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88
36	7.40	5.25	4.38	3.89	3.57	3.35	3.18	3.05	2.95	2.86
37	7.37	5.23	4.36	3.87	3.56	3.33	3.17	3.04	2.93	2.84
38	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.92	2.83
39	7.33	5.19	4.33	3.84	3.53	3.31	3.14	3.01	2.90	2.81
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
41	7.30	5.16	4.30	3.82	3.50	3.28	3.11	2.98	2.88	2.79
42	7.28	5.15	4.29	3.80	3.49	3.27	3.10	2.97	2.86	2.78
43	7.26	5.14	4.27	3.79	3.48	3.25	3.09	2.96	2.85	2.76
44	7.25	5.12	4.26	3.78	3.47	3.24	3.08	2.95	2.84	2.75
45	7.23	5.11	4.25	3.77	3.45	3.23	3.07	2.94	2.83	2.74
46	7.22	5.10	4.24	3.76	3.44	3.22	3.06	2.93	2.82	2.73
47	7.21	5.09	4.23	3.75	3.43	3.21	3.05	2.92	2.81	2.72
48	7.19	5.08	4.22	3.74	3.43	3.20	3.04	2.91	2.80	2.72
49	7.18	5.07	4.21	3.73	3.42	3.20	3.03	2.90	2.79	2.71
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.79	2.70

Index

A. Momeni et al., *Introduction to Statistical Methods in Pathology* ,
DOI 10.1007/978-3-319-60543-2