Introduction to Biostatistics

Antonella Iuliano and Monica Franzese, Institute for Applied Mathematics "Mauro Picone", Napoli, Italy

Introduction

Biostatistics is a branch of applied statistics with applications in many areas of biology including epidemiology, medical sciences, health sciences, educational research and environmental sciences. The principles and methods of statistics, which is the science that deals with the collection, classification, analysis, and interpretation of numerical data for the purpose of data description and decision making, are applied to the biological areas. The first application of statistics appeared during the seventeenth century in political science to describe the various aspects of the affairs of a government or state (hence the term "statistics"). At the same time, the development of probability theory, thanks to the contribution of many mathematicians, such as Blaise Pascal (1623–1662), Pier Fermat (1601–1665), Jacques Bernoulli (1654–1705) and others, has provided the basis for the modern statistics. However, the first scientist to introduce biostatistics concepts was the astronomer and mathematician Adolphe Quetelet (1796–1874), who in his work combined the theory and practical methods of statistics in biological, medical and sociological applications. Later, Francis Galton (1822–1911) tried to solve the problem of heredity on the basis of Darwin's genetic theories with the statistics. In particular, Galton's contribution to biology was the application of statistics to the analysis of biological variation, using correlation and regression techniques. For this reason, he has been defined as the father of biostatistics and his methodology has become the basis for the use of statistics in biology. Karl Person (1860–1906) continued in the tradition of Galton's theory contributing significantly to the field of biometrics, meteorology, theories of social Darwinism and eugenics. The dominant figure in biostatistics during the twentieth century was Ronald Fischer (1890–1962), who used mathematics to combine Mendelian genetics and natural selection. In particular, he developed the analysis of variance (ANOVA) to analyze large amount of biological data. Today, statistics is an active field whose applications touch many aspects of biology and medicine. In particular, we can distinguish two types of statistical approaches, which aim to provide precise conclusions and significant information from a set of data collected during a biological experiment. The first approach is called descriptive statistics and it is used to analyze a collection of data without assuming any underlying structure for such data (Spriestersbach *et al.*, 2009), while the second one, called inferential statistics, works on the basis of a given structure for the observed data and involves hypothesis testing to draw conclusions about a population when only a part of the data is observed (Altman and Krzywinski, 2017; Gardner and Altman, 1986). In fact, when a biologist conducts an experiment, he must make sure that the possible conclusions are statistically significant. In addition, the necessity to perform long and laborious arithmetic computations, as part of the statistical analysis of data with the use of computers, has contributed to improve the quality of the data and the interpretation of the results. In fact, the large amount of available statistical software programs, such as the R Project for Statistical Computing, SAS, SPSS and others, have further revolutionized statistical computing in the field of bioinformatics and computational biology.

The aim of this work is to provide statistical concepts that help biologists to correctly prepare experiments, verify conclusions and properly interpret results. We first introduce several descriptive statistical techniques for organizing and summarizing data, and then we discuss some procedures to infer the population parameters using the data contained in a sample that has been drawn from that population. Finally, an illustrative example is analyzed to give a general understanding of the nature and relevance of biostatistics in clinical research.

Statistical Analysis

The descriptive analysis is the starting point in any applied research providing a numerical summary of the collected data (Spriestersbach *et al.*, 2009). The main steps of a scientific investigation are: collection of data, organization and visualization of data, calculation of descriptive statistics, and interpretation of statistics (see **Fig. 1**). Such data are usually available from one or more sources, such as routinely kept records (hospital medical records or hospital accounting records), surveys (questionnaires and interviews), experiments (treatment decision to investigate the effects of the assigned therapy, treatment and control groups), clinical trials (to test efficacy or toxicity of a treatment with respect to control group) and external sources (published reports or data banks). The set of all elements in a data is known as statistical population, while a sample consists of one or more observations extracted from the population. Each member (or element) of the data under investigated is called statistical unit. A statistical variable is each aspect or characteristic of the statistical unit that is considered for the study. A statistical variable can be qualitative or quantitative, depending on whether their nature is countable or not. Quantitative variables can be characterized as discrete or continuous. Examples of discrete variables are the number of daily admissions to a general hospital or the number of decayed, while examples of continuous variables are the diastolic blood pressure, the heart rate, the heights of the adult males or the ages of patients. On the other hand, qualitative or categorical variables involve observations that can be grouped into categories. In particular, these data can be statistically divided into three groups: nominal (when exist a natural ordering among the

 doi:10.1016/B978-0-12-809633-8.20353-1

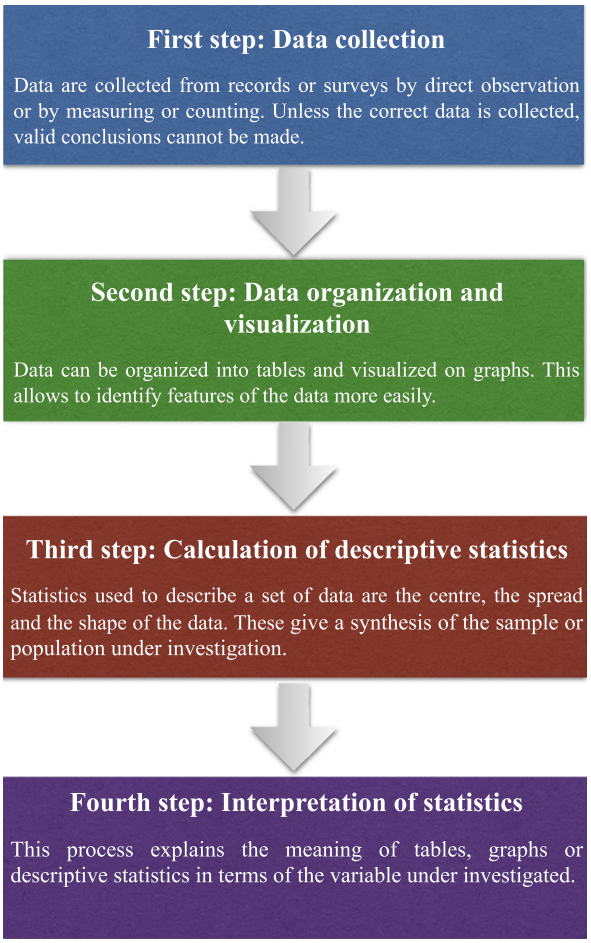

Fig. 1 Main steps of a statistical investigation.

categories), binary or dichotomous (when there are only two possible levels), and ordinal data (when there is a natural order among the categories). Examples of categorical variables involve the sex (male or female), the genotype (AA, Aa, or aa), or the ankle condition (normal, sprained, torn ligament, or broken).

Sometimes in statistics, we use the word measurement (or measurement scale) to categorize and quantify variables. In particular, a measurement is the assignment of numbers to objects or events according to a set of rules. Different measurement scales are distinguished on the relationships assumed to exist between object having different scale values. The most important types of measurement scales are: nominal, ordinal, interval, and ratio. The lowest measurement scale is the nominal scale. It consists of naming observations or classifying them into various mutually exclusive and collectively exhaustive categories. The practice of using numbers to distinguish among the various medical diagnoses constitutes measurement on a nominal scale. When the observations are ranked according to some criterion (from lowest to highest observations), they are said to be measured on an ordinal scale. Convalescing patients may be characterized as unimproved, improved, and much improved. The interval scale is a more sophisticated scale than the nominal and ordinal scale. Here, not only is it possible to order measurements, but also to known the distance between any two measurements. We know that the difference between a measurement of 20 and a measurement of 30 is equal to the difference between measurements of 30 and 40. The ability to do this implies the use of a unit distance and a zero point, both of which are arbitrary. A clear example is the Fahrenheit or Celsius scale. The unit of measurement is the degree of temperature, and the point of comparison is the arbitrarily chosen "zero degrees", which does not indicate a lack of heat. The interval scale unlike the nominal and ordinal scales is a truly quantitative scale. The highest level of measurement is the ratio scale. This scale is characterized by the fact that equality of ratios as well as equality of intervals may be determined. Fundamental to the ratio scale is a true zero point. The measurement of such familiar traits as height, weight, and length makes use of the ratio scale. Interval and ratio data are sometimes referred to as parametric and nominal and ordinal data are referred to as nonparametric. Parametric means that it meets certain requirements with respect to parameters of the population (normal or bell curve). Parametric data are analyzed using statistical techniques called parametric statistics. Nonparametric data are lacking those same parameters and are investigated by using non-parametric statistics.

Collection, Organization and Visualization of Data

The first step of explaining a biological or biomedical phenomenon is the collection of data under investigated (first step in Fig. 1). Then, the second step is the organization of observed data into tables or data matrix in order to visualize their distribution (second step in Fig. 1). Let N be the number of element or individuals in the population and let X be a random variable assuming the values x_i, for $i=1,2,\ldots,n$. We denote the number of individuals presenting the value (or characteristic) x_i as n_i. This number n_i is the absolute frequency of the observed value x_i, while the relative frequency fi of the observed value x_i is the proportion on the total population N of values presenting the value x_i. In symbols, the relative frequency is the ratio

$$f_i = \frac{n_i}{N} \tag{1}$$

for $i=1,\ldots,n$. Note that the sum of the absolute frequencies is equal to the total number of data N and the sum of the relative frequencies is equal to 1:

$$\sum_{i=1}^{n} n_i = N, \quad \sum_{i=1}^{n} f_{ii} = 1$$

The observed values x_i, the absolute and relative frequencies are usually organized in tables, called statistical tables or frequency distribution. These tables show the way in which the values of the variable are distributed among the specified characteristics. An example of a univariate statistical table is given in Table 1. When we have a large amount of information, data are grouped into class intervals. A common rule used to choose the number of intervals is to take no fewer than 5 intervals and no more than 15. The better way to decide how many class intervals to employ is to use the Sturges's formula given by $k=1+3.322\log_{10}(N)$, where k stands for thre number of class intervals and N is the number of values in the data set. The width of the class intervals is determined by dividing the range R by k. The range R is given by the difference between the smallest x_{min} and the largest observation x_{max} in the data set. The number of values, falling within each class interval, is the absolute frequency n_i, while the proportion of values falling within each class interval is the relative frequency fi (see Table 2). Starting from the definition of (absolute and relative) frequency distribution, it is possible to calculate the (absolute and relative) cumulative frequency (N_i and F_i, $i=1,\ldots,n$) distribution, which indicates the number of elements in the data set that lie above (or below) a particular value in a data set. For instance, see Tables 1 and 2. Usually, the information contained in these tables can be presented graphically under the form of histograms and cumulative frequency curves. A histogram is a graphical representation of the absolute or relative frequencies for each value of the characteristic or class intervals. It is commonly used for quantitative variables. A cumulative frequency curve is a plot of the number or percentage of individuals falling in or below each value of the characteristic or class intervals. Other types of graphical representations are the pie chart or the bar plot. The first graph is a circular chart divided into sectors, showing the relative magnitudes in frequencies or percentages. The second one, often used to display categorical data, is a

Table 1 Frequency distribution

Values of characteristics x_i	*Absolute frequency* n_i	*Relative frequency* f_i	*Cumulative absolute frequency* N_i	*Cumulative relative frequency* F_i
x_1	n_1	$f_1=\frac{n_1}{N}$	$n_1=N_1$	$f_1=F_1$
x_2	n_2	$f_2=\frac{n_2}{N}$	$n_1+n_2=N_2$	$f_1+f_2=F_2$
...	...	...	...	...
x_i	n_i	$f_i=\frac{n_i}{N}$	$n_1+n_2+\ldots+n_i=N_i$	$f_1+f_2+\ldots+f_i=F_i$
...	...	...	...	
x_k	n_k	$f_k=\frac{n_k}{N}$	$n_1+n_2+\ldots+n_k=N$	$f_1+f_2+\ldots+f_k=1$
Total	N	1		

Source: UF Health – UF Biostatistics. Available at: http://bolt.mph.ufl.edu/2012/08/02/learn-by-doing-exploring-a-dataset/.

Table 2 Frequency distribution based on class intervals. The symbol $-|$ means that only the superior limit is included into the class interval

Class intervals $x_i-\|x_{i+1}$	*Absolute frequency* n_i	*Relative frequency* f_i	*Cumulative absolute frequency* N_i	*Cumulative relative frequency* F_i
$x_1-\|x_2$	n_1	$f_1=\frac{n_1}{N}$	$n_1=N_1$	$f_1=F_1$
$x_2-\|x_3$	n_2	$f_2=\frac{n_2}{N}$	$n_1+n_2=N_2$	$f_1+f_2=F_2$
...	...	...	...	...
$x_{i-1}-\|x_i$	n_i	$f_i=\frac{n_i}{N}$	$n_1+n_2+\ldots+n_i=N_i$	$f_1+f_2+\ldots+f_i=F_i$
...	...	...	...	
$x_{n-1}-\|x_n$	n_n	$f_n=\frac{n_n}{N}$	$n_1+n_2+\ldots+n_n=N$	$f_1+f_2+\ldots+f_n=1$
Total	N	1		

Source: Reproduced from Daniel, W.W., Cross, C.L., 2013. Biostatistics: A Foundation for Analysis in the Health Sciences, tenth ed. John Wiley & Sons.

chart with rectangular bars with lengths proportional to the values they represent. They can be plotted vertically or horizontally. Histogram, pie chart and bar plot are graphs very useful for presenting data in a comprehensible way to a statistical and non-statistical audience.

