STATISTICAL METHODS

Arnaud Delorme, Swartz Center for Computational Neuroscience, INC, University of San Diego California, CA92093-0961, La Jolla, USA. Email: arno@salk.edu.

Keywords: statistical methods, inference, models, clinical, software, bootstrap, resampling, PCA, ICA

Abstract: Statistics represents that body of methods by which characteristics of a population are inferred through observations made in a representative sample from that population. Since scientists rarely observe entire populations, sampling and statistical inference are essential. This article first discusses some general principles for the planning of experiments and data visualization. Then, a strong emphasis is put on the choice of appropriate standard statistical models and methods of statistical inference. (1) Standard models (binomial, Poisson, normal) are described. Application of these models to confidence interval estimation and parametric hypothesis testing are also described, including two-sample situations when the purpose is to compare two (or more) populations with respect to their means or variances. (2) Non-parametric inference tests are also described in cases where the data sample distribution is not compatible with standard parametric distributions. (3) Resampling methods using many randomly computer-generated samples are finally introduced for estimating characteristics of a distribution and for statistical inference. The following section deals with methods for processing multivariate data. Methods for dealing with clinical trials are also briefly reviewed. Finally, a last section discusses statistical computer software and guides the reader through a collection of bibliographic references adapted to different levels of expertise and topics.

Statistics can be called that body of analytical and computational methods by which characteristics of a population are inferred through observations made in a representative sample from that population. Since scientists rarely observe entire populations, sampling and statistical inference are essential. Although, the objective of statistical methods is to make the process of scientific research as efficient and productive as possible, many scientists and engineers have inadequate training in experimental design and in the proper selection of statistical analyses for experimentally acquired data. John L. Gill [1] states: "…statistical analysis too often has meant the manipulation of ambiguous data by means of dubious methods to solve a problem that has not been defined." The purpose of this article is to provide readers with definitions and examples of widely used concepts in statistics. This article first discusses some general principles for the planning of experiments and data visualization. Then, since we expect that most readers are not studying this article to learn statistics but instead to find practical methods for analyzing data, a strong emphasis has been put on choice of appropriate standard statistical model and statistical inference methods (parametric, non-parametric, resampling methods) for different types of data. Then, methods for processing multivariate data are briefly reviewed. The section following it deals with clinical trials. Finally, the last section discusses computer software and guides the reader through a collection of bibliographic references adapted to different levels of expertise and topics.

DATA SAMPLE AND EXPERIMENTAL DESIGN

Any experimental or observational investigation is motivated by a general problem that can be tackled by answering specific questions. Associated with the general problem will be a population. For example, the population can be all human beings. The problem may be to estimate the probability by age bracket for someone to develop lung cancer. Another population may be the full range of responses of a medical device to measure heart pressure and the problem may be to model the noise behavior of this apparatus.

Often, experiments aim at comparing two sub-populations and determining if there is a (significant) difference between them. For example, we may compare the frequency occurrence of lung cancer of smokers compared to non-smokers or we may compare the signal to noise ratio generated by two brands of medical devices and determine which brand outperforms the other with respect to this measure.

How can representative samples be chosen from such populations? Guided by the list of specific questions, samples will be drawn from specified sub-populations. For example, the study plan might specify that 1000 presently cancer-free persons will be drawn from the greater Los Angeles area. These 1000 persons would be composed of random samples of specified sizes of smokers and non-smokers of varying ages and occupations. Thus, the description of the sampling plan will imply to some extent the nature of the target sub-population, in this case smoking individuals.

Choosing a random sample may not be easy and there are two types of errors associated with choosing representative samples: *sampling errors* and *non-sampling errors*. *Sampling errors* are those errors due to chance variations resulting from sampling a population. For example, in a population of 100,000 individuals, suppose that 100 have a certain genetic trait and in a (random) sample of 10,000, 8 have the trait. The experimenter will estimate that 8/10,000 of the population or 80/100,000 individuals have the trait, and in doing so will have underestimated the actual percentage. Imagine conducting this experiment (i.e., drawing a random sample of 10,000 and examining for the trait) repeatedly. The observed number of sampled individuals having the trait will fluctuate. This phenomenon is called the *sampling error*. Indeed, if sampling is truly random, the observed number having the trait in each repetition will fluctuate "randomly" about 10. Furthermore, the limits within which most fluctuations will occur are estimable using standard statistical methods. Consequently, the experimenter not only acknowledges the presence of *sampling errors*, but he can estimate their effect.

In contrast, variation associated with improper sampling is called *non-sampling error*. For example, the entire target population may not be accessible to the experimenter for the purpose of choosing a sample. The results of the analysis will be biased if the accessible and non-accessible portions of the population are different with respect to the characteristic(s) being investigated. Increasing sample size within the accessible portion will not solve the problem. The sample, although random within the accessible portion, will not be "representative" of the target population. The experimenter is often not aware of the presence of *non-sampling errors* (e.g., in the above context, the experimenter may not be aware that the trait occurs with higher frequency in a particular ethnic group that is less accessible to sampling than other groups within the population). Furthermore, even when a source of *non-sampling error* is identified, there may not be a practical way of assessing its effect. The only recourse when a source of *non-sampling error* is identified is to document its nature as thoroughly as possible. Clinical trials involving survival studies are often associated with specific non-sampling errors (see the section dealing with clinical trials below).

DESCRIPTIVE STATISTICS

Descriptive statistics are tabular, graphical, and numerical methods by which essential features of a sample can be described. Although these same methods can be used to describe entire populations, they are more often applied to samples in order to capture population characteristics by inference.

We will differentiate between two main types of data samples: qualitative data samples and quantitative data samples. Qualitative data arises when the characteristic being observed is not measurable. A typical case is the "success" or "failure" of a particular test. For example, to test the effect of a drug in a clinical trial setting, the experimenter may define two possible outcomes for each patient: either the drug was effective in treating the patient, or the drug was not effective. In the case of two possible outcomes, any sample of size n can be represented as a sequence of n nominal outcome x_1, x_2,…, x_n that can assume either the value "success" or "failure".

By contrast, quantitative data arise when the characteristics being observed can be described by numbers. Discrete quantitative data is countable whereas continuous data may assume any value, apart from any precision constraint imposed by the measuring instrument. Discrete quantitative data may be obtained by counting the number of each possible outcome from a qualitative data sample. Examples of discrete data may be the number of subjects sensitive to the effect of a drug (number of "success" and number of "failure"). Examples continuous data are weight, height, pressure, and survival time. Thus, any quantitative data sample of size n may be represented as a sequence of n numbers x_1, x_2, …, x_n and sample statistics are functions of these numbers.

Satisfaction rank	Number of responses
0	38
1	144
2	342
3	287
4	164
5	25
Total	1000

Table 1. Result of a hearing aid device satisfaction survey in 1000 patients showing the frequency distribution of each response.

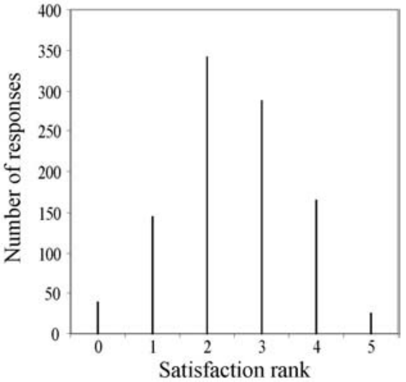

Fig. 1. Frequency histogram for the hearing aid device satisfaction survey of Table 1.

Discrete data may be preprocessed using frequency tables and represented using histograms. This is best illustrated by an example. For discrete data, consider a survey in which 1000 patients fill in a questionnaire for assessing the quality of a hearing aid device. Each patient has to rank product satisfaction from 0 to 5, each rank being associated with a detailed description of hearing quality. Table 1 represents the frequency of each response type. A graphical equivalent is the frequency histogram illustrated in Fig. 1. In the histogram, the heights of the bars are the frequencies of each response type. The histogram is a powerful visual aid to obtain a general picture of the data distribution. In Fig. 1, we notice a majority of answers corresponding to response type "2" and a 10-fold frequency *drop* for response types "0" and "5" compared to response type "2".