Descriptive Measures

After the organization and visualization of data, several statistical functions can be computed to describe and summarize the set of data (third step in **Fig. 1**) and to interpret the obtained results (fourth step in **Fig. 1**). These functions are called descriptive measures or statistics. There are three general types of statistics: measures of central tendency (or location), measures of dispersion (or variability), and measures of symmetry (or shape). The measures of central tendency convey information about the average value of a data. The most commonly used measures of location are the mean, the median, and the mode (Manikandan, 2011a,b; Wilcox and Keselman, 2003). These descriptive measures are called location parameters, because they can be used to designate specific positions on the horizontal axis when the distribution of a variable is graphed. Let X be a random variable and let x be the observed values of the random variable X. The mean is obtained by adding all the values in a population or sample and dividing by the number of values that are observed. We use the Greek letter μ to stand for the population mean, while we use the symbol $\overline{x}$ to define the sample mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i, \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{N} x_i \tag{2}$$

The value N indicates the population size, the quantity n is the number of observed values in the sample. Similarly, we can compute the mean for (simple and grouped) frequency distributions. An alternative to the mean is the computation of the median. The median is a numerical value that divides the ordered set of values (from lowest to highest value) into two equal parts, such that the number of values equal to or greater than the median is equal to the number of values equal to or less than the median. If the number of values is odd, the median is the middle value of the ordered set of data. If the number of values is even, the median is the mean of the two middle values. In addition, when the median is close to the mean, then we use as statistics the mean, even if the median is usually the better choice. Similarly, we can compute the median for (simple and grouped) frequency distributions. Other location parameters include percentiles or quartiles. These descriptive measures divide the data set into four equal parts each containing 25% of the total observations. The 50th percentile Q_2 is the median. The 25th percentile is the first quartile Q_1. The 75th percentile is the third quartile Q_3. Finally, the mode of a variable is the value that occurs most frequently into the data. The mode may not exist, and even if it does, it may not be unique. This happens when the data set has two or more values of equal frequency, which is greater than the other values. The mode is usually used to describe a bimodal distribution. In a bimodal distribution, the taller peak is called the major mode and the shorter one is the minor mode. For continuous data, the mode is the midpoint of the interval with the highest rectangle in the histogram. If the data are grouped into class intervals, than the mode is defined in terms of class frequencies. The mode is used also for describing qualitative data. For example, suppose that the patients in a mental health clinic, during a given year, received one of the following diagnoses: mental retardation, organic brain syndrome, psychosis, neurosis, and personality disorder. The diagnosis, occurring most frequently in the group of patients, is called modal diagnosis. The computational formulas of location measures are shown in **Table 3**.

Dispersion measures describe the spread or variation present in a set of data. The most important statistics of variability are the range, the variance, the standard deviation and the coefficient of variation (Manikandan, 2011c). The range R is the difference between the largest and smallest value in a set of observations. The variance and the standard deviation are two very popular measures of dispersion.

The variance is defined as the average of the squared differences from the mean, while the standard deviation is a measure of how the data are spread out across the mean. We use the Greek letter σ^2 to indicate the population variance, while we use the symbol s^2 to define the sample variance:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (X_i - \mu)^2, \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{N} (x_i - \overline{x})^2 \tag{3}$$

The standard deviation is the square root of the variance. The more variation there is into the data, the larger is the standard deviation. The standard deviation is useful as a measure of variation within a given set of data. When two distributions are taken into account and their measures are expressed in different units, compare the two standard deviations may lead to false results. For example, for a certain population we wish to know whether serum cholesterol levels, measured in milligrams per 100 mL, are more variable than body weight, measured in pounds. Therefore, in this case, we use a measure of relative variation rather than absolute variation. Such measure is called coefficient of variation CV, which expresses the standard deviation as a percentage of the mean: it is a unit-free measure. The CV is small if the variation is small and it is unreliable if the mean is near zero. Hence, if we consider two groups, the one with less CV is said to be more consistent.

Another measure of dispersion is the interquartile range (IQR). It is the difference between the third and first quartiles, i.e., $IQR = Q3 - Q1$. A large IQR indicates a large amount of variability among the middle 50% of the relevant observations, and a small IQR indicates a small amount of variability among the relevant observations. A useful graph for summarize the information contained in a data set is the box-and-whisker plot. The construction of a box-and- whisker plot makes use of the quartiles of a

Table 3 Summary of the main descriptive statistics used to describe the basic features of the data in a study

Statistics	*For population*	*For sample*	*For simple frequency distribution*	*For class frequency distribution*
Mean	$\mu = \frac{\sum_{i=1}^{N} xi}{N}$ where N is the population size	$\overline{x} = \frac{\sum_{i=1}^{n} xi}{n}$ where n is the sample size	$\overline{x} = \frac{\sum_{i=1}^{n} xini}{n}$ where n_i is the ith absolute frequency	$\overline{x} = \frac{\sum_{i=1}^{n} cini}{n}$ where $c_i = \frac{x_{i+1}+x_i}{2}$ are the central values of each class, for $i=1,\ldots,n$
Median	Odd size: $M_e = \frac{N+1}{2}$ th ordered observation X_i	Odd size: $M_e = \frac{n+1}{2}$ th ordered observation x_i	$M_e = x_i$, such that, $F_i \geq 0.50$	$M_e \approx x_i + (x_{i+1} - x_i)\frac{0.5-F_{i-1}}{F_i - F_{i-1}}$, where $F_i \geq 0.50$
	Even size: $M_e = \frac{\left(\frac{N}{2}\right)+\left(\frac{N}{2}+1\right)}{2}$ th ordered observation X_i	Even size: $M_e = \frac{\left(\frac{n}{2}\right)+\left(\frac{n}{2}+1\right)}{2}$ th ordered observation x_i	F_i is the ith relative cumulative frequency	F_i is the ith relative cumulative frequency
Quartiles	$Q_1 = \frac{N+1}{4}$ th ordered observation X_i	$Q_1 = \frac{n+1}{4}$ th ordered observation x_i	$Q_1 = x_i$, such that, $F_i \geq 0.25$	$Q_1 \approx x_i + (x_{i+1} - x_i)\frac{0.25-F_{i-1}}{F_i - F_{i-1}}$, where $F_i \geq 0.25$
	$Q_2 = M_e$	$Q_2 = M_e$	$Q_2 = M_e$	$Q_2 = M_e$
	$Q_3 = \frac{3(N+1)}{4}$ th ordered observation X_i	$Q_3 = \frac{3(n+1)}{4}$ th ordered observation x_i	$Q_3 = x_i$, such that, $F_i \geq 0.75$ F_i is the ith relative cumulative frequency	$Q_3 \approx x_i + (x_{i+1} - x_i)\frac{0.75-F_{i-1}}{F_i - F_{i-1}}$, where $F_i \geq 0.75$
Mode	$M_o = x_i$ is the value that occurs most frequently in the data	$M_o = x_i$ is the value that occurs most frequently in the data	$M_o = \{x_i : n_i = max\}$, where n_i is the ith absolute frequency	$m_c = \{x_i - \lvert x_{i+1} : h_i = max\}$, where $h_i = \frac{n_i}{x_{i+1}-x_i}$ is the intensity class $M_o \approx \frac{x_i + x_{i+1}}{2}$
Range	$R = x_{max} - x_{min}$	$R = x_{max} - x_{min}$	$R = x_{max} - x_{min}$	$R = x_{max} - x_{min}$
Interquartile range	$IQR = Q_3 - Q_1$	$IQR = Q_3 - Q_1$	$IQR = Q_3 - Q_1$	$IQR = Q_3 - Q_1$
Variance	$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{N}$	$s^2 = \sum_{i=1}^{n} \frac{(x_i-\overline{x})^2}{n-1}$	$s^2 = \sum_{i=1}^{n} \frac{(x_i-\overline{x})^2 n_i}{n-1}$	$s^2 = \sum_{i=1}^{n} \frac{(c_i\overline{x})^2 n_i}{n-1}$
Deviazione standard	$\sigma = \sqrt{\sigma}$	$s = \sqrt{s}$	$s = \sqrt{s}$	$s = \sqrt{s}$
Correlation of variation	$CV = \frac{\sigma}{\mu} \times 100$	$CV = \frac{s}{x} \times 100$	$CV = \frac{s}{x} \times 100$	$CV = \frac{s}{x} \times 100$
Skewness	$\Gamma_1 = \frac{1}{N\sigma^3} \sum_{i=1}^{N} (x_i - \mu)^3$	$\gamma_1 = \frac{\sqrt{n}}{(n-1)\sqrt{n-1}s^3} \sum_{i=1}^{n} (x_i - \overline{x})^3$	$\gamma_1 = \frac{\sqrt{n}}{(n-1)\sqrt{n-1}s^3} \sum_{i=1}^{n} (x_i - \overline{x})^3 n_i$	$\gamma_1 = \frac{\sqrt{n}}{(n-1)\sqrt{n-1}s^3} \sum_{i=1}^{n} (c_i - \overline{x})^3 n_i$
Kurtosis	$\Gamma_2 = \frac{1}{N\sigma^4} \sum_{i=1}^{N} (x_i - \mu)^4$	$\gamma_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)\sigma^4} \sum_{i=1}^{n} (x_i - \overline{x})^4$	$\gamma_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)\sigma^4} \sum_{i=1}^{n} (x_i - \overline{x})^4 n_i$	$\gamma_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)\sigma^4} \sum_{i=1}^{n} (c_i - \overline{x})^4 n_i$

data set (see Fig. 2). An outlier is an observation whose value either exceeds the value of the third quartile by a magnitude greater than $1.5 \times (IQR)$ or is less than the value of the first quartile by a magnitude greater than $1.5 \times (IQR)$. Similarly, we can compute the (simple and grouped) dispersion measures for frequency distributions. The computational formulas of variability measures are shown in Table 3.

An attractive property of a data distribution occurs when the mean, median, and mode are all equal. The well-known "bell-shaped curve" is a graphical representation of a distribution for which the mean, median, and mode are equal among them. Much statistical inference is based on this distribution called normal distribution (see Fig. 3). Generally, a data distribution can be classified on the basis of their form (symmetric or asymmetric). A symmetric distribution is a type of distribution where the left side of the distribution mirrors the right side (see Fig. 3). When the left half and right half of the graph of a distribution are not mirror images of each other, the distribution is asymmetric. In this case, the distribution is said to be skewed. In other words,

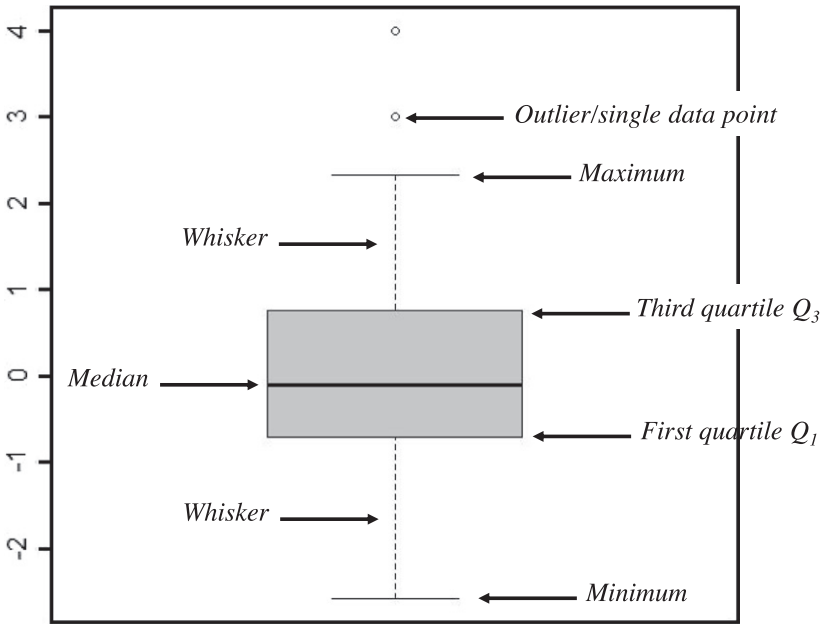

Fig. 2 Boxplot or box plot whisker diagram.