For continuous data, consider the data sample in Table 2, which represents amounts of infant serum calcium in mg/100 ml for a random sample of 75 week-old infants whose mothers received vitamin D supplements during pregnancy. Little information is conveyed by the list of numbers. To depict the central tendency and variability of the data, Table 3 groups the data into six classes, each of width 0.03 mg/100 ml. The "frequency" column in Table 3 gives the number of sample values occurring in each class. The picture given by the frequency distribution Table 3 is a clearer representation of central tendency and variability of the data than that presented by Table 2. In Table 3, data are grouped in six classes of equal size and it is possible to see the "centering" of the data about the 9.325–9.355 class and its variability—the measurements vary from 9.27 to 9.44 with about 95% of them between 9.29 and 9.41. The advantage of grouped frequency distributions is that grouping smoothes the data so that essential features are more discernible. Fig. 2 represents the corresponding

9.37	9.34	9.38	9.32	9.33	9.28	9.34
9.29	9.36	9.30	9.31	9.33	9.34	9.35
9.35	9.36	9.30	9.32	9.33	9.35	9.36
9.32	9.37	9.34	9.38	9.36	9.37	9.36
9.36	9.33	9.34	9.37	9.44	9.32	9.36
9.38	9.39	9.34	9.32	9.30	9.30	9.36
9.29	9.41	9.27	9.36	9.41	9.37	9.31
9.31	9.33	9.35	9.34	9.35	9.34	9.38
9.40	9.35	9.37	9.35	9.32	9.36	9.35
9.35	9.36	9.39	9.31	9.31	9.30	
9.31	9.36	9.34	9.31	9.32	9.34	

Table 2. Serum calcium (mg/100 ml) in a random sample of 75 week-old infants whose mother received vitamin D supplement during pregnancy.

Serum calcium (mg/100 mL)	Frequency
9.265–9.295	4
9.295–9.325	18
9.325–9.355	24
9.355–9.385	22
9.385–9.415	6
9.415–9.445	1
Total	75

Table 3. Frequency distribution of infant serum calcium data.

histogram. The sides of the bars of the histogram are drawn at the class boundaries and their heights are the frequencies or the relative frequencies (frequency/sample size). In the histogram, we clearly see that the distribution of the data centered about the point 9.34. Although grouping smoothes the data, too much grouping (that is choosing too few classes) will tend to mask rather than enhance the sample's essential features.

There are many numerical indicators for summarizing and describing data. The most common ones indicate central tendency, variability, and proportional representation (the sample mean, variance, and percentiles, respectively). We shall assume that any characteristic of interest in a population, and hence in a sample, can be represented by a number. This is obvious for measurements and counts, but even qualitative characteristics (described by discrete variables) can be numerically represented. For example, if a population is dichotomized into those individuals who are carriers of a particular disease and those who are not, a 1 can be assigned to each carrier and a 0 to each non-carrier. The sample can then be represented by a sequence of 0s and 1s.

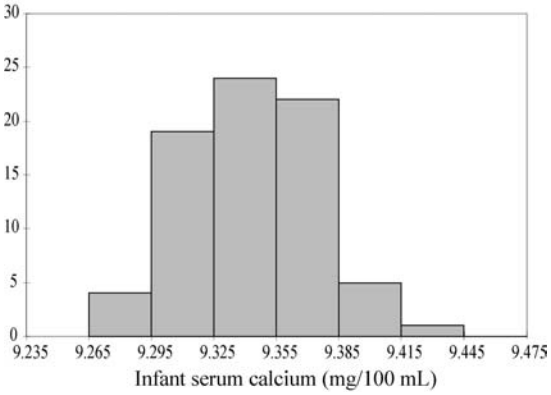

Fig. 2. Frequency histogram of infant serum calcium data of Table 2 and 3. The curve on the top of the histogram is another representation of probability density for continuous data.

The most common measure of central tendency is the sample mean:

$$M = (x_1 + x_2 + ... + x_n)/n \qquad \text{also noted } \overline{X} \tag{1}$$

where $x_1, x_2, ..., x_n$ is the collection of numbers from a sample of size n. The sample mean can be roughly visualized as the abscissa of the horizontal center of gravity of the frequency histogram. For the serum calcium data of Table 2, M=9.34 which happens to be the midpoint of the highest bar of the histogram (Fig. 2). This histogram is roughly symmetric about a vertical line drawn through M but this is not necessarily true of all histograms. Histograms of counts and survival times data are often skewed to the right (long-tailed with concentrated "mass" at the lower values). Consequently, the idea of M as a center of gravity is important to bear in mind when using it to indicate central tendency. For example, the median (described later in this section) may be a more appropriate index of centrality depending on the type of data and the kind of information one wishes to convey.

The sample variance, defined by

$$s^2 = \frac{1}{n-1}\left[(x_1 - M)^2 + (x_2 - M)^2 + ... + (x_n - M)^2\right] = \sum_{i=1}^{n} \frac{(x_i - M)^2}{n-1} \tag{2}$$

is a measure of variability or dispersion of the data. As such it can be motivated as follows: x_i-M is the deviation of the ith data sample from the sample mean, that is, from the "center" of the data; we are interested in the amount of deviation, not its direction, so we disregard the sign by calculating the squared deviation $(x_i$-$M)^2$; finally, we "average" the squared deviations by summing them and dividing by the sample size minus 1. (Division by n – 1 ensures that the sample variance is an unbiased estimate of the population variance.) Note that an equivalent and often more practical formula for computing the variance may be obtained by developing Equation (2):

$$s^2 = \frac{\sum x_i^2 - nM^2}{n-1} \tag{3}$$

A measure of variability in the original units is then obtained by taking the square root of the sample variance. Specifically, the sample standard deviation, denoted s, is the square root of the sample variance.

For the serum calcium data of Table 2, s^2 = 0.0010 and s = 0.03 mg/100 ml. The reader might wonder how the number 0.03 gives an indication of variability. Note that for the serum calcium data M±s=9.34±0.03 contains 73% of the data, M±2s=9.34±0.06 contains 95% and M±3s=9.34±0.09 contains 99%. It can be shown that the interval M±$3s$ will include at least 89% of any set of data (irrespective of the data distribution).

An alternative measure of central tendency is the median value of a data sample. The median is essentially the sample value at the middle of the list of sorted sample values. We say "essentially" because a particular sample may have no such value. In an odd-numbered sample, the median is the middle value; in an even-numbered sample, where there is no middle value, it is conventional to take the average of the two middle values. For the serum calcium data of Table 3, the median is equal to 9.34.

By extension to the median, the sample p percentile (say 25th percentile for example) is the sample value at or below which p% (25%) of the sample values lie. If there is no value at a specific percentile, the average between the upper and lower closest existing round percentile is used. Knowledge of a few sample percentiles can provide important information about the population.

For skewed frequency distributions, the median may be more informative for assessing a population "center" than the mean. Similarly, an alternative to the standard deviation is the interquartile range: it is defined as the 75th minus the 25th percentiles and is a variability index not as influenced by outliers as the standard deviation.

There are many other descriptive and numerical methods (see for instance [2]). It should be emphasized that the purpose of these methods is usually not to study the data sample itself but rather to infer a picture of the population from which the sample is taken. In the next section, standard population distributions and their associated statistics are described.

PROBABILITY, RANDOM VARIABLES, AND PROBABILITY DISTRIBUTIONS

The foundation of all statistical methodology is probability theory, which progresses from elementary to the most advanced mathematics. Much of the misunderstanding and abuse of statistics comes from the lack of understanding of its probabilistic foundation. When assumptions of the underlying probabilistic (mathematical) model are grossly violated, derived inferential methods will lead to misleading and irrational conclusions. Here, we only discuss enough probability theory to provide a framework for this article.

In the rest of this article, we will study experiments that have more than one possible outcome, the actual outcome being determined by some chance mechanism. The set of possible outcomes of an experiment is called its sample space; subsets of the sample space are called events, and an event is said to occur if the actual outcome of the experiment is a member of that event. A simple example follows.

The experiment will be the toss of a pair of fair coins, arbitrarily labeled coin number 1 and coin number 2. The outcome (1,0) means that coin #1 shows a head and coin #2 shows a tail. We can then specify the sample space by the collection of all possible outcomes:

$$S = \{(0,0)\ (0,1)\ (1,0)\ (1,1)\}$$

There are 4 ordered pairs so there are 4 possible outcomes in this coin-tossing experiment. Consider the event A "toss one head and one tail," which can be represented by $A = \{(1,0)\ (0,1)\}$. If the actual outcome is (0,1) then the event A has occurred.

In the example above, the probability for event A to occur is obviously 50%. However, in most experiments it is not possible to intuitively estimate probabilities, so the next step in setting up a probabilistic framework for an experiment is to assign, through some mathematical model, a probability to each event in the sample space.

Definition of Probability

A probability measure is a rule, say P, which associates with each event contained in a sample space S a number such that the following properties are satisfied:

1: For any event, A, $P(A) \geq 0$.

2: $P(S) = 1$ (since S contains all the outcomes, S always occurs).

3: $P(not\ A)+P(A)=1$.

4: If A and B are mutually exclusive events (that cannot occur simultaneously) and independent events (that are not linked in any way), then

$$P(A \text{ or } B) = P(A) + P(B) \qquad \text{and}$$

$$P(A \text{ and } B) = 0$$

Many elementary probability theorems (rules) follow directly from these definitions.

Probability and relative frequency

The axiomatic definition above and its derived theorems dictate the properties that probability must satisfy, but they do not indicate how to assign probabilities to events. The major classical and cultural interpretation of probabilities is the relative frequency interpretation. Consider an experiment that is (at least conceptually) infinitely repeatable. Let A be any event and let n_A be the number of times the event A occurs in n repetitions of the experiment; then the relative frequency of occurrence of A in the n repetitions is n_A/n. For example, if mass production of a medical device reliably yields 7 malfunctioning devices out of 100, the relative frequency of occurrence of a defective device is 7/100.

The probability of A is defined by $P(A) = \lim n_A/n$ as $n \to \infty$, where this limit is assumed to exist. The number $P(A)$ can never be known, but if the experiment can in fact be repeated a "large" number of times, it can be estimated by the relative frequency of occurrence of A.

The relative frequency interpretation is an objective interpretation because the probability of an event is assumed to be independent of judgment by the observer. In the subjective interpretation of probability, a probability is assigned to an event according to the assigner's strength of belief that the event will occur, on a scale of 0 to 1. The "assigner" could be an expert in a specific field, for example, a cardiologist that provides the probability for a sample of electrocardiograms to be pathological.

Probability distribution definition and probability mass function

We have assumed that all data can be numerically represented. Thus, the outcome of an experiment in which one item will be randomly drawn from a population will be a number, but this number cannot be known in advance. Let the potential outcome of the experiment be denoted by X, which is called a random variable in statistics. When the item is drawn, X will be realized or observed. Although the numerical values that X will take cannot be known in advance, the random mechanism that governs the outcome can perhaps be described by a probability model. Using the model, we may calculate the probability that the random variable *X* will take a value within a set or range of numbers.

One such popular mathematical model is the probability distribution of a discrete random variable *X*. It can be best described as a mathematical equation or table that gives, for each value *x* that *X* can assume, the probability associated with this value $P(X = x)$. For instance, if *X* represents the outcome of the tossing of a coin, there are two possible outcomes, "tail" and "head". If it is a fair coin *P(X="tail")*=0.5 and *P(X="head")*=0.5. In statistics, the function $P(X = x)$ is called the probability mass function of *X*.

It follows from the relative frequency interpretation of probability that, for a discrete random variable or for the frequency distribution of a continuous variable, relative frequency histograms estimate the probability mass functions of this variable. For example, in Table 3, if the random variable *X* indicates the serum calcium measure, then

$$\widehat{P}(X\ is\ in\ the\ first\ bin) = \widehat{P}(9.265 \le X < 9.295) = 4/75$$

the ^ symbol on *P* indicating estimated probability values, since actual probabilities describe the population itself and cannot be calculated from data samples. Similarly the probability that *X* is in the 2nd bin, the 3rd bin, … can be estimated and the collection of these probabilities constitute an estimated probability mass function.

Probability density function for continuous variables

The probability mass function above best describes discrete events but what probabilities can we assign to continuous variables? Since a continuous variable *X* can assume any value on a continuum, the probability that *X* assumes a *particular* value is 0 (except in very particular cases that will not be discussed here). Consequently, associated with a continuous random variable *X*, is a function f_X, called its probability density function that can be used to compute probability. The probability that a continuous random variable *X* assumes a value between values x_1 and x_2 is the area under the graph of f_X over the interval x_1 and x_2; mathematically

$$P(x_1 \le X \le x_2) = \int_{x_1}^{x_2} f_X(x)\ dx \qquad (5)$$

For example, for the infant serum data of Table 2 (see also Table 3), we would estimate that the probability that an infant whose mother received vitamin D supplement during pregnancy has between 9.35 and 9.38 mg/100 ml calcium is 22/75 or 0.293, which is the relative frequency of the 9.355–9.385 class in the sample. For continuous data, a smooth curve passing through the midpoint of a histogram bars' upper limit should resemble the probability density function of the underlying population.

There are many mathematical models of probability distribution. Three of the most commonly used probability distribution models described below are the binomial distribution and the Poisson distribution for discrete variables, and the normal distribution for continuous variables.

The binomial distribution

The scenario leading to the binomial distribution is an experiment that consists of *n* independent, repeated trials, each of which can end in only one of two ways arbitrarily labeled "success" or "failure." The probability that any trial ends in a "success" is *p* (and hence $q = 1 - p$ for a "failure"). Let the random variable *X* denote the total number of successes in the *n* trials, and *x* denote a number in {0; …; n}. Under these assumptions:

$$P(X = x) = \binom{n}{x} p^x q^{n-x} \qquad x = 0,\ 1,\ ....\ n \qquad (6)$$

with

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \qquad (7)$$

where *n*!=1*2*3…*n is *n* factorial.

For example, suppose the proportion of carriers of an infectious disease in a large population is 10% ($p = 0.1$) and that the number of carriers follows a binomial distribution. If 20 individuals are sampled ($n = 20$) and *X* is the number of carriers ("successes") in the sample, then the probability that there will be exactly one carrier in the sample is

$$P(X = 1) = \binom{20}{1}(0.10)^1(0.90)^{20-1} = 0.27$$

More complex probabilities may be calculated with the help of probability rules and definitions. For instance the probability that there will be at least two carriers in the sample is

$$\begin{aligned} P(X \ge 2) &= 1 - P(X < 2) && \text{see 3rd probability definition} \\ &= 1 - P(X = 0\ or\ X = 1) \\ &= 1 - \left(P(X = 0) + P(X = 1)\right) && \text{see 4th probability definition} \\ &= 1 - \binom{20}{0}(0.10)^0(0.90)^{20} - \binom{20}{1}(0.10)^1(0.90)^{19} \\ &= 1 - 0.12 - 0.27 = 0.61 \end{aligned}$$

Historically, single trials of a binomial distribution are called Bernoulli variates after the Swiss mathematician James Bernoulli who discovered it at the end of the seventeenth century.

The Poisson distribution

The Poisson distribution is often used to represent the number of successive independent events of a specified type (for example cases of flu) with low probability of occurrence (less than 10%) in some specified interval of time or space. The Poisson distribution is also often used to represent the number of occurrence of events of a specified type where there is no natural upper limit, for example the number of radioactive particles emitted by a sample over a set time period. Specifically, *X* is a Poisson random variable if it obeys the following formula:

$$P(X = x) = e^{-\lambda}\lambda^x / x! \qquad x = 0,\ 1,\ 2,\ \ldots \qquad (8)$$

where e = 2.178…is the natural logarithmic base and λ is a given constant. For example, suppose the number of a particular type of bacteria in a standard area (e.g., 1 cm^2) can be described by a Poisson distribution with parameter λ = 5. Then, the probability that there are no more than 3 bacteria in the standard area is given by

$$P(X \leq 3) = P(X=0)+P(X=1)+P(X=2)+P(X=3)$$
$$= e^{-5}\,5^0/0!+e^{-5}\,5^1/1!+e^{-5}\,5^2/2!+e^{-5}\,5^3/3! = 0.265$$

Note that the Poisson and the binomial distributions are closely related. In the case of a rare event (p<10%), the binomial distribution (described by probability p and n events) is well approximated by the Poisson distribution with the constant $\lambda=np$. The Poisson distribution was named after the French mathematician Siméon-Denis Poisson, who discovered it in the early part of the nineteenth century.

The normal distribution

The binomial and Poisson distributions describe discrete events but there are also many distributions describing continuous variables. The most important one is the normal distribution (also called Laplace-Gauss distribution as it was discovered by the French astronomer Pierre-Simon Laplace and the German mathematician Karl Friedrich Gauss in the early nineteenth century). Normal distributions arise as a result of many small random fluctuations about some general average (for example, repeated recordings of a constant body temperature using a noisy electronic thermometer). A random variable X is said to be a normal or Gaussian random variable with mean parameter μ and standard deviation parameter σ if its probability density function is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty \tag{9}$$

The normal probability density function graphed in Fig. 3, is bell shaped with tails rather rapidly receding to zero height. Because f_X represents probability density, the total area bounded by the curve is 1 (see Equation (9)). The area between two values of variable X (x_1 and x_2 where x_1<x_2) represents the probability that X lies between x_1 and x_2 (Equation (5)).