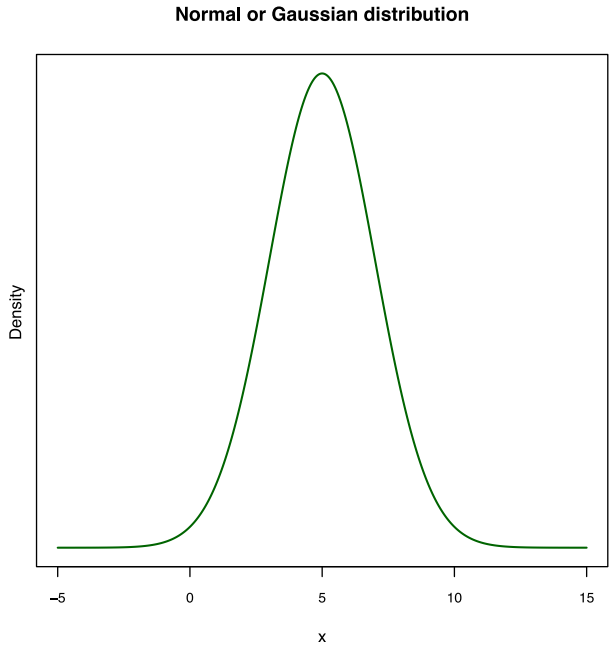

Fig. 3 Normal or Gaussian distribution with mean $\mu = 5$ and standard deviation $\sigma = 2$.

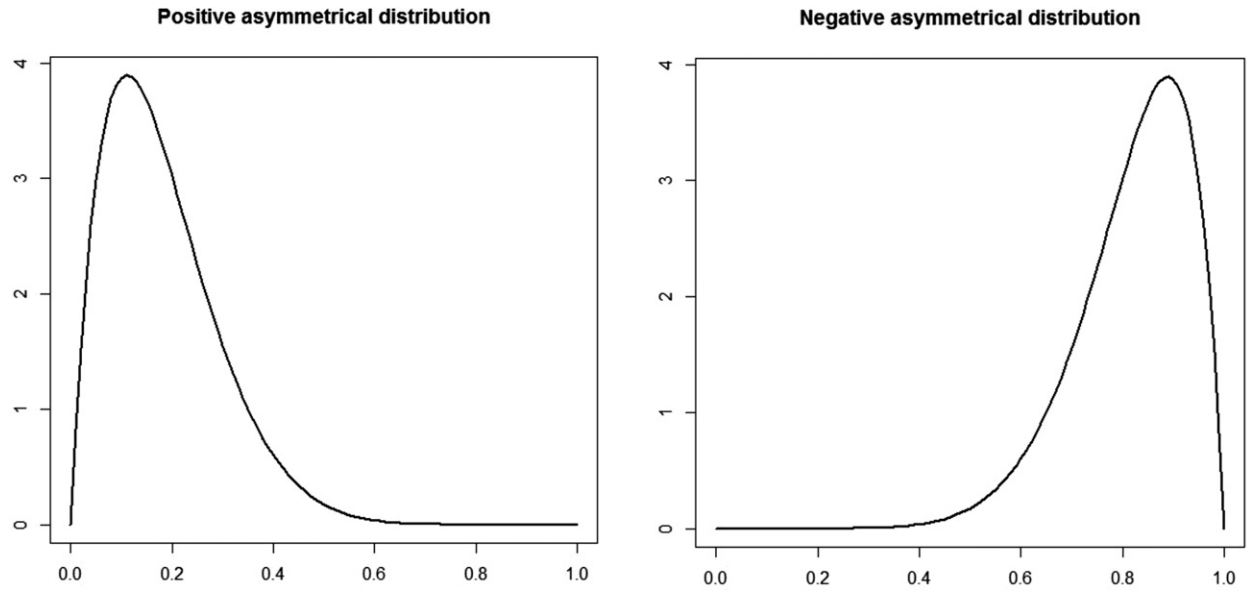

Fig. 4 Positively Skewed Distribution (to the right) and negatively Skewed Distribution (to the left).

mean, median and mode occur at different points of the distribution. In particular, there are two kinds of skewness. The distribution is said to be left-skewed (or negatively skewed) if the distribution appears to be skewed to the left, i.e. its mean is less than its mode. On the contrary, the distribution is said to be right-skewed (positively skewed) if the distribution is skewed to the right, i.e., its mean is greater than its mode. Most computer statistical packages (e.g., The R Project for Statistical Computing) include this statistic as part of a standard printout. A value of skewness >0 indicates positive skewness and a value of skewness <0 indicates negative skewness (see **Fig. 4**). As skewness involves the third moment of the distribution, kurtosis involves the fourth moment. Usually, kurtosis is quoted in the form of excess kurtosis (kurtosis relative to normal distribution kurtosis). Excess kurtosis is simply kurtosis less 3. In fact, kurtosis for a standard normal distribution is equal to three. There are three different ways to define the kurtosis. A distribution with excess kurtosis equal to zero (and kurtosis exactly 3) is called *meso*kurtic, or *meso*kurtotic. A distribution with positive excess kurtosis (and $\gamma_2 > 3$) is called leptokurtic, or leptokurtotic. A distribution with negative excess kurtosis (and $\gamma_2 > 3$) is called platykurtic, or platykurtotic. For instance, see **Fig. 5**. In terms of shape, a leptokurtic distribution has fatter tails while a platykurtic distribution has thinner tails. The computational formulas of skewness and kurtosis are shown in **Table 3**.

Inferential Statistics

Statistical inference is the procedure by which we obtain a conclusion about a population on the basis of the information contained in a sample drawn from that population. The basic assumption in statistical inference is that each element, within the population of interest, has the same probability of being included in a specific sample. Therefore, the knowledge of the probability distribution of a random variable provides the clinician and researcher with a powerful tool for summarizing and describing a set of data and for reaching conclusions about a population of data on the basis of a sample drawn from that population. In this section, we discuss two general areas of statistical inference, estimation and hypothesis testing, used to infer the population parameters under the assumption that the sample estimates follow the normal or Gaussian distribution. These types of statistical inference procedures are classified as parametric statistics.

Continuous Probability Distributions

To understand the nature of the distribution of a continuous random variable, we consider the probability density function which is the area under a smooth curve between any two points a and b, i.e., the definite integral between a and b. Thus, the probability of a continuous random variable to assume values between a and b is denoted by P(a < X < b). The graph of probability density function is shown in **Fig. 6**.

The normal distribution is the most important continuous probability distribution in statistics. It describes well the distribution of random variables that arise in practice, such as the heights, weights, blood pressure, body mass, etc. Let X be a random variable normally distributed, the probability density function of X is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < +\infty \tag{4}$$

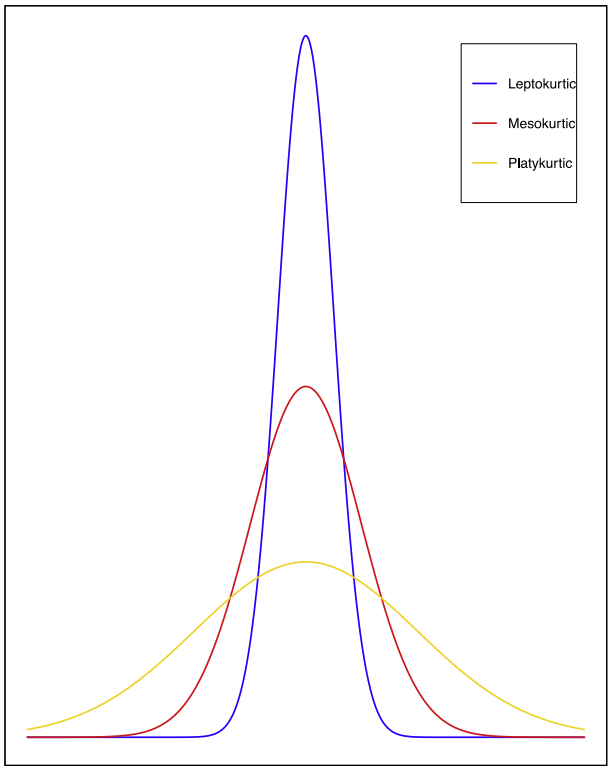

Fig. 5 Kurtosis distributions: a distribution with kurtosis equal to zero is called *meso*kurtic, or *meso*kurtotic (red line); a distribution with positive kurtosis is called leptokurtic, or leptokurtotic (blue line); a distribution with negative kurtosis is called platykurtic, or platykurtotic (gold line).

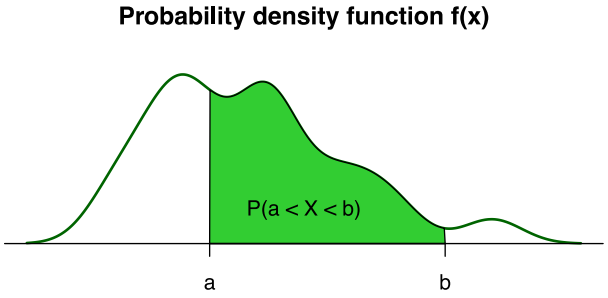

Fig. 6 Graph of a continuous distribution showing area between a and b. The probability of a continuous random variable to assume values between a and b is denoted by P(a<X<b).

where the parameters μ and σ^2 are the mean and variance of X, respectively. Generally, we write $X \sim N(\mu, \sigma^2)$. Which means that X follows the normal distribution (or X is normally distributed) with mean μ, and variance σ^2. The graph of the normal distribution produces the familiar bell-shaped curve shown in **Fig. 7**. It is symmetrical about its mean μ. Mean, median and mode are equal. The total area under the curve above the x-axis is one square unit. In particular, the 68% of observations lie between $(\mu \pm \sigma)$, 95% of observations lie between $(\mu \pm 2\sigma)$ and 99.7% of observations lie between $(\mu \pm 3\sigma)$. For instance, see **Fig. 8**. The normal distribution is completely determined by the parameters μ and σ^2. Different values of μ and σ shift the graph of the distribution along the x-axis. Different values of σ determine the degree of flatness or peakedness of the graph of the distribution. Because of the characteristics of these two parameters, μ is often referred to as a location parameter and σ is often referred to as a shape parameter. The normal distribution with mean $\mu = 0$ and $\sigma^2 = 1$ is called standard normal distribution. It is obtained from Eq. (5) by setting

$$z = \frac{(x - \mu)}{\sigma}$$

This value is called z-transformation (or z-score). Hence, the probability density function of the standard normal distribution is

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{z^2}{2}}, -\infty < x < +\infty \tag{5}$$

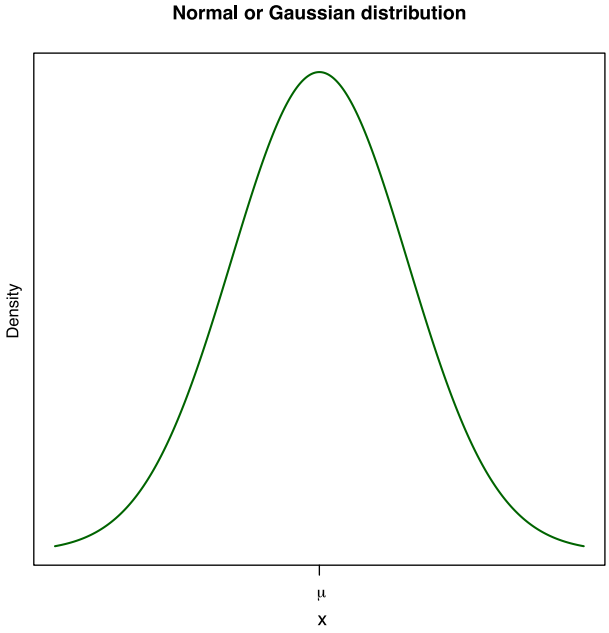

Fig. 7 Normal or Gaussian distribution with mean μ and standard deviation σ.

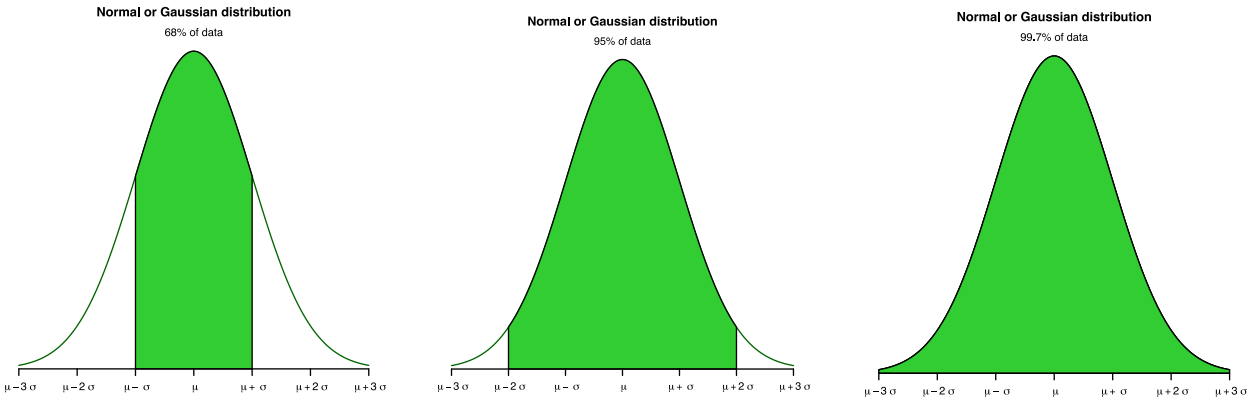

Fig. 8 Standard deviation and coverage. About 68% of values drawn from a normal distribution are within one standard deviation away from the mean; about 95% of the values lie within two standard deviations; and about 99.7% are within three standard deviations.