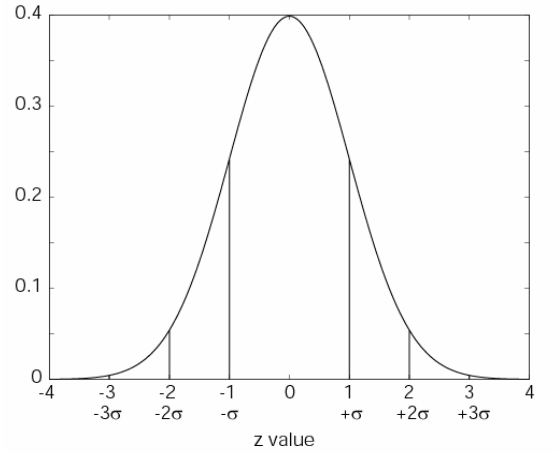

Fig. 3. The normal probability density function showing symmetry about a vertical line through μ and the role of σ as a variability parameter. Vertical bars indicates $\pm\sigma$, $\pm 2\sigma$, $\pm 3\sigma$.

As shown in Fig. 3, if X is normal (μ, σ), it can be calculated that $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.997$, which, according to the relative frequency interpretation of probability, states that about 99.7% of a large sample from a "normally distributed population" will be contained in the interval mean plus or minus three standard deviations ($\mu \pm 3\sigma$).

Note that there is a relation between the normal and the binomial distribution. Using the same notation as in Equation (6), if n, the number of samples, is large enough then the variable z defined as

$$z = \frac{x - np}{\sqrt{npq}} \tag{10}$$

is approximately normally distributed with mean 0 and standard deviation 1. In a coin throwing experiment, throwing the coin a large number of times and counting the number of heads x, then building a histogram for the value z, the histogram will be close to a normal distribution (as shown in Fig. 3). Similarly, there is a relation between the Poisson and the normal distribution, the variable z defined as $z=(x-\lambda)/\lambda$ is normally distributed for large values of λ.

Many statistical inferential methods described in the next section assume that the data is approximately normally distributed. Much abuse occurs, however, when these methods are applied blindly with no verification of the normality assumption. Incidentally, methods that incorporate assumptions of normality often can be applied to non-normal situations because under certain conditions, the normal distribution can approximate other distributions, such as the binomial and the Poisson distributions. Sometimes, the data can also be preprocessed to fit the normal distribution. For example, a histogram might indicate non-normality, while a histogram of the logarithms of the data would fit the normal distribution, indicating that normal-based models can be applied to the log-transformed data. These transformations are discussed in most experimental design textbooks.

The importance of the normal distribution in statistics is also due to the central limit theorem in statistics which states that the distribution of any linear mixture of two or more independent random variables is more normal (has a shape closer to the normal distribution) than the distribution of the random variables themselves. This property is used by some algorithms processing multivariate data (as described in a later section).

There are many other continuous probability distributions besides the normal distribution. For example, the most commonly used distribution in survival analysis is the Weibull distribution. The von Mises distribution allows parametric statistical tests for periodic data (i.e., seasonal).

Characteristics of probability distributions

Just as there are numerical indexes for sample description, for example, sample means, variances, and percentiles, there are numerical characteristics of probability distributions. The expectation or mean (not sample mean) of an random variable X is

$$E(X) \equiv \sum_x xP(X = x) \qquad X \text{ discrete}$$
$$\equiv \int_{-\infty}^{\infty} xf_X(x)\,dx \qquad X \text{ continuous} \tag{11}$$

The expectation E is a measure of central tendency for a population (i.e., the center of gravity of the probability distribution about the y axis). The variance of X is defined in terms of expectation by

$$\text{Var}(X) \equiv E\{[X - E(X)]^2\} \tag{12}$$

In words, Var(X) is the expected squared deviation of X from $E(X)$, and in this sense is a measure of variability or dispersion for a population. The standard deviation of X is the square root of its variance. Table 4 indicates mean and variance for the binomial, the Poisson, and the normal distribution.

	Binomial	Poisson	Normal
Mean	$\mu = Np$	$\mu = \lambda$	μ
Variance	$\sigma^2 = Npq$	$\sigma^2 = \lambda$	σ^2

Table 4. Mean and variance for standard distributions (see text for details).

Numerical descriptors of populations are often the very things we want to know about populations. They should not be confused with their sample counterparts; the sample numerical descriptors are the basis for drawing inferences regarding their population counterparts, which are of primary interest.

STATISTICAL INFERENCE

A statistical hypothesis is a statement about the probability distribution of populations using one or more data samples. Typical questions are "is this single data sample consistent with this theoretical distribution of values?", "are these two data samples originating from the same population?", "are these n data samples originating from the same population?". Associated with each of these questions, in statistics, two hypotheses are usually formulated.

Hypothesis H_0: All data samples originate from the same population (or the single data sample is consistent with a given theoretical distribution).

Hypothesis H_1: Some data samples do not originate from the same population (or the single data sample is not consistent with the given theoretical distribution).

The test is called significant if we reject hypothesis H_0 with respect to a user-defined confidence interval (for instance 5% of chance of wrongly rejecting H_0). It is important to remember that inference tests can never disprove hypothesis H_0. Instead, based on the significance threshold and on the inference test we have chosen, we can say that the data support rejecting H_0. The test is called non-significant if we accept hypothesis H_0 and reject hypothesis H_1. Accepting H_0 means that we failed to find any significant difference with respect to our user-defined confidence interval. Because the error in accepting H_0 is usually large (see error types below), in general we should avoid drawing any conclusion about the experiment when accepting H_0.

Degree of freedom: Elementary tests usually depend on the data sample size as well as the number of parameters, (e.g. mean or variance) that have to be estimated from the sample, to run the test. Specifically, the number of degrees of freedom of a statistics is defined as the number of independent observations minus the number of population parameters which must be estimated from sample observations. Details will be provided for each test.

p-values: Once hypotheses H_0 and H_1 have been defined, that a test has been chosen to address these hypotheses (see below), and that parameters for this test have been calculated, one must choose a level of significance. $p<0.05$ is the arbitrary value that is generally accepted to be significant. It means that there must be less than a 5% possibility of falsely detecting a significant difference. We will now describe how the p value relates to the different types of errors associated with elementary tests.

Type I and type II errors: If we reject a hypothesis H_0 when it should be accepted, we say that a type I error has been made. If we accept a hypothesis H_0 when it should be rejected we say that a type II error has been made. In either case a wrong decision or judgment has occurred. This is not a simple matter because decreasing one error type usually leads to increasing the other error type. One way of getting around this problem is just to set your significance level at .05 (and not at .01 or .001). In this way you are balancing between type I and type II errors in your decision making process. One way to decrease both error types is to increase the size of the sample. However, two ways of analyzing the same size dataset (i.e. two types of inference test) might have different efficiency, so that the more efficient might give better performance on both error types. As an example of type I and type II errors, let's imagine that there is a significant difference between the average of blood pressure measured from a population of patients and the general population at p=0.05. Then there will be a 5% chance that our statement is false (type I error). This means that if we repeat the test 100 times, when in fact no real effects are present, we will draw a wrong conclusion about 5% of the time that we observe a significant difference. In contrast, if we state that there is no such difference between population of patients at p=0.05, there is not a 5% chance of being wrong but usually more (type II error). This is why, in general, when accepting hypothesis H_0, we should not draw any conclusion about the results of an experiment. The exact calculation of type II error usually depends on the size of the actual effect in the population, hence it is usually described by curves as a function of effect magnitude.

Correction for multiple comparisons: When multiple tests are performed, the probability that one of them is significant by chance becomes larger. As for type I error, if 100 tests are performed with significance threshold of p=0.05, when in fact no real effects are present, then on average about 5 of them will indicate significance, but will be false positives. This is the case for instance when processing biophysical images such as magnetic resonance imaging data: a collection of values is acquired for each coordinate on a 3-D grid and a statistical test must be performed on this data. The same problem may arise

Goal	Dataset		
	Binomial or Discrete	**Continuous measurement (from a normal distribution)**	**Continuous measurement, Rank, or Score (from non-normal distribution)**
Example of data sample	List of patients recovering or not after a treatment	Readings of heart pressure from several patients	Ranking of several treatment efficiency by one expert
Describe one data sample	Proportions	Mean, SD	Median
Compare one data sample to a hypothetical distribution	χ^2 or binomial test	One-sample t test	Sign test or Wilcoxon test
Compare two paired samples	Sign test	Paired t test	Sign test or Wilcoxon test
Compare two unpaired samples	χ^2 square Fisher's exact test	Unpaired t test	Mann-Whitney test
Compare three or more unmatched samples	χ^2 test	One-way ANOVA	Kruskal-Wallis test
Compare three or more matched samples	Cochrane Q test	Repeated-measures ANOVA	Friedman test
Quantify association between two paired samples	Contingency coefficients	Pearson correlation	Spearman correlation

Table 5. Which statistical inference test to use for which type of data. All statistical tests in this table are described in the text and often instantiated using a numerical example.

when processing time series data. The standard conservative approach developed by Bonferroni [3] consists of dividing the *p*-value threshold by the number of comparisons performed. For example, for 100 tests performed at p=0.05, the corrected p value is 0.05/100=0.0005. This is a conservative approach and a less stringent method has been developed by Holm [4]: first choose a significance level $p=\alpha$ (e.g., p=0.05). Then compute the exact *p*-value for each test (which is usually possible using modern computerized approaches). Rank the collection of *p*-values from smallest to largest. The smallest *p*-value is tested against α/N, where N is the number of tests. If the smallest *p*-value is not less than α/N, stop the procedure. However, if it is less than α/N, proceed to test the second smallest *p*-value against $\alpha/(N-1)$, etc... A variant of the Holm's procedure consist of testing the first *p*-value against α/N, the second one against $2\alpha/N$, the third one against $3\alpha/N$, etc. Technical details and theory about multiple comparisons may be found in [5].