The graph of the standard normal distribution is shown in Fig. 9. The probability of the random variables z between two points (or to the left/right of a given z-score) on the z-axis is the areas located under the curve of the standard normal distribution. The probability of z can be calculated by using standard normal distribution tables well known in literature.

Three important distributions related to the normal distribution are: Chi-square distribution, t distribution and F distribution. Let $X_1, X_2, \ldots, X_m$ be m independent random variables having standard normal distribution, i.e., $X_i \sim N(0,1)$, the new random variable

$$Z = \sum_{i=1}^{m} X_i^2 \sim \chi_m^2$$

follows a Chi-Square distribution with m degrees of freedom (i.e., the number of random variables). Its mean is m, and its variance is $2m$. The probability density function of Z is given by

$$f(z) = \frac{1}{2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)} \exp^{-\frac{z}{2}} z^{\frac{m}{2}-1},\ 0 < z < +\infty \tag{6}$$

where the gamma function Γ is the integral

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-y} dy,\ \alpha > 0$$

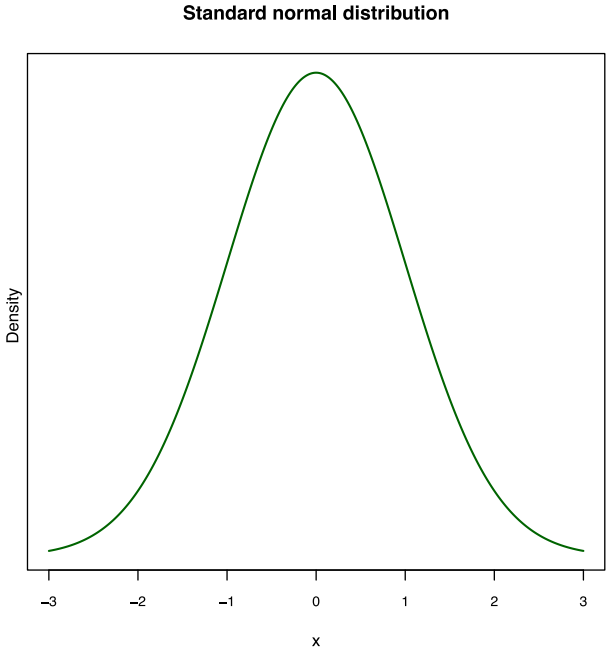

Fig. 9 Standard normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

Note that as the degrees of freedom increase, the chi-square curve approaches a normal distribution. The graph of the Chi-squares X^2 distribution is shown in **Fig. 10**.

The Student's t distribution is a probability distribution that is used to estimate population parameters when the sample size is small and/or when the population variance is unknown. Let Z be a random variable with standard normal distribution, i.e., $Z \sim N(0,1)$, and let V be a random variable having a Chi-square distribution with m degrees of freedom, i.e., $V \sim \chi^2_m$. Assume further that Z and V are independent. Define a new random variable T by

$$T = \frac{Z}{\sqrt{V/m}} \sim T_m$$

called Student t distribution with m degrees of freedom. The probability density function of the t distribution with m degrees of freedom is

$$f(t) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{\pi m}\Gamma\left(\frac{m}{2}\right)}\left(1 + \frac{t^2}{m}\right)^{-\frac{m+1}{2}}, -\infty < t < +\infty \tag{7}$$

It is symmetrical about the mean equal to zero. It has a variance greater than 1, but the variance approaches 1 as the sample size becomes large, i.e., $Var(T) = \frac{m}{m-2}$. The shape of the t-distribution curve depends on the number of degrees of freedom. Compared to the normal distribution, the t distribution is less peaked in the center and has thicker tails. Finally, the t distribution approaches the standard normal distribution as m tends to infinity. The graph of the Student t distribution is shown in **Fig. 11**.

Let U and V be independent chi-square random variables with m and n degrees of freedom, respectively. The variable

$$F = \frac{U/m}{V/n} \sim F_{m,n}$$

follows the Fisher F distribution with numerator degree of freedom m and denominator degree of freedom n. The probability density function of the F distribution is

$$f(x) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\left(\frac{m}{n}\right)^{\frac{m}{2}} x^{\frac{n}{2}-1}\left(1 + \frac{m}{n}x\right)^{-\frac{(m+n)}{2}}, \; 0 < t < +\infty \tag{8}$$

Then, the mean is $E(X) = \frac{n}{n-2}$ and the variance $\mathrm{Var(X)} = \frac{2n^2(m+n-2)}{m(n-2)^2(m-4)}$. In general, the F distribution is skewed to the right. The graph of the Fisher F distribution is shown in **Fig. 12**. The X^2,t,F distributions, like the standard normal, has been extensively tabulated.

Estimation

The estimation process consists of estimate sample statistics in order to give an approximation of the corresponding parameters of the population from which the sample is drawn. For example, we suppose that the administrator of a hospital is interested in the

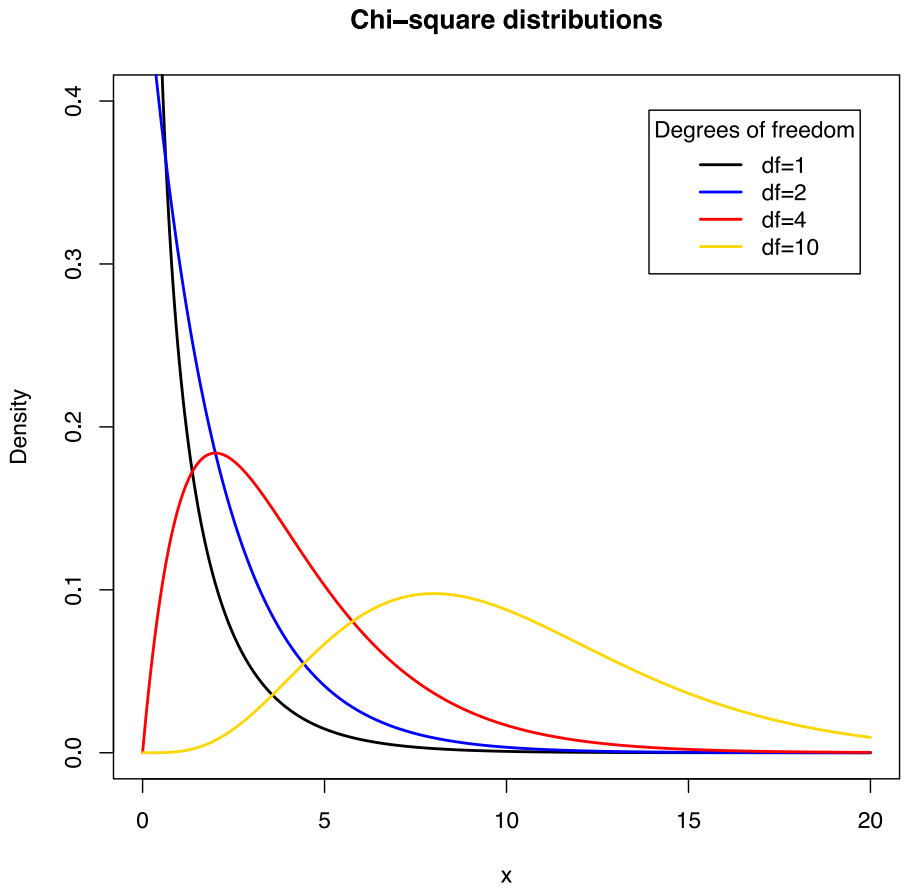

Fig. 10 Chi-square distribution for different degrees of freedom.

mean age of patients admitted to the hospital during a given year. He decides to examine only a sample of the records to conduct his analysis in order to determine a mean age estimation of all patients admitted to the hospital during the year. This statistic is an estimation of the corresponding population mean. Typically, we expect the estimate to differ by some amount from the parameter it estimates.

For each of the population parameters, we can compute two types of estimate: a point estimate and an interval estimate. A point estimate is a single numerical value used to estimate the corresponding population parameters, while an interval estimate is an interval that, with a specified degree of confidence, most likely includes the parameters being estimated (Gardner and Altman, 1986). For example, let X be a random variable that follows the normal distribution with mean μ, and variance σ^2. The computed sample mean $\bar{x}$ is a point estimator of the population mean μ. Similarly, the computed sample variance s^2 is a point estimation of the of the population variance σ^2. Interval estimation is an alternative procedure to point estimation. It consists to replace the point estimator of the population parameter by using a statistic that allows to calculate an interval of the parameter space. Let Θ be the parameter space and let X be a random variable from a distribution that belongs to a family of distributions with a parameter $\theta \in \Theta$. θA confidence interval (CI) is an interval composed by two numerical values, called lower and upper limit, that with a specified degree of confidence $\alpha \in (0,1)$ includes the unknown parameter θ. In other words, it is an interval that with probability $1-\alpha$, include the unknown parameter θ. The probability $1-\alpha$ is called the confidence coefficient and represents the area under the probability distribution between the two limits of the CI. Usually, $1-\alpha$ is taken to be 0.90, 0.95 or 0.99. To construct a confidence interval CI, we generally consider the following steps:

1. the sample statistic is identified to estimate a population parameter θ (for example, the population mean or the population variance);
2. the confidence level $1-\alpha$ is fixed to compute the margin of error, i.e., the product between the critical value (which is a term that splits the area under the probability distribution in two regions) and the standard deviation;
3. the limits of the confidence interval are determined as follow.

$$\text{CI} = \text{sample statistic} \pm \text{margin of error} \tag{9}$$

In particular, when X is a random variable normally distributed with mean μ, and variance σ^2, we can construct different type of CI for the mean μ with known or unknown variance σ^2.

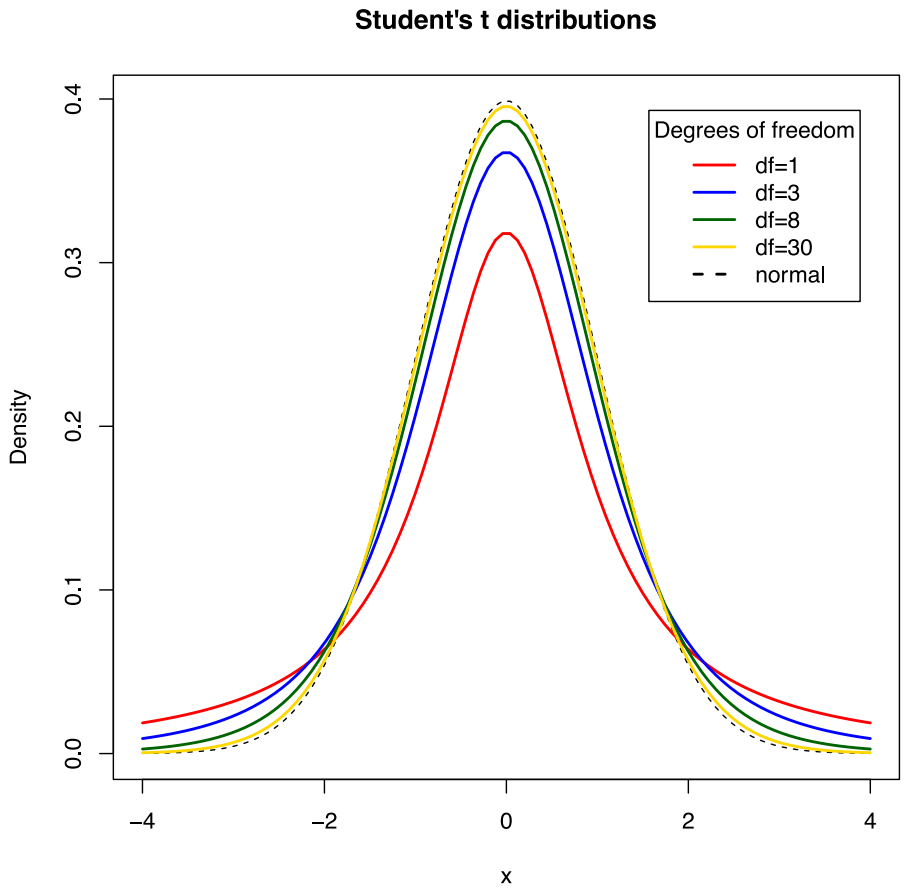

Fig. 11 Student distribution for different degree of freedom.