Paired/unpaired samples: Table 5 distinguishes between paired versus unpaired data samples. For unpaired data samples, there is no direct correspondence between values. This may be the case when a specific measure (e.g., blood pressure) is taken from two distinct populations of patients (e.g., patients suffering from heart failure and control patients). The two data samples corresponding to the two groups of patients are said to be unpaired because there is no relationship between them. In contrast, for paired samples, each value in one sample corresponds to a value in the other sample. In the previous example, it could be the case if each patient tested had a twin volunteering to be a control patient. This would also be the case if two assessments were performed on the same patients (e.g., measure of blood pressure before and after taking a drug). Note that paired groups must necessarily be of the same size. Matched/unmatched data samples are an extension of paired/unpaired data samples when there are more than two samples.

Sampling with or without replacement: Sampling with replacement means that each item is put back in the data sample after being sampled (so it may be sampled more than once and appear twice or more in a data sample). Sampling with replacement satisfies the requirement that the trials are independent, but when the sample size is small relative to the size of the population, sampling with or without replacement makes little difference. In elementary statistics, a representative sample is synonymous with the concept of a "random" sample. When sampling from a population of finite size, a sample of n items is a random sample if it is chosen in such a way that any other sample of size n would be equally likely to be chosen. Sampled items can be chosen with or without replacement. Although impractical in many situations, sampling with replacement leads to easier mathematical analysis. When the population is large relative to the sample size, the analytical methods developed for sampling with replacement yield good approximations. A random sample can be chosen by assigning a number to each member of the population, and then choosing at random n numbers (with or without replacement). This can be done by the so-called Monte-Carlo method (consisting of random draws) that uses computer-generated (pseudo) random numbers.

Table 5 indicates which statistical test should be used depending on data type and question type. We have already described most types of questions when we defined hypotheses H_0 and H_1. We did not cover the last row of Table 5 which is concerned with the relationship between data samples (or more specifically the relationship between variables underlying two paired data samples). The corresponding question may be formulated as “is there any relationship between the two variables (e.g., two paired measurements)?” The H_0 hypothesis is that there is no relationship between the two variables.

Columns of Table 5 contain elementary tests for different types of variables. Elementary tests cover confidence interval estimation and parametric hypothesis testing for situations involving normally distributed samples, including two-sample situations where the purpose is to compare two populations with respect to their means or variances. Other types of elementary confidence intervals are for proportions and difference of proportions, usually based on the binomial distribution or based on the normal approximation to the binomial distribution. Confidence interval estimation for parameters of non-normal distributions are much more difficult and closed form formulas often do not exist. In these cases, experimenters must use non-parametric statistical tests that only take into account rank ordering of data samples. They may also use resampling statistical tests, which estimates confidence intervals using many computer-generated random resamplings. For practicality, in Table 5, we divided hypothesis testing into three main categories: hypothesis testing on discrete variables, parametric statistical testing on continuous variables, and non-parametric statistical testing on continuous variables. We will deal in a separate section with resampling methods since it may be applied to any type of data. The list of tests is not exhaustive but instead seeks to provide, within the limited scope of this short article, a range of methods to perform statistical inference on different types of data.

Which type of test to use is often one of the most delicate choices an experimenter has to make. For continuous data for instance, one could use at least three tests: a parametric, a non-parametric, or a resampling inference test. Different tests make different assumptions: parametric test such as the *t*-test make the hypothesis that the data is normally distributed. Non-parametric tests make fewer assumptions about the population distribution but require more data samples. Resampling tests make the assumption that the data samples are an accurate representation of the population. There is no ideal test (although some applied statisticians would argue that resampling methods are indeed superior to other methods), and the test to choose often depends on the type of data being processed or common usage in one specific field of research.

Testing hypothesis on discrete variables

For discrete variables, data is most often represented by proportions of different outcomes. As shown in the first column of Table 5, specific tests have been designed to deal and compare proportions between data samples. Some of these tests (as indicated below) can only deal with binomial data samples (“success” or “failure”).

Goodness of fit to distribution for one data sample

A goodness of fit test may be used to compare one data sample to a hypothetical value or distribution. In a goodness-of-fit test the hypotheses are concerned with the distribution itself. For example, a drug has been repeatedly tested on adults and has shown minor side effects in 2.5% of the cases in which it was administered. To validate this drug for treating children, it is given to a sample of 300 children. The goal of this study is to determine if children showed more or less side effects than adults. The hypothesis H_0 is that the distribution of sample data values for children is generally the same as the hypothetical distribution for adults. The hypothesis H_1 is that the distribution of sample data values for children generally differs from the hypothetical distribution for adults. Table 6 indicates that 13 out of 300 children showed an abnormal reaction to the drug. The second column in Table 6 indicates the expected values from the theoretical distribution (2.5% of cases for 300 subjects is 7.5 individuals; it is not so important that the expected value is not a whole number since this distribution is only theoretical).

	Children	Expected value
Side effect	13	7.5
No side effects	287	292.5
Total	300	300

Table 6. Measured and expected frequencies of side effect for 300 children treated with a test drug.

The χ^2 value is then simply calculated by comparing the expected frequencies e_1 (7.5 individuals showing side effects) and e_2 (292.5 individuals showing no side effects) to the observed frequencies O_1 (13) and O_2 (287) using the formula:

$$\chi^2 = \frac{(O_1 - e_1)^2}{e_1} + \frac{(O_2 - e_2)^2}{e_2}$$

or more generally

$$\chi^2 = \sum_i \frac{(O_i - e_i)^2}{e_i} \qquad (13)$$

where O_i is the frequency observation in row *i* and e_i is the corresponding expected frequency. The degrees of freedom is equal to *(n - 1)*, where *n* is the number of rows in the table. Once the χ^2 value and the degrees of freedom have been calculated, the critical value for χ^2_{crit} can be looked up in Table 7 for a given level of significance. If $\chi^2 > \chi^2_{\text{crit}}$, we reject

df	p=0.05	p=0.01	p=0.001	df	p=0.05	p=0.01	p=0.001
1	3.84	6.64	10.83	20	31.41	37.57	45.32
2	5.99	9.21	13.82	21	32.67	38.93	46.80
3	7.82	11.35	16.27	22	33.92	40.29	48.27
4	9.49	13.28	18.47	23	35.17	41.64	49.73
5	11.07	15.09	20.52	24	36.42	42.98	51.18
6	12.59	16.81	22.46	25	37.65	44.31	52.62
7	14.07	18.48	24.32	26	38.89	45.64	54.05
8	15.51	20.09	26.13	27	40.11	46.96	55.48
9	16.92	21.67	27.88	28	41.34	48.28	56.89
10	18.31	23.21	29.59	29	42.56	49.59	58.30
11	19.68	24.73	31.26	30	43.77	50.89	59.70
12	21.03	26.22	32.91	35	49.80	57.34	66.62
13	22.36	27.69	34.53	40	55.76	63.69	73.41
14	23.69	29.14	36.12	50	67.51	76.15	86.66
15	25.00	30.58	37.70	60	79.08	88.38	99.62
16	26.30	32.00	39.25	70	90.53	100.42	112.31
17	27.59	33.41	40.79	80	101.88	112.33	124.84
18	28.87	34.81	42.31	90	113.15	124.12	137.19
19	30.14	36.19	43.82	100	124.34	135.81	149.48

Table 7. χ^2 distribution of critical values. To use this table, choose a *p* value (column) and read the value for your calculated degrees of freedom (*df*). If your calculated χ^2 value is larger than the one you read in the table, the test you performed is significant (see text for details).

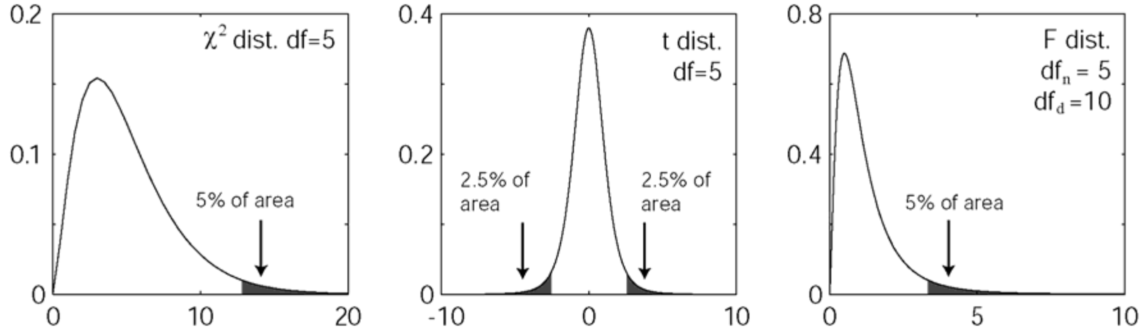

Fig. 4. Standard distributions (χ^2, t, and F). Tails of these distributions are used to determine significance thresholds (see text).

hypothesis H_0 in favor of hypothesis H_1 and conclude that the data support the hypothesis that there is a difference between the sample data and the theoretical distribution at the 5% level of significance.

In the example shown in Table 6,

$$\chi^2 = \frac{(13-7.5)^2}{7.5} + \frac{(287-292.5)^2}{292.5} = 4.13$$

with 1 degree of freedom (2 rows minus 1). For a test at the 5% level of significance (p=0.05) with 1 degree of freedom, χ^2_{crit} in the χ^2 table (Table 7) is equal to 3.84. Since 4.13 > 3.84, the hypothesis that the proportion of children having side effects in the same as that of adults is rejected. Comparing actual and expected frequencies in Table 6, we conclude that children have higher occurrences of side effects than adults for this drug.

Note that the construction of the χ^2 table is relatively simple. One can simply assume that a known population (whose expected distribution is known) is sampled several times and that the χ^2 value is computed for each of these samples. The histogram of these observed χ^2 values, when in fact no real effects are present, is an approximation to the χ^2 distribution for the null hypothesis (Fig. 4). The tails of this distribution may be used to set thresholds for significance testing (if an observed χ^2 value ends up in the tail of the distribution, then it is likely that it does not originate from the known population). For example, the χ^2 value for a data sample is significantly different from the χ^2 standard distribution at p=0.05 if it lies in the lower or upper tails each containing only 2.5% of the values of the standard χ^2 distribution.

Binomial test for binomial variables

For data samples that we assume are obeying the binomial distribution, it is possible to compute exact p values as explained in a previous section. For example, a coin is tossed 10 times to determine if it returns fair results or not. It returns 9 heads. The hypothesis H_0 is that the coin is fair and that the probability of obtaining a head is 0.5. The H_1 hypothesis is that the coin is biased towards head. Using the binomial distribution, we need to compute the probability of obtaining an equal or more extreme number of heads than the one we measured. The probability of obtaining 9 heads or more is

$$\begin{aligned} P(X \geq 9) &= P(X=9) + P(X=10) \\ &= \binom{10}{9}.5^9(1-.5)^{10-9} + \binom{10}{10}.5^{10} = 0.011 \end{aligned}$$

It appears that this result would appear by chance in only 1.1% of coin tossing trials if the coin is returning fair results. If we consider p<0.05 to be the standard threshold for significance, we can conclude that the coin does return more heads than a fair coin at the 5% level of significance. Note that this was a one-sided test, assuming that we have prior knowledge that the coin will be biased towards heads (based for instance on the aspect of the coin): for a two-sided test that would assess if the coin is "fair" in returning both faces and heads, we would need to add probabilities of obtaining both 9 to 10 heads and 9 to 10 faces.

Sign test to compare paired samples

This test is best illustrated by an example. To determine if drug A is more effective than a drug B for pain control, 10 patients are tested with these 2 drugs (with an interval of several days to prevent carry over effects) and asked if the drug was effective in controlling their pain. Hypothetical results are shown in Table 8, with "+" signs indicating a positive effect of the drug and "-" signs indicating no effect of the drug. The last row indicates the sign of the difference between the first two rows: a "+" sign indicates that drug A is performing better than drug B and a "-" sign indicates that drug B is performing better than drug A. When the outcome is the same, the cell is left empty. If the two drugs are equally effective, and if the sample is large enough, then there should be approximately equal numbers of "+" and "-" signs in the last row. We can test the expected number of "+" signs (6 out of 7 non-empty cells) using binomial probability (note that for a large number of values, the approximation of the binomial distribution by the normal distribution may be used). We need to compute the probability of obtaining an equally or more extreme number of "+" or "-" than the one we obtained, hence to compute P(0, 1, 2, 5, 6, 7):

$$P(\text{"+"}=0,1,2,5,6,7) = \binom{7}{0}.5^0(1-.5)^{7-0} + \binom{7}{1}.5^1(1-.5)^{7-1} + ... = 0.45$$

Patient	1	2	3	4	5	6	7	8	9	10
Drug A	+	+	+	-	+	+	+	+	+	+
Drug B	+	-	-	+	+	-	-	+	-	-
Sign		+	+	-		+	+		+	+

Table 8. Success (+) or failure (-) of drug A and B for reducing pain in 10 patients.

Although it seems that drug *A* is a better pain killer than drug *B*, the *p* value did not reach significance ($p>0.05$). In other words, H_0, the hypothesis that the two drugs are performing equally well cannot be rejected. This type of test applied to binomial variables is also sometimes called the Mc Nemar's test.

χ2 test to compare two or more unpaired samples

The χ^2 test allows the comparison of proportions observed in several groups under two or more conditions. Suppose that we wish to determine which of four prosthetic devices perform better for improving muscular response. Each of the four devices is implanted in 4 random samples of 100 patients each. For each patient, a clinician then estimates if there has been "no improvement" or "partial to full restoration". Data is cross-classified as shown in Table 9. The test described here is usually called the χ^2 test of independence because it aims at finding if results from different groups can or cannot originate from the same population.

Here, the objective is to determine whether improvement is independent of the type of device. If it is the case, then the proportion of responses with "no improvement" and "partial to full restoration" should be similar for all four types of devices. The χ^2 test allows the comparison of the actual proportion of responses to each type of device to the idealized proportions where all types of devices perform equally well. These proportions (also called expected frequencies) are calculated by pooling the responses for all types of devices. For instance, in Table 9, irrespective of the device type, there are 120 patients showing no restoration and 280 patients showing some degree of restoration, so the expected frequency for "no restoration" is 30% and the expected frequency for "partial to full restoration" is 70%.

As for the simpler example earlier in this section comparing a sample data distribution to a theoretical distribution, the χ^2 is simply calculated by comparing the expected frequencies, denoted by $e_{i,j}$ for device *i* (where *i* ranges from 1 to 4) and outcome *j* (where *j*=1 indicates "no restoration" and *j*=2 indicates "partial to full restoration"), to the observed frequencies $O_{i,j}$ using the formula:

$$\chi^2 = \sum_{i,j} (O_{ij} - e_{ij})^2 / e_{ij} \tag{14}$$

The degrees of freedom is equal to *(r - 1) (c - 1)*, where *r* and *c* are the number of rows and columns in the table. Once the χ^2 value and the degrees of freedom have been calculated, the critical value for χ^2_{crit} can be looked up in Table 7 for a given level of significance. If $\chi^2 > \chi^2_{\text{crit}}$, then there is a significant difference between the groups being compared.

	Type of device				
	1	2	3	4	Total
No improvement	35(30)	40(30)	35(30)	10(30)	120
Partial to full restoration	65(70)	60(70)	65(70)	90(70)	280
Total	100	100	100	100	400

Table 9. Results of improvement in muscular response following implantation of an electronic device available in four types. Numbers are observed frequencies and number in parentheses are expected frequencies.

For the example shown in Table 9,

$$\chi^2 = \frac{(35-30)^2}{30} + \frac{(65-70)^2}{70} + \frac{(40-30)^2}{30} + ... = 26.2$$

The degrees of freedom is (4-1)(2-1)=3. In this example, for a test at the 5% level of significance (p=0.05) and 3 degrees of freedom, Table 7 indicates that χ^2_{crit}= 7.82. Since 26.2 > 7.82, the hypothesis that all four devices are equally effective is rejected. It can be seen that device type 4 is most effective. In fact, further analysis supports the conclusion that differences between the other device types can be explained by sampling variation, and that there is a statistically significant difference between the first three device types taken together and the 4th device type. The additional analysis is sensible because the first three types are different "vintages" of essentially the same design, whereas the type 4 device is an experimental version of a fundamentally different design.