Confidence interval CI for the population mean μ

When the variance σ^2 is known, the statistic used to construct a $100(1-\alpha)$ confidence interval CI for the population mean μ is the quantity

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \tag{10}$$

where σ is the known population standard deviation. Then, an interval estimate for μ is expressed as

$$\left(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \tag{11}$$

where the critical value, denoted by $z_{\alpha/2}$, is the value of z to the left of which lies $-\alpha/2$ and to the right of which lies $\alpha/2$ of the area under its curve, i.e., $P(Z \geq |z_{\alpha/2}|) = \frac{\alpha}{2}$ with $Z \sim N(0,1)$. For instance, see **Fig. 13**. When the variance σ^2 is unknown and the sample size n is large $(n > 30)$, we consider the Student's t distribution. In this case, the statistic used to construct a $100(1-\alpha)$ confidence interval CI for the population mean μ is given by

$$t = \frac{x - \mu}{s/\sqrt{n}} \sim T_{(n-1)} \tag{12}$$

where s is the sample standard deviation to replace σ in Eq. (11). This statistics follows a Student's distribution with $n-1$ degrees of freedom. An interval estimate for μ is expressed as

$$\left(\overline{x} - t_{\alpha/2;n-1}\frac{s}{\sqrt{n}},\ \overline{x} + t_{\alpha/2;n-1}\frac{s}{\sqrt{n}}\right) \tag{13}$$

where the critical value, denoted by $t_{\alpha/2,n-1}$, is the value of t to the left of which lies $-\alpha/2$ and to the right of which lies $\alpha/2$ of the area under its curve, i.e. $P(T \geq |t_{\alpha,n-1}|) = \frac{\alpha}{2}$ with $T \sim T_{(n-1)}$. For instance, see **Fig. 14**.

Confidence interval CI for the difference between the population means $\mu_1 - \mu_2$

Sometimes we are interested in estimating the difference between two population means. From each of the populations an independent random sample is drawn and, from the data of each, the sample means $\overline{x}_1$ and $\overline{x}_2$ respectively, are computed.

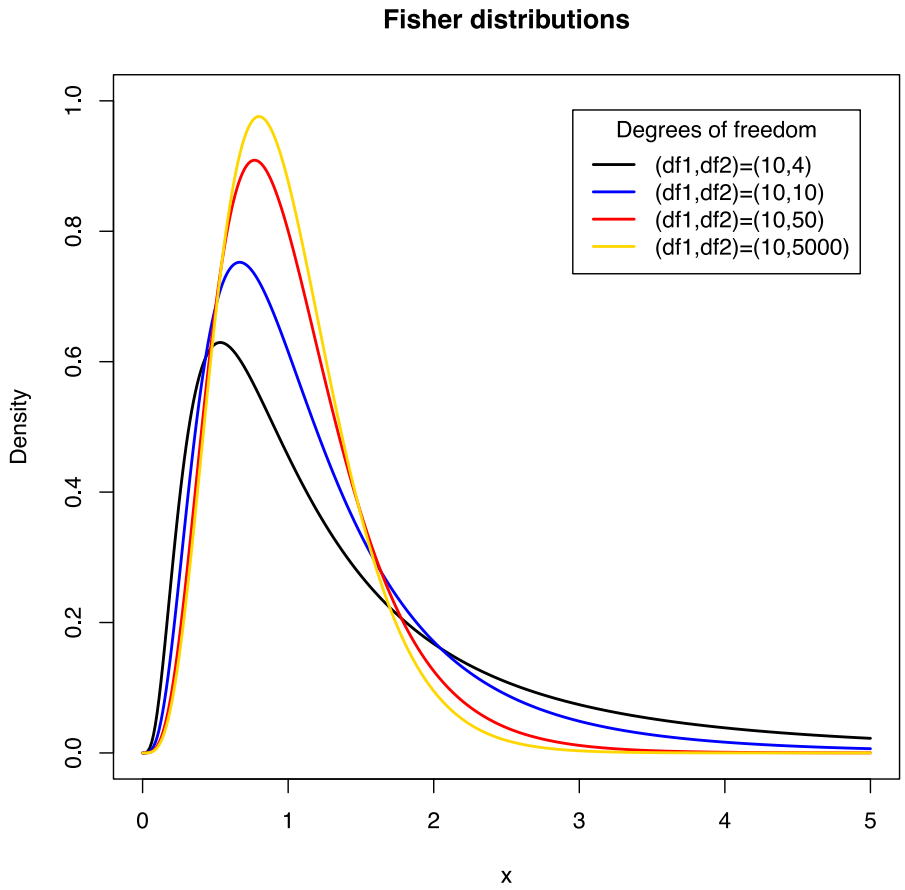

Fig. 12 Fisher distribution for different degree of freedom.

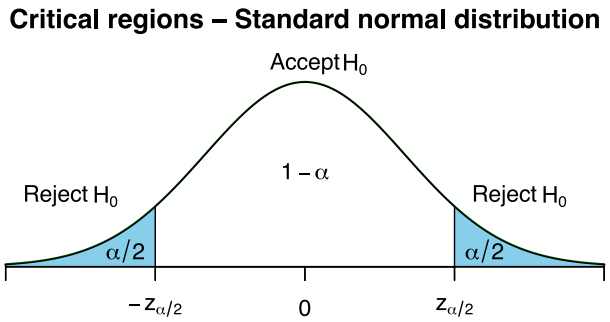

Fig. 13 Critical regions for the standard normal distribution.

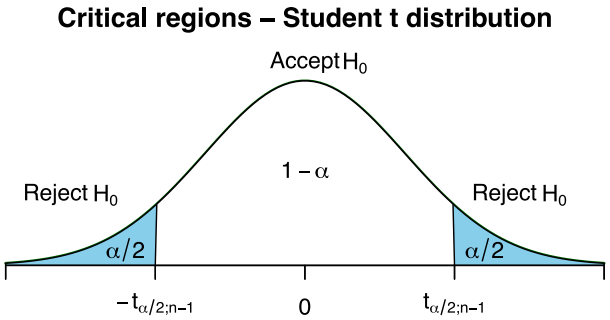

Fig. 14 Critical regions for the Student distribution.

An unbiased estimate of the difference between the population means, $\mu_1 - \mu_2$, is the difference between the sample means, $\bar{x}_1 - \bar{x}_2$. The variance of the estimator is $\left(\frac{\sigma_1^2}{n}\right) + \left(\frac{\sigma_2^2}{m}\right)$, where n and m are the sample sizes. The statistic used to construct a $100(1-\alpha)$ confidence interval CI for the difference between the population means, $\mu_1 - \mu_2$ is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim \mathrm{N}(0,1) \tag{14}$$

Hence, a confidence interval CI for $\mu_1 - \mu_2$ is given by

$$\left((\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}},\ (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\right) \tag{15}$$

where the critical value, denoted by $z_{\alpha/2}$, is the value of z to the left of which lies $-\alpha/2$ and to the right of which lies $\alpha/2$ of the area under its curve, i.e. $P(Z \geq |z_{\alpha/2}|) = \frac{\alpha}{2}$ with $Z \sim N(0,1)$. For instance, see **Fig. 13**. An investigation of a confidence interval CI for the difference between population means provides information that is helpful in deciding whether or not it is likely that the two population means are equal. When the constructed interval does not include zero, we say that the interval provides evidence that the two population means are not equal. When the interval includes zero, we say that the population means may be equal. When population variances are unknown, we use the t distribution to estimate the difference between two population means with a confidence interval CI. We assume that the two sampled populations are normally distributed. With regard to the population variances, we distinguish two cases: (1) the population variances are equal, and (2) the population variances are not equal. Let us consider each situation separately. If the population variances are equal, the two sample variances that we compute from two independent samples are estimates of the same quantity, the common variance. This estimation is called pooled estimate and it is obtained by computing the weighted average of the two sample variances. Each sample variance is weighted by its degrees of freedom. The pooled estimate is given by the formula

$$s_p^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2} \tag{15}$$

where n and m are the sample sizes. The statistic used to construct a $100(1-\alpha)$ confidence interval CI for the difference between the population means, $\mu_1 - \mu_2$ is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim \mathrm{T}_{(n+m-2)} \tag{16}$$

Hence, a confidence interval CI for μ_1–μ_2, when population variances are unknown and equal, is given by

$$\left((\bar{x}_1 - \bar{x}_2) - t_{\alpha/2;\ n+m-2}s_p\sqrt{\frac{1}{n} + \frac{1}{m}}, (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2;\ n+m-2}s_p\sqrt{\frac{1}{n} + \frac{1}{m}}\right) \tag{17}$$

where the critical value, denoted by $t_{\alpha/2;n+m-2}$, is the value of t to the left of which lies $-\alpha/2$ and to the right of which lies $\alpha/2$ of the area under its curve, i.e., $P(T \geq |t_{\alpha,n+m-2}|) = \frac{\alpha}{2}$ with $T \sim T_{(n+m-2)}$.

If the population variances are not equal, the solution, proposed by Cochran (1964) consists of computing the reliability factor, $t'_{\alpha/2}$ by the following formula:

$$t'_{\alpha/2} = \frac{w_1 t_1 + w_2 t_2}{w_1 + w_2}$$

where $w_1 = \frac{s_1^2}{n}$, $w_2 = \frac{s_2^2}{m}$, t_1–$t_{\alpha/2}$ for n–1 degrees of freedom, and t_2–$t_{\alpha/2}$ for m–1 degrees of freedom. Hence, an approximate a 100 (1–α) confidence interval CI for the difference between the population means, μ_1–μ_2 is given by

$$\left((\bar{x}_1 - \bar{x}_2) - t'_{\alpha/2}s_p\sqrt{\frac{1}{n} + \frac{1}{m}}, (\bar{x}_1 - \bar{x}_2) + t'_{\alpha/2}s_p\sqrt{\frac{1}{n} + \frac{1}{m}}\right) \tag{18}$$

Confidence interval CI for a population proportion

Many questions of interest to the health science are related to population proportions. For example, the proportion of patients who receive a particular type of treatment, or the proportion of some population who has a certain disease or the proportion of a population who is immune to a certain disease. In this case, we consider the binomial distribution frequently used to model the number of successes p in a sample of size n drawn with replacement from a population of size n. Hence, the binomial distribution is characterized by two parameters, n and p. When the sample size is large, the distribution of sample proportions is approximately normally distributed by virtue of the central limit theorem. The mean of the distribution, $\mu_{\hat{p}}$, that is, the average of all the possible sample proportions, is equal to the true population proportion, p, and the variance of the distribution, $\sigma_{\hat{p}}^2$, is equal to $\frac{p(1-p)}{n}$. To estimate the population proportion, we compute the sample proposition $\hat{p}$. This sample proportion is used as the point estimator of the population proportion. In particular, when both np and $n(1-p)$ are greater than 5, we can say that the sampling distribution of $\hat{p}$ is approximately normally distributed with mean $\mu_{\hat{p}} = p$. Hence, the statistic used to construct a 100(1–α) confidence interval

CI for the population proportion p is given by

$$z = \frac{p - \hat{p}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1) \tag{19}$$

A confidence interval CI for the population proportion p is

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \tag{20}$$

where the critical value, denoted by $z_{\alpha/2}$, is the value of z to the left of which lies $-\alpha/2$ and to the right of which lies $\alpha/2$ of the area under its curve, i.e., $P(Z \geq |z_{\alpha/2}|) = \frac{\alpha}{2}$ with $Z \sim N(0,1)$. For instance, see **Fig. 13**.

Confidence interval CI for the difference between two population proportions

Often there are two population proportions in which we are interested and we desire to assess the probability associated with a difference in proportions computed from samples drawn from each of these populations. The relevant sampling distribution is the distribution of the difference between the two sample proportions. If independent random samples of size n and m are drawn from two populations of dichotomous variables where the proportions of observations with the characteristic of interest in the two populations are p_1 and p_2, respectively, the distribution of the difference between sample proportions, $\hat{p}_1 - \hat{p}_2$, is approximately normal with mean and variance equal to

$$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2, \text{ and } \sigma^2_{\hat{p}_1 - \hat{p}_2} = \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}$$

respectively, when n and m are large (i.e., np_1, mp_2, $n(1-p_1)$ and $m(1-p_2)$ are greater than 5). Hence, an unbiased point estimator of the difference between two population proportions is provided by the difference between sample proportions, $\hat{p}_1 - \hat{p}_2$. The statistic used to construct a $100(1-\alpha)$ confidence interval CI for the difference between two population proportions $p_1 - p_2$ is

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}} \sim N(0,1) \tag{21}$$

A confidence interval CI for $p_1 - p_2$ is given by

$$\left((\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}, (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}\right) \tag{22}$$

where the critical value, denoted by $z_{\alpha/2}$, is the value of z to the left of which lies $-\alpha/2$ and to the right of which lies $\alpha/2$ of the area under its curve, i.e., $P(Z \geq |z_{\alpha/2}|) = \frac{\alpha}{2}$ with $Z \sim N(0,1)$. For instance, see **Fig. 13**.