The χ^2 test may be used on a table of any size and not necessarily on binomial variables. For the example shown in Table 9, we could imagine three possible outcomes - "no improvement," "partial restoration," and "full restoration.". This would have added a row to Table 9 but the χ^2 formula (Equation (14)) would still apply.

Quantify relationship between variables

Classification in a table often reflects characteristics of individuals or objects, so they are often referred to as attributes. A measure of the degree of relationship, association, or dependence of two attributes (and the associated variables in the population) is called the coefficient of correlation. It is given by

$$r = \sqrt{\frac{\chi^2}{N(\min(\#rows, \#columns) - 1)}} \tag{15}$$

where χ^2 represents the value computed from the χ^2 table; *N* is the total number of observations, and min(*#rows,#columns*) represents the smaller number between the number of rows (*#row*) and the number of columns (*#columns*). *r* can only take values between 0 and 1. The closer *r* is to 1, the greater the association between the two (or more) columns of the table. To determine if a value of *r* is significant or not, χ^2 tests previously described in this section may be used.

Parametrical hypothesis testing on continuous variables

A parametric statistical hypothesis assumes that the data sample originates from a population that fits a specific model (most often the normal model). This is usually the case when recording a measure that fluctuates around a fixed mean because of environmental noise. Before running any statistical tests, one must verify that the data distribution is consistent with the normal distribution. First, plot the histogram to check that the distribution's overall shape is similar to that of the normal distribution. You may then perform a goodness of fit test with the normal distribution. In a goodness-of-fit test, the hypotheses are concerned not with the parameters but with the distribution itself. For example, H_0: *X* has a normal distribution; H_1: *X* does not have a normal distribution. This may be done using the χ^2 goodness of fit test (mentioned in the previous section) applied to the data histogram frequencies compared to expected values calculated from the normal distribution (by integrating Equation (9) using Equation (5)). Other goodness-of-fit tests are the Kolmogorov-Smirnov, Cramer-Von Mises, and Anderson-Darling. There are also tests when H_0 involves some specific distribution, for example, the Shapiro-Wilk test for normality. Most computer packages incorporate such tests.

One-sample t-test to compare one data sample to a hypothetical distribution

This test is used to determine if a data sample belongs to a population with mean μ and standard deviation σ (the hypothesis H_0 is that it does belong to this population). This test applies to continuous or non-continuous data that have a distribution that is not significantly different from normal. First, check that the standard deviation of the data sample is similar to the population's standard deviation σ (within a two-fold range). For a data sample containing *N* values that has a mean *M* and standard deviation *SD*, the variable *t* is defined as

$$t = \frac{M - \mu}{SD}\sqrt{N} \qquad (16)$$

The degrees of freedom associated with *t* is equal to *df=N-1*. After calculating *t* and *df*, set up a threshold for significance (i.e. $p<0.05$) and look up t_{crit} critical value in Table 10. In the *t*-test table, you may choose either one-tailed or two-tailed *t*-test critical values. One-tailed *t*-tests are used when there is some prior knowledge to predict the direction of the difference. Most commonly, two-tailed *t*-tests are used when there is no such knowledge. If the calculated *t* value is greater than t_{crit}, there is a statistically significant difference between the data sample and the hypothetical distribution (the null hypothesis H_0 is rejected).

As for the χ^2 table, building the *t*-test table is straightforward. One may assume that, for a given degree of freedom, a known population (with a normal distribution) is sampled several times and that the *t* value is computed for each of these samples (*M* should on average be the same as μ since it is the theoretical population which is being sampled). The histogram of these observed *t* values, obtained when in fact no real effects are present, is an approximation to the *t* distribution (Fig. 4, middle panel). The tails of this distribution allow for threshold-setting for significance (as for the χ^2, if an observed value ends up in the tail of the distribution, then it is likely that it does not belong to this distribution). Note that for an infinite number of degrees of freedom, the *t* distribution is equal to the normal distribution.

For example, in the past, a machine has been producing washers having a thickness of 0.06 inches on average. To test if the machine is still working properly, we produce 10 washers of size (0.065; 0.062; 0.060; 0.059; 0.061; 0.064; 0.067; 0.064; 0.061; 0.062). The sample mean is 0.0625 and the sample standard deviation is 0.0025. The *t* value is equal to

One-tailed	.1	.05	.025	.01	.005	.0001
Two-tailed	.2	.1	.05	.02	.01	.0002
df						
1	3.078	6.314	12.71	31.82	63.66	318.3
2	1.886	2.920	4.303	6.965	9.925	22.33
3	1.638	2.353	3.182	4.541	5.841	10.21
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.611
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.704	3.307
50	1.295	1.676	2.009	2.403	2.678	3.261
60	1.296	1.671	2.000	2.390	2.660	3.232
80	1.292	1.664	1.990	2.374	2.639	3.195
100	1.290	1.660	1.984	2.364	2.626	3.174
1000	1.282	1.646	1.962	2.330	2.581	3.098
inf.	1.282	1.64	1.960	2.326	2.576	3.091

Table 10. *t* distribution of critical values. To use this table, find your degrees of freedom in the *df* column (or a lower one if yours is not present in the table). Then, look up the probability in the top row (*p*=0.05 is a test of significance at 5%). If your calculated *t* value is larger than the one you read in the table, the test you performed is significant (see text for details).

$$t = \frac{0.0625 - 0.06}{0.0025}\sqrt{10} = 3.16$$

We use a two-tailed *t*-test since we have no *a priori* knowledge about the sampled distribution. At the 5% significance level, $t_{crit5\%}$ is equal to 1.83. Since t>1.83*,* we can conclude that there is a significant difference between the expected washer thickness and the observed one (we reject hypothesis H_0 which assumes that the sample distribution has a mean of 0.06 inches). However at .5% significance level, $t_{crit1\%}$ is equal to 3.25. Since t< 3.25, we cannot conclude that such difference exist at this level of significance (we cannot reject hypothesis H_0).

Paired t-test to compare paired data samples

This test applies to two paired samples of continuous or non-continuous data that have a distribution non-significantly different from normal and similar standard deviations (with less than 2-fold difference). First calculate the difference between each pair and average them (D_{av}) (note that differences in values also have to be normally distributed). Then calculate the value of *t* using

$$t = \frac{D_{av}}{SD}\sqrt{N} \tag{17}$$

where *SD* is the standard deviation of the difference between each pair. Since the accuracy of a statistic is influenced by the population size, we must then calculate the degrees of freedom (*df*) or the number of independent parameters used in the calculation of the test statistic. The degrees of freedom is equal to the degrees of freedom used in calculation the sample *SD*, that is, the number of pairs minus 1: *df=N-1.*

Finally, as for the one sample *t*-test, set up a threshold for significance, look up t_{crit} critical value in Table 10, and compare it to the calculated value. If the calculated *t* value is greater than t_{crit}, there is a statistically significant difference between the two groups (the null hypothesis H_0 is rejected).

For example, to test if a newly designed electronic blood pressure (BP) device returns similar (hypothesis H_0) or different (hypothesis H_1) readings compared to an old manual blood pressure device, readings on 10 patients are performed and presented in Table 11 (only systolic pressure in Hg.mm is reported in the table).

We must first ensure that the two standard deviations are similar (14.1 for the electronic BP device and 13.5 for the manual BP device). To calculate the *t* value, we compute the difference between each pair, check that their distribution is normal, and then average them. D_{av}= ((121-115)+(130-131)+…)/10=2.8. The standard deviation of the difference is *SD*=2.57*,* and the degrees of freedom is 9 (10 readings minus 1). Thus the *t* value is equal to

$$t = \frac{2.8}{2.57}\sqrt{10} = 3.44$$

At the 5% level of significance, for 9 degrees of freedom, $t_{crit5\%}$ is equal to 2.26. Since *t*> 2.26, we can conclude that the two devices return different averages (we can reject hypothesis H_0). The newly devised electronic BP device probably has to be recalibrated to better match the readings of the manual one.

Patient	1	2	3	4	5	6	7	8	9	10
Electronic BP device	121	130	129	113	145	132	110	116	125	155
Manual BP device	115	131	127	111	140	131	111	111	121	150

Table 11. Systolic blood pressure in Hg.mm measured in 10 patients using either a new electronic device or an old manual device.

HF patients	78	81	88	76	93	112	83	96		
Control patients	80	71	68	80	95	67	85	69	85	77

Table 12. Heart rate in beats/second of control and test patients suffering from heart failure.