Confidence interval CI for the population variance σ^2

The statistic used to construct a $100(1-\alpha)$ confidence interval CI for the population variance σ^2 is

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1) \tag{23}$$

This statistics follows a chi-square X^2 distribution with $n-1$ degrees of freedom. An interval estimate for σ^2 is expressed as

$$\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2;n-1}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2;n-1}}\right) \tag{24}$$

where s is the sample variance and $\chi^2_{1-\alpha/2;n-1)}$ and $\chi^2_{\alpha/2;n-1)}$ are the values from the X^2 table to the left and right of which, respectively, lies $\alpha/2$ of the area under the curve, i.e., $P(Y \leq \chi^2_{1-\alpha/2;n-1}) = 1 - \frac{\alpha}{2}$ and $P(Y \geq \chi^2_{\alpha/2;n-1)}) = \frac{\alpha}{2}$ with $Y \sim \chi^2_{(n-1)}$. For instance, see **Fig. 15**. If we take the square root of each term in Eq. (23), we have the confidence interval for σ, the population standard deviation.

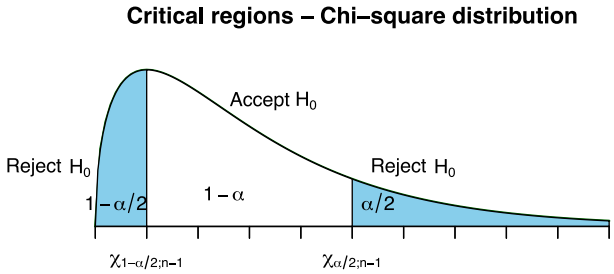

Fig. 15 Critical regions for the Chi-square distribution.

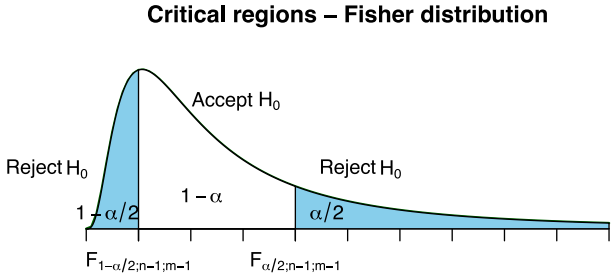

Fig. 16 Critical regions for the Fisher distribution.

Confidence interval CI for the ratio of the variances of two normally distributed populations

Generally, we consider the ratio σ_1^2/σ_2^2 to compare the variances of two normally distributed populations. If two variances are equal, their ratio will be equal to 1. We usually will not know the variances of populations of interest, and, consequently, any comparisons we make will be based on sample variances. In other words, we may wish to estimate the ratio of two population variances. The assumptions are that s_1^2 and s_2^2 are computed from independent samples of size n and m respectively, drawn from two normally distributed populations. The variance s_1^2 is designed as the larger of the two sample variances. Hence, the statistic used to construct a $100(1-\alpha)$ confidence interval CI for the ratio of the variances of two normally distributed populations is

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{(n-1;m-1)} \tag{25}$$

This statistic follows a F distribution depending on two-degrees-of-freedom values, one corresponding to the value of $n-1$ used in computing s_1^2 and the other corresponding to the value of $m-1$ used in computing s_2^2. These are usually referred to as the numerator degrees of freedom and the denominator degrees of freedom. An interval estimate for the ratio σ_1^2/σ_2^2 is expressed as

$$\left(\frac{s_1^2/s_2^2}{F_{1-\alpha/2;n-1;m-1}}, \frac{s_1^2/s_2^2}{F_{\alpha/2;n-1;m-1}}\right) \tag{26}$$

where $F_{1-\alpha/2;n-1;m-1}$ and $F_{\alpha/2;n-1;m-1}$ are the values from the F table to the left and right of which, respectively, lies $\alpha/2$ of the area under the curve i.e., $P(F \leq F_{1-\alpha/2;n-1;m-1}) = 1 - \frac{\alpha}{2}$ and $P(Y \geq F_{\alpha/2;n-1;m-1}) = \frac{\alpha}{2}$ with $F \sim F_{(n-1;m-1)}$. For instance, see **Fig. 16**.

Note that the tables of all the critical values can be used for both one-sided (lower and upper) and two-sided tests with specific values of α.

Hypothesis Testing

The aim of hypothesis testing is to aid the clinician and researcher in reaching a conclusion concerning a population by examining a sample from that population. Interval estimation, discussed in the preceding section, and hypothesis testing are based on similar concepts. In fact, confidence intervals can be used to obtain the same conclusions that are reached through the use of hypothesis tests. There are two statistical hypotheses involved in hypothesis testing: the null hypothesis and the alternative hypothesis. The null hypothesis is the hypothesis to be tested and it is designated by the symbol H_0. The null hypothesis is the hypothesis of no difference (or equality, either $=, \leq$, or $\geq$), since it is a statement of agreement with conditions supposed to be true in the population under investigated. Consequently, the conclusion that the researcher is seeking to reach is to reject the null hypothesis. If the null hypothesis is not rejected, we conclude that the data not provide sufficient evidence that the null hypothesis is not in reality true. The alternative hypothesis, designed by the symbol H_1, is the statement that researchers hope to be true. In other words, it is the hypothesis of effect or real difference. Based on the sample data, the test determines whether to reject the null hypothesis. In particular, we can follow two types of decisional strategies. The first approach is based on the computation of test statistic, the second one is called p-value approach (Altman and Krzywinski, 2017; Gardner and Altman, 1986). All possible values that the test statistic can assume are arranged on the horizontal axis of the probability distribution and are divided into two regions: the rejection region and non rejection region. The values of the test statistic forming the rejection region are those values that are less likely to occur if the null hypothesis is true, while the values making up the acceptance region are more likely to occur if the null hypothesis is true. The decision rule tells us to reject the null hypothesis if the value of the test statistic that we compute from the sample falls in the rejection region or not. The decision to reject or accept the null hypothesis is based on the level of significance α. A computed value of the test statistic that falls in the rejection region is said to be significant. Generally, a small value of α is selected to make the probability of rejecting a true null hypothesis small. The more frequently values used for α are 0.01, 0.05, and 0.10. The relationship between the (unknown) reality if the null hypothesis is true or not and the decision to accept or reject the null hypothesis is shown in **Table 4**. The error committed when a true null hypothesis is rejected is called the type I error. The type II error is the error committed when a false null hypothesis is not rejected. The probability of committing a type II error is designated by β. The II error is know as the statistical power, which is the ability of a test to detect a true effect, i.e., reject the null hypothesis if the alternative hypothesis is true. The second strategy is based on the concept of p-value, which is the probability that the computed test statistic is at least as extreme as a specified value of the test statistic when the null

Table 4 Conditions under which type I and type II errors may be committed

Decision rules	*The truth*	
	H_0 true	*H_1 true*
Accept H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

Table 5 Confidence intervals (CI) and hypothesis test for the single population mean μ and for the difference between two population means μ_1 and μ_2 when sampling from normally distributed populations

Statistics	σ^2 *Known*		
	CI at level α	*Hypthesis test*	*Critical regions*
$z=\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\sim N(0,1)$ *z*-transformation	$\left(\overline{x}-z_{a/2}\frac{\sigma}{\sqrt{n}},\overline{x}+z_{a/2}\frac{\sigma}{\sqrt{n}}\right)$		Reject H_0 $z\leq -z_{a/2}$ez $\geq t_{a/2}$ *p*-value $<\alpha$
		$H_0:\mu=\mu_0$ $H_1:\mu\neq\mu_0$ $(\mu>\mu_0$o$\mu<\mu_0)$	Accept H_0 $-z_{a/2}\leq z\leq z_{a/2}$ *p*-value $>\alpha$
$z=\frac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}}\sim N(0,1)$ *n* and *m* sample sizes	$\left(\overline{X}_1-\overline{X}_2-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}},\overline{X}_1-\overline{X}_2+z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}\right)$		Reject H_0 $z\leq -z_{a/2}$ez $\geq z_{a/2}$ *p*-value $<\alpha$
		$H_0:\mu_1=\mu_2$ $H_1:\mu_1\neq\mu_2$ $(\mu_1>\mu_2$o$\mu_1<\mu_2)$	Accept H_0 $-z_{a/2}\leq z\leq z_{a/2}$ *p*-value $>\alpha$

hypothesis is true. Thus, the *p*-value is the smallest value (typical less than 5%) for which we reject a null hypothesis. To calculate the *p*-value, we first calculate the value of the test statistic, and then, using the known distribution of the test statistic, calculate the *p*-value. Note that the probability of rejecting a true null hypothesis is the significance level α, we conclude that, if the *p*-value is less than (or equal to) α, then the null hypothesis is rejected, while, if the *p*-value is greater than α, then the null hypothesis is not rejected (or accepted). For instance, see **Tables 5** and **6**.

Non-Parametric Statistics

In this section, we explore the most common non-parametric techniques used when the underlying assumptions of traditional hypothesis tests are violated. These statistical procedures allow for the testing of hypotheses that are not statements about population parameter values and are applied when the form of the sampled population is unknown.

Wilcoxon signed-rank test for location

Suppose to test a null hypothesis about a population mean, but neither z nor t is an appropriate test statistic because the sampled population does not follow or approximate a normal distribution (Wilcoxon, 1945). When confronted with such a situation we use a non-parametric statistical procedure called Wilcoxon signed-rank test for location. It makes use of the magnitudes of the differences between measurements and a hypothesized location parameter rather than just the signs of the differences. The Wilcoxon test is based on the following assumptions about the data: (i) the sample is random; (ii) the variable is continuous; (iii) the population is symmetrically distributed about its mean μ; (iv) the measurement scale is at least interval. After the formulation of null mean H_0 and alternative hypothesis H_1,

$$H_0:\mu=\mu_0(\leq,\ \geq)\ \text{vs}\ H_1:\mu\neq\mu_0(>,<)$$

we perform the Wilcoxon test when the population mean μ_0 is unknown.

1. Subtract the Hypothesized Mean μ_0 from Each Observation x_i, to Obtain $d_i=x_i-\mu_0$.
   If any x_i is equal to the mean, so that, $d_i=0$, eliminate that d_i from the calculations and reduce n accordingly.
2. Rank the usable d_i from the smallest to the largest without regard to the sign of d_i. That is, consider only the absolute value of the d_i, designated $|d_i|$, when ranking them. If two or more of the $|d_i|$ are equal, assign each tied value the mean of the rank positions the tied values occupy. If, for example, the three smallest $|d_i|$ are all equal, place them in rank positions 1, 2, and 3, but assign each a rank of $(1+2+3)/3=2$.

Table 6 Confidence intervals (CI) and hypothesis test for the single population mean μ, for the difference between two population means μ_1 and μ_2, for population variance σ^2 and for the ratio of the variances when sampling from normally distributed populations

Statistics	σ^2 *Unknown*		
	CI at level α	*Hypthesis test*	*Critical regions*
$t = \frac{\overline{X}-\mu_0}{s/\sqrt{n}} \sim T_{(n-1)}$ *t*-transformation	$\left(\overline{X} - t_{\alpha/2;n-1}\frac{s}{\sqrt{n}}, \overline{X} + t_{\alpha/2;n-1}\frac{s}{\sqrt{n}}\right)$	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$ $(\mu > \mu_0 o \mu < \mu_0)$	Reject H_0 $t \leq -t_{\alpha/2;n-1} e t \geq t_{\alpha/2;n-1}$ *p*-value $< \alpha$ Accept H_0 $-t_{\alpha/2;n-1} \leq t \leq t_{\alpha/2;n-1}$ *p*-value $> \alpha$
$t = \frac{\overline{x}_1-\overline{x}_2}{s_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim T_{(n+m-2)}$ where $s_p^2 = \frac{(n-1)s_1^2+(m-1)s_2^2}{n+m-2}$ n and m sample sizes	$\left(\overline{X}_1 - \overline{X}_2 - t_{\alpha/2;n+m-2}s_p\sqrt{\frac{1}{n}+\frac{1}{m}}, \overline{X}_1 - \overline{X}_2 + t_{\alpha/2;n+m-2}s_p\sqrt{\frac{1}{n}+\frac{1}{m}}\right)$	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$ $(\mu_1 > \mu_2 o \mu_1 < \mu_2)$	Reject H_0 $t \leq -t_{\alpha/2;n+m-2} e t \geq t_{\alpha/2;n+m-2}$ *p*-value $< \alpha$ Accept H_0 $-t_{\alpha/2;n+m-2} t \leq t_{\alpha/2;n+m-2}$ *p*-value $> \alpha$
$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2_{n-1}$	$\left(\frac{n-1}{\chi^2_{\alpha/2;n-1}}s^2, \frac{n-1}{\chi^2_{1-\alpha/2;n-1}}s^2\right)$	$H_0 : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 \neq \sigma_0^2$ $(\sigma^2 < \sigma_0^2 o \sigma^2 > \sigma_0^2)$	Reject H_0 $0 \leq \chi^2 < \leq \chi^2_{1-\alpha/2;n-1} e \chi^2 \geq \chi^2_{\alpha/2;n-1}$ Accept H_0 $\chi_{1-\alpha/2;n-1} \leq \chi^2 \leq \chi^2_{\alpha/2;n-1}$
$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{(n-1,\ m-1)}$	$\left(\frac{s_1^1/s_2^2}{F_{1-\alpha/2;n-1;m-1}}, \frac{s_1^1/s_2^2}{F_{\alpha/2;n-1;m-1}}\right)$	$H_0 : \sigma_1^2 = \sigma_2^2$ $H_0 : \sigma_1^2 \neq \sigma_2^2$ $(\sigma_1^2 > \sigma_2^2, \sigma_1^2 < \sigma_2^2)$	Reject H_0 $0 \leq F \leq F_{1-\alpha/2;n-1;m-1} e F \geq F_{\alpha/2;n-1;m-1}$ Accept H_0 $F_{1-\alpha;n-1;m-1} \leq F \leq F_{\alpha/2;n-1;m-1}$

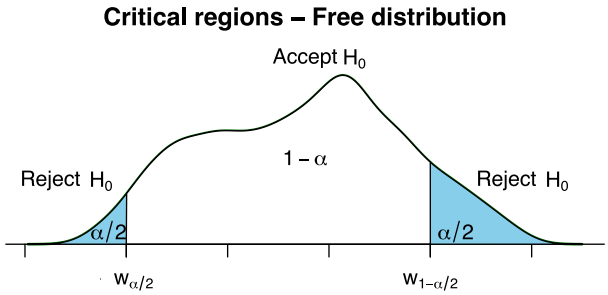

Fig. 17 Critical regions for a free distribution.

3. Assign each rank the sign of the $|d_i|$ that yields that rank.
4. Find T_+ the sum of the ranks with positive signs, and, T_-, the sum of the ranks with negative signs.

The Wilcoxon test statistic is either T_+ or T_-, depending on the nature of the alternative hypothesis. If the null hypothesis is true, that is, if the true population mean is equal to the hypothesized mean, the probability of observing a positive difference $d_i = x_i - \mu_0$ of a given magnitude is equal to the probability of observing a negative difference of the same magnitude. Then, in repeated sampling, when the null hypothesis is true and the assumptions are met, the expected value of T_+ is equal to the expected value of T_-. However, when H_0 is true, we do not expect a large difference in their values. Consequently, a sufficiently small value of T_+ or a sufficiently small value of T_- will cause rejection of H_0. When the alternative hypothesis is two-sided ($\mu \neq \mu_0$), either a sufficiently small value of T_+ or a sufficiently small value of T_- will cause rejection of H_0. The test statistic, then, is T_+ or T_-, whichever is smaller. To simplify notation, we call the smaller of the two T. Similarly, when the one-sided alternative hypothesis is true a sufficiently small (or large) value of T_+ (or T_-) will cause rejection of H_0, and T_+ (or T_-) is the test statistic. Critical values of the Wilcoxon test statistic are given in probability tables well known in literature. The following are the decision rules for the three possible alternative hypotheses:

1. $H_1: \mu \neq \mu_0$. Reject H_0 at the level of significance α if the calculated T is smaller than or equal to the tabulated T for n and preselected $\alpha/2$ (see Fig. 17).
2. $H_1: \mu < \mu_0$. Reject H_0 at the level of significance α if T_+ is less than or equal to the tabulated T for n and preselected α.
3. $H_1: \mu > \mu_0$. Reject H_0 at the level of significance α if T_- is less than or equal to the tabulated T for n and preselected α.

The Mann–Whitney test

Another important non-parametric test is the Mann–Whitney test based on the ranks of the observations (Mann and Whitney, 1947). The assumptions underlying the Mann–Whitney test are as follows: (i) the two samples, of size n and m, respectively, available for analysis have been independently and randomly drawn from their respective populations; (ii) the measurement scale is at least ordinal; (iii) the variable of interest is continuous; (iv) if the populations differ at all, they differ only with respect to their medians. When these assumptions are met we test the null hypothesis that the two populations have equal medians against either of the three possible alternatives: (1) the populations do not have equal medians (two-sided test), (2) the median of population 1 is larger than the median of population 2 (one-sided test), or (3) the median of population 1 is smaller than the median of population 2 (one-sided test). If the two populations are symmetric, so that within each population the mean and median are the same, the conclusions we reach regarding the two population medians will also apply to the two population means. In particular, the null and alternative hypotheses are given by:

$$H_0 : M_X = M_Y (\leq, \geq) \text{ vs } H_1 : M_X \neq M_Y (>, <)$$

where M_X is the median of a population of population 1 and M_Y is the median of population 2. For a fixed significance level α, we compute the test statistic combining the two samples and rank all observations from smallest to largest while keeping track of the sample to which each observation belongs. Tied observations are assigned a rank equal to the mean of the rank positions for which they are tied. The test statistic is

$$T = S - \frac{n(n+1)}{2}$$

where n is the number of sample X observations and S is the sum of the ranks assigned to the sample observations from the population of X values. The choice of which sample's values we label X is arbitrary. If the median of the X population is smaller than the median of the Y population, as specified in the alternative hypothesis, we would expect (for equal sample sizes) the sum of the ranks assigned to the observations from the X population to be smaller than the sum of the ranks assigned to the observations from the Y population. A sufficiently small value of T will cause rejection of H_0. Critical values of the Mann-Whitney test statistic are given in probability table well known in literature. The following are the decision rules for the three possible alternative hypotheses:

1. $H_1: M_X \neq M_Y$. Reject H_0 if the computed T is either less than $w_{\alpha/2}$, or greater than $w_{1-\alpha/2}$, where $w_{\alpha/2}$, is the tabulated critical value of T for n, the number of X observations; m, the number of Y observations; and $\alpha/2$, the chosen level of significance, and $w_{1-\alpha/2} = nm - w_{\alpha/2}$. For instance, see Fig. 17.

Table 7 Clinical depression data. For each patient, the dataset contains the following characteristic or variables. Hospt: the patient's hospital with 1, 2, 3, 5, or 6; Treat: the treatment received by the patient (Lithium, Imipramine, or Placebo); Outcome: recurrence or no recurrence occurred during the patient's treatment; Time: the length (in days) of the patient's participation in the study in terms of recurrence or no recurrence; AcuteT: the time (in days) that the patient was depressed prior to the study; Age: the age of the patient in years, when the patient entered the study; Gender: The patient's gender (1 = Female, 2 = Male). The number of patients are 109

Hospt	*Treat*	*Outcome*	*Time*	*AcuteT*	*Age*	*Gender*
1	Lithium	Recurrence	36,143	211	33	1
1	Imipramine	No Recurrence	105,143	176	49	1
1	Imipramine	No Recurrence	74,571	191	50	1
1	Lithium	Recurrence	49,714	206	29	2
1	Lithium	No Recurrence	14,429	63	29	1
1	Placebo	Recurrence	5	70	30	2
1	Lithium	No Recurrence	104,857	55	56	1
1	Placebo	Recurrence	2,857	512	48	1
1	Placebo	No Recurrence	102,429	162	22	2
1	Placebo	Recurrence	55,714	306	61	2
1	Imipramine	No Recurrence	106,429	165	58	1
1	Imipramine	No Recurrence	105,143	129	31	1
1	Imipramine	No Recurrence	83	428	44	1
1	Imipramine	Recurrence	27,286	256	55	2
1	Lithium	No Recurrence	105,857	197	57	2
1	Lithium	Recurrence	5,571	227	46	1
1	Imipramine	No Recurrence	98	168	58	1
1	Lithium	No Recurrence	16,286	194	57	1
2	Lithium	Recurrence	1,286	173	54	1
2	Lithium	No Recurrence	2,143	48	23	1
2	Imipramine	No Recurrence	100	47	65	1
2	Imipramine	Recurrence	27,143	95	27	1
2	Lithium	Recurrence	4	148	50	1
2	Lithium	Recurrence	74,143	127	41	2
2	Placebo	No Recurrence	104,857	129	65	1
2	Placebo	Recurrence	0,143	182	52	1
2	Placebo	Recurrence	1,429	90	60	1
2	Placebo	Recurrence	45,857	177	25	2
2	Imipramine	Recurrence	17,429	234	27	2
2	Imipramine	No Recurrence	78	322	32	1
2	Imipramine	Recurrence	66,857	141	43	2
2	Placebo	No Recurrence	78,429	165	20	2
2	Lithium	No Recurrence	78,429	239	23	2
2	Imipramine	No Recurrence	78,143	147	36	2
2	Imipramine	No Recurrence	15,857	348	22	2
3	Lithium	No Recurrence	79	274	49	2
3	Imipramine	No Recurrence	32,571	130	40	2
3	Lithium	Recurrence	9	98	54	2
3	Lithium	Recurrence	3,286	77	26	1
3	Imipramine	No Recurrence	206	90	48	1
3	Lithium	Recurrence	30	280	51	2
3	Placebo	Recurrence	7,143	167	35	2
3	Placebo	Recurrence	31	181	28	1
3	Placebo	Recurrence	17,286	399	23	1
3	Placebo	Recurrence	0,143	289	57	2
5	Lithium	Recurrence	3,286	182	47	1
5	Imipramine	No Recurrence	1,571	159	31	2
5	Lithium	Recurrence	19,714	122	27	1
5	Imipramine	No Recurrence	126,714	115	61	1
5	Placebo	Recurrence	8	343	60	1
5	Lithium	Recurrence	71,714	114	28	1
5	Placebo	Recurrence	63,714	249	36	1
5	Placebo	No Recurrence	96,286	140	29	1
5	Lithium	No Recurrence	50,857	110	34	1
5	Imipramine	No Recurrence	155	214	49	1
5	Imipramine	No Recurrence	39,571	224	45	1
5	Lithium	Recurrence	36,286	294	28	1

(*Continued*)

Table 7 Continued

Hospt	*Treat*	*Outcome*	*Time*	*AcuteT*	*Age*	*Gender*
5	Placebo	No Recurrence	102,571	162	24	2
5	Placebo	Recurrence	8,143	140	33	2
5	Imipramine	No Recurrence	28	147	34	1
5	Imipramine	No Recurrence	38	138	60	1
5	Imipramine	No Recurrence	111,571	196	23	2
5	Lithium	No Recurrence	165	139	35	1
5	Placebo	Recurrence	16	246	45	1
5	Lithium	No Recurrence	124,571	105	46	1
5	Lithium	No Recurrence	68	160	38	2
5	Placebo	No Recurrence	39,571	146	32	2
5	Placebo	No Recurrence	131	187	33	1
5	Imipramine	Recurrence	3,429	372	52	1
5	Lithium	No Recurrence	42	146	50	2
5	Imipramine	No Recurrence	26	131	38	1
5	Lithium	Recurrence	37,857	237	47	1
5	Imipramine	Recurrence	92,714	105	23	1
5	Imipramine	No Recurrence	106,714	140	31	1
5	Placebo	Recurrence	11,143	136	55	1
5	Placebo	No Recurrence	115	147	39	1
5	Placebo	Recurrence	44	160	41	1
5	Imipramine	No Recurrence	75	175	62	2
5	Placebo	No Recurrence	77,857	261	50	2
5	Lithium	Recurrence	0,286	146	46	1
5	Imipramine	No Recurrence	86	195	33	2
5	Placebo	No Recurrence	12,429	476	22	1
5	Lithium	No Recurrence	22	441	37	2
6	Lithium	Recurrence	5,429	86	40	2
6	Lithium	No Recurrence	67	201	22	1
6	Imipramine	Recurrence	3,429	130	30	2
6	Lithium	Recurrence	6,286	86	63	2
6	Imipramine	No Recurrence	5	209	40	1
6	Lithium	Recurrence	5,286	214	23	1
6	Imipramine	Recurrence	1	72	52	1
6	Placebo	Recurrence	3,429	238	23	1
6	Placebo	Recurrence	6,571	133	22	2
6	Placebo	Recurrence	1	128	23	1
6	Placebo	No Recurrence	45	139	30	2
6	Imipramine	No Recurrence	109,571	148	26	2
6	Lithium	Recurrence	0,857	285	46	1
6	Placebo	Recurrence	4,714	141	61	1
6	Imipramine	Recurrence	0,571	212	30	2
6	Imipramine	No Recurrence	9,143	168	39	1
6	Imipramine	No Recurrence	102	305	49	1
6	Lithium	Recurrence	46,286	204	57	1
6	Lithium	Recurrence	0,571	140	51	1
6	Lithium	Recurrence	6,429	182	53	1
6	Placebo	Recurrence	0	162	31	1
6	Placebo	Recurrence	20,857	207	43	1
6	Placebo	Recurrence	18,286	102	29	1
6	Imipramine	Recurrence	31,857	154	28	1
6	Imipramine	Recurrence	22	203	51	1
6	Lithium	Recurrence	2	176	33	1

2. H_1:$M_X > M_Y$. Reject H_0 if the computed T is less than $w_{1-\alpha}$, where $w_{1-\alpha} = nm - w_\alpha$ is the tabulated critical value for n, the number of X observations; m, the number of Y observations; and α, the chosen level of significance.
3. H_1:$M_X < M_Y$. Reject H_0 if the computed T is less than w_α, where w_α is the tabulated critical value of T for n, the number of X observations; m, the number of Y observations; and α, the chosen level of significance.

When either n or m is greater than 20 we compute the following test statistic

$$z = \frac{T - nm/2}{\sqrt{nm(n+m+1)/12}} \sim N(0,1)$$

and compare the result, for significance, with critical values of the standard normal distribution. Finally, many computer packages give the test value of both the Mann–Whitney test (U) and the Wilcoxon test (W). These two tests are algebraically equivalent tests, and are related by the following equality when there are no ties in the data:

$$U + W = \frac{m(m + 2n + 1)}{2}$$

Case Studies

In this section, we consider two datasets as case studies. The first is the Clinical depression dataset downloaded from http://bolt.mph.ufl.edu (for instance, see **Table 7**). The depression is the most common mental illness in the United States, affecting 19 million adults each year (Source: NIMH, 1999). Nearly 50% of individuals who experience a major episode will have a recurrence within 2–3 years. In a study conducted by the National Institutes of Health, 109 clinically depressed patients were separated into three groups, and each group was given one of two active drugs (imipramine or lithium) or no drug at all. For each patient, the dataset contains the following characteristic or variables:
Hospt: The patient's hospital, represented by a code for each of the 5 hospitals (1, 2, 3, 5, or 6).
Treat: The treatment received by the patient (Lithium, Imipramine, or Placebo).
Outcome: Whether or not a recurrence occurred during the patient's treatment (Recurrence or No Recurrence).
Time: Either the time (days) till recurrence, or if no recurrence, the length (days) of the patient's participation in the study.
AcuteT: The time (days) that the patient was depressed prior to the study.
Age: The age of the patient in years, when the patient entered the study.
Gender: The patient's gender (1 = Female, 2 = Male).

Using these data, researchers are interested in comparing therapeutic solutions that could delay or reduce the incidence of recurrence.

In the second dataset a researcher designed an experiment to assess the effects of prolonged inhalation of cadmium oxide. Fifteen laboratory animals served as experimental subjects, while 10 similar animals served as controls. The variable of interest was hemoglobin level following the experiment. The results are shown in **Table 8**. We wish to know if we can conclude that prolonged inhalation of cadmium oxide reduces hemoglobin level.

Results

In this section, we describe the main results obtained from the descriptive and inferential analysis using the information contained **Tables 7** and **8**.

The first dataset (Clinical depression dataset) is composed by four categorical variables (Hospt, Treat, Outcome, Gender) and two numerical variables (Time, Time, AcuteT). Bar plots for Hospt, Outcome, and Gender variables are plotted for a qualitative analysis of these data (see **Fig. 18**). On the contrary, a quantitative descriptive analysis is performed for the variable age of patients. **Table 9** summarizes the main results of some statistical measures computed using the formulas shown in **Table 3**. Finally, the confidence intervals (CI) and hypothesis test for the difference between two population means μ_1 and μ_2

Table 8 Hemoglobin determinations (grams) for 25 laboratory animals

Exposed animals	*Unexposed animals*
14.4	17.4
14.2	16.2
13.8	17.1
16.5	17.5
14.1	15.0
16.6	16.0
15.9	16.9
15.6	15.0
14.1	16.3
15.3	16.8
15.7	–
16.7	–
13.7	–
15.3	–
14.0	–

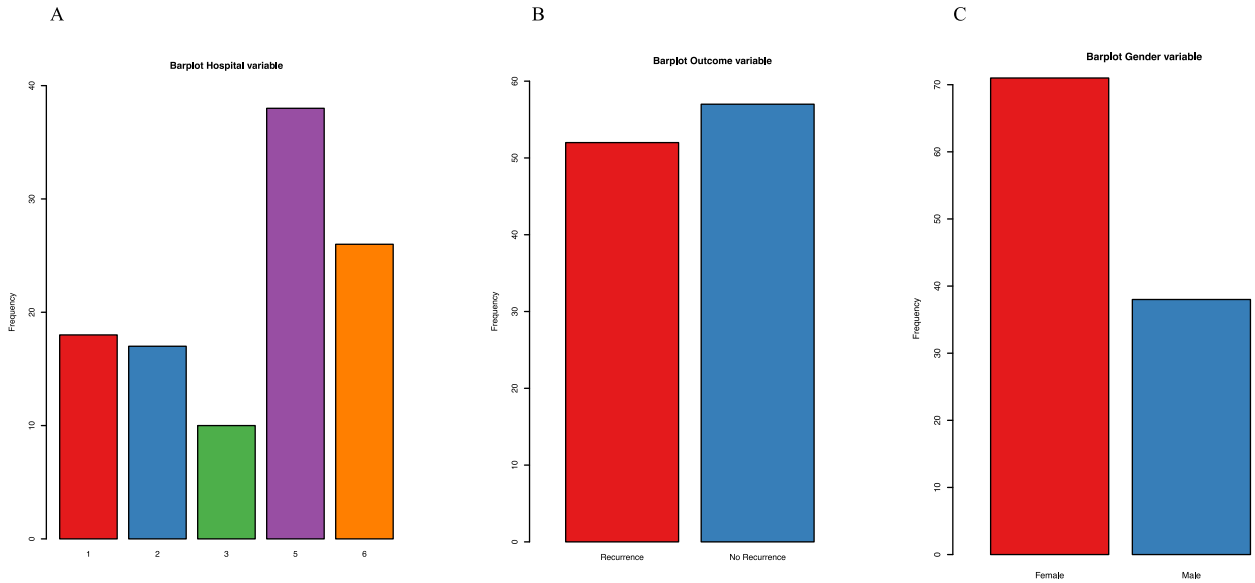

Fig. 18 Bar plots of the qualitative variables: Host (A), Outcome (B) and Gender (B). The first plot indicates that the number of patients from hospital 5 is greater than the others. The second plot shows that the outcome of the treatment for patients with no recurrence exceeds that with recurrence. The third plot displays that the number of female patients is greater than that of males.

Table 9 Data synthesis of patients grouped by age

Class intervals	n_i	f_i	$f_i(\%)$	F_i	c_i	$c_i\, n_i$	$c_i^2\, n_i$	
19 – 25	16	0.15	15	0.15	22	352	7,744	
25 – 31	21	0.20	20	0.34	28	588	16,464	Mode class
31 – 37	14	0.13	13	0.47	34	476	16,184	
37 – 43	11	0.10	10	0.57	40	440	17,600	Median class
43 – 49	15	0.14	14	0.70	46	690	31,740	
49 – 55	15	0.14	14	0.84	52	780	40,560	
55 – 71	17	0.15	15	1	63	1,071	67,473	
Total	109	1	100			4,397	197,765	

Mean	*Median*	*Mode*	*Variance*	*Standard deviation*	*First quartile*	*Third quartile*	*IQR*
40.34	38.8	28	187.04	13.68	28.16	51.14	22.98

are detected using the second listed test statistic illustrated in **Table 6**. In particular, we test the effect of the treatment (Imipramine) with respect to the control group (Placebo) during the participation of patients in the study (Time). The sample mean estimates are 37.58 and 63.06 for each group under investigated ($n=34$-Imipramine and $m=38$-Placebo). The test statistic (two-tailed test) is $t=-2.38$ with 70 degree of freedom, the confidence interval at level $\alpha=0.05$ is (–46.84,–4.12) and the p-value is significant (p-value$=0.0201<0.05$). Hence, the null hypothesis H_0 is rejected which means that the true difference in means is not equal to zero. This means that there is an evidence on the effects of therapy in the treatment of patients with Imipramine during the study.

In the second dataset, we consider the hemoglobin levels (measured in grams) for 25 laboratory animals, divided in two groups: exposed (X) and not exposed (Y) to cadmium oxide. We assume that the assumptions of the Mann–Whitney test are applicable. Therefore, with $n=15$, $m=10$ and $\alpha=0.05$, we find the statistic test (two-tailed test) $T=25$ and the p-value equal to 0.006008 (p-value <0.05, statistically significant). We conclude that M_X is smaller than M_Y. This leads to the conclusion that prolonged inhalation of cadmium oxide does reduce the hemoglobin level.

Software

We use the R statistical software (see Relevant Websites section) to plot the graphs and to perform the descriptive statistics and statistical inference. In particular, we apply the common used statistical packages in R.

Conclusions

Biostatistics can be defined as the application of mathematics used in statistics to the fields of biological sciences and medicine. When research activities involve data collection on a sample of a population, an understanding of descriptive and inferential analysis become essential for an accurate study of the phenomenon to draw conclusions and make inferences about the entire population. The two major areas of statistics are the descriptive statistics and the inferential statistics. The aim of the first areas is to collect data and obtain a synthesis of this information in order to give a descriptive overview of the data. On the other hand, the goal of the statistical inference is to decide whether the findings of an investigation reflect chance or real effects at a given level of probability. Both estimation and testing hypothesis are covered. These statistical tools are useful for researchers in order to decide what type of study to use for their research project, how to execute the study on patients and well people, and how to evaluate the final results.

See also: Natural Language Processing Approaches in Bioinformatics

References

Altman, N., Krzywinski, M., 2017. Points of significance: P values and the search for significance. Nature Methods 14 (1), 3–4.
Cochran, W.G., 1964. Approximate significance levels of the Behrens–Fisher test. Biometrics 20, 191–195.
Gardner, M.J., Altman, D.G., 1986. Confidence intervals rather than P values: Estimation rather than hypothesis testing. British Medical Journal (Clinical Research Edition) 292 (6522), 746–750.
Manikandan, S., 2011a. Measures of central tendency: The mean. Journal of Pharmacology and Pharmacotherapeutics 2 (2), 140.
Manikandan, S., 2011b. Measures of central tendency: Median and mode. Journal of Pharmacology and Pharmacotherapeutics 2 (3), 214.
Manikandan, S., 2011c. Measures of dispersion. Journal of Pharmacology and Pharmacotherapeutics 2 (4), 315–316.
Mann, H.B., Whitney, D.R., 1947. On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics 18, 50–60.
Spriestersbach, A., *et al*., 2009. Descriptive statistics: The specification of statistical measures and their presentation in tables and graphs. Part 7 of a series on evaluation of scientific publications. Deutsches Ärzteblatt International 106 (36), 578–583.
Wilcox, R.R., Keselman, H.J., 2003. Modern robust data analysis methods: Measures of central tendency. Psychological Methods 8 (3), 254.
Wilcoxon, F., 1945. Individual comparisons by Ranking methods. Biometrics 1, 80–83.

Further Reading

Daniel, W.W., Cross, C.L., 2013. Biostatistics: A Foundation for analysis in the Health Sciences, tenth ed. John Wiley & Sons.
Dehmer, M., Emmert-Streib, F., Graber, A., Salvador, A., 2011. Applied Statistics for Network Biology: Methods in Systems Biology. Wiley-Blackwell.
Dunn, O.J., Clark, V.A., 2009. Basic Statistics: A Primer for the Biomedical Sciences. John Wiley & Sons.
Heumann, C., Schomaker, M., 2016. Introduction to Statistics and Data Analysis. With Exercises, Solutions and Applications in R. Springer.
Hoffman, J.I.E., 2015. Biostatistics for Medical and Biomedical Practitioners. Elsevier.
Indrayan, A., Malhotra, R.K., 2017. Medical Biostatistics, fourth ed. Chapman and Hall/CRC.

Relevant Websites

https://www.r-project.org
The R Project for Statistical Computing.
http://bolt.mph.ufl.edu
UF Health: Biostatistics.