Unpaired t-test to compare unpaired data samples

An unpaired *t*-test aims to compare two unpaired data samples and applies to continuous or non-continuous data that have a distribution not significantly different from normal. Sample sizes should be similar (with less than 2-fold difference) for the two groups and, if *n*<30, variances should also be similar (with less than 2-fold difference). If the *t*-test is used in other circumstances, the results will have no meaning.

The most common way of calculating the *t*-statistics for unpaired data samples is to use the pooled variance estimate (it is also possible to use unpooled variance estimates but this is less common and will not be presented here). First calculate the unbiased pooled variance estimate:

$$V = \frac{V_A(N_A - 1) + V_B(N_B - 1)}{N_A + N_B - 2} \tag{18}$$

Then estimate the standard error of the difference of the means:

$$SE = \sqrt{V(1/N_A + 1/N_B)} \tag{19}$$

Then the *t* statistics is the difference of the means divided by its estimated standard error:

$$t = \frac{M_A - M_B}{SE} \tag{20}$$

where M_A, and M_B are the means of groups *A* and *B* respectively and where V_A and V_B are the variances of groups *A* and *B* respectively. For this test, the number of degrees of freedom is equal to the total number of points minus 2, because two means are estimated.

$$df = (N_A + N_B) - 2$$

Finally, set up a threshold for significance (*p*<0.05 for example), and look up the critical value t_{crit} in Table 10 (see the section above on one sample *t*-test for the difference between one-tailed and two-tailed *t*-tests). If the calculated *t* value is greater than t_{crit}, there is a statistically significant difference between the two groups (the null hypothesis H_0 is rejected).

For example, to test if patients diagnosed with heart failure have similar (hypothesis H_0) or higher (hypothesis H_1) heart rates than control patients, 15 readings are performed at rest for these two groups of patients *A* and *B* of matched age, sex, and ethnicity. Heart rate is reported in beating per minutes in Table 12.

After testing for normality (see for how to test for normality at the beginning of this section), we ensure that standard deviations for the two data samples are similar (SD_A=11.5 and SD_B=9.1). To calculate the *t* value, we then need to compute the mean heart rate for each group. For patients suffering from heart failure, M_A=88.4, and for control patients, M_B=77.7 (variances are V_A=140.3 and V_B=82.9). Thus

df2\df1	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	100	∞
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.62	8.59	8.57	8.55	8.54
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.50	4.46	4.43	4.41	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.81	3.77	3.74	3.71	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.86	2.83	2.79	2.76	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.70	2.66	2.62	2.59	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.57	2.53	2.49	2.46	2.41
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.47	2.43	2.38	2.35	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.38	2.34	2.30	2.26	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.31	2.27	2.22	2.19	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.25	2.20	2.16	2.12	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.19	2.15	2.11	2.07	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.15	2.10	2.06	2.02	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.11	2.06	2.02	1.98	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.07	2.03	1.98	1.94	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.04	1.99	1.95	1.91	1.84
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	1.98	1.94	1.89	1.85	1.78
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.94	1.89	1.84	1.80	1.73
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.90	1.85	1.80	1.76	1.69
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.87	1.82	1.77	1.73	1.66
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.84	1.79	1.74	1.70	1.62
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.04	1.96	1.88	1.79	1.74	1.68	1.63	1.56
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.74	1.69	1.64	1.59	1.51
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05	1.97	1.89	1.81	1.71	1.66	1.60	1.55	1.47
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.95	1.87	1.78	1.69	1.63	1.58	1.52	1.44
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.65	1.59	1.53	1.48	1.39
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.89	1.81	1.72	1.62	1.57	1.50	1.45	1.35
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.88	1.79	1.70	1.60	1.54	1.48	1.43	1.33
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.85	1.77	1.68	1.57	1.52	1.45	1.39	1.28
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.80	1.72	1.62	1.52	1.46	1.39	1.32	1.19
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.77	1.69	1.59	1.48	1.42	1.35	1.28	1.12
1000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.76	1.68	1.58	1.47	1.41	1.33	1.26	1.08
∞	3.84	3.00	2.61	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.46	1.40	1.32	1.25	1.03

Table 13. *F* distribution of critical values at *p*=0.05 for ANOVA tests. To use this table, read the value at the intersection of the numerator's degrees of freedom (*df1*) and the denominator's degrees of freedom (*df2*). If your calculated *F* value is larger than the one you read in the table, the test you performed is significant (see text for details).

the pooled variance estimate is *V*=108, the standard error of the mean is 4.93 and the *t* value is equal to

$$t = \frac{88.4 - 77.7}{4.93} = 2.17$$

At 5% significance level for 16 degrees of freedom (10 heart failure patients plus 8 control patients minus 2), $t_{crit1\%}$ is equal to 2.12. Since *t*>2.12, the data support the fact that patients with heart failure have higher heart rate than controls (we can reject hypothesis H_0).

One-way ANOVA for unmatched samples

One-way ANOVA (Analysis of Variance) is used to test the hypothesis that two or more samples are drawn from the same distribution of values and have the same mean and variance. Unpaired student *t*-test is a particular case of one-way ANOVA applied to two data samples. As for *t*-test, ANOVA test applies to continuous or non-continuous data that have a distribution that is not significantly different from normal. Sample sizes should be similar (with less than 2-fold difference) for all sample groups and, if n<30, variances should also be similar (less than 2-fold difference). If the test is used in other circumstances, the test outcome will lead to erroneous conclusions. The basis of ANOVA is the *F* (Fisher) variable, which combines the unbiased variance between sample groups ($V_{interGroup}$) and the variance within sample groups ($V_{withinGroup}$).

$$F = \frac{V_{interGroup}}{V_{WithinGroup}} \qquad (21)$$

For several data samples *A, B, C, ...* of the same size, inter-group variance is defined as

$$V_{interGroup} = \frac{N_A(M_A)^2 + N_B(M_B)^2 + N_C(M_C)^2 + ... - N_T(M_G)^2}{N_G - 1} \qquad (22)$$

where M_A, M_B , and M_C are the means of sample *A, B, C, …* and N_A, N_B , N_C, … are the number of values in samples *A, B, C, …* M_G is the average of all values from all sample groups and N_G is the number of samples. The within sample group variance is defined as

$$V_{withinGroup} = \frac{(N_A - 1)(SD_A)^2 + (N_B - 1)(SD_B)^2 + (N_C - 1)(SD_C)^2 + ...}{N_T - N_G} \qquad (23)$$

where SD_A, SD_B, SD_C, … are the standard deviations of group *A, B, C, ...* and N_T represents the total number of observations (for all data sample pooled together). Degrees of freedom for the numerator of *F* and the denominator of *F* are defined as:

$$df_{numerator} = N_G - 1$$

$$df_{denominator} = N_T - N_G$$

Note that each variance in Equation (22) and (23) is divided by the appropriate degrees of freedom to give unbiased estimate of population variance (assuming the null hypothesis H_0 is true). As for other inference tests, the computed F value is tested against critical F values (Table 13) obtained from the tail of null-hypothesis F distribution (Fig. 4, right panel).

For example, a clinician planning to purchase equipment for electro-encephalography compares the signal to noise ratio for three sets of electro-encephalographic equipment. For each system that has been made available to him, he records 10 new patients performing standard psychophysical tasks and measures the signal to background noise ratio of the encephalographic equipment (Table 14).

After testing for normality, we must ensure that standard deviations are similar (i.e., no 2-fold differences). Standard deviation for Brand A is equal to SD_A=1.11; Brand B: SD_B=0.75; Brand C: SD_C=0.94. After calculating $V_{intergroup}$=0.44 and $V_{winthinGroup}$=0.89, we may calculate F using Equation (21)

$$F = \frac{0.44}{0.89} = 0.49$$

The degrees of freedom for the numerator is $df_{numerator}$=N_G-1=2. The degrees of freedom for the denominator is $df_{denominator}$=30-3=27. Reading F_{crit}=2.95 in Table 13, we may conclude that there is no significant difference (since F<2.95) in terms of signal to noise ratio between the three sets of EEG equipments (we accept hypothesis H_0).

One-way ANOVA for matched samples

One-way ANOVA may also be used to compare paired sample groups. In fact, since for matched samples, one may analyses either the rows or the columns of a table, the formula given here may be used both for rows or columns, and are usually associated with two-way ANOVA. The formula for the F (Fisher) variable is now equal to

$$F = \frac{V_{interGroup}}{V_{error}} \qquad (24)$$

The variance due to error or chance is defined as

$$V_{error} = \frac{\sum_{j,k}\left(x_{jk} - M_{j.} - M_{.k} - M\right)^2}{(N_R - 1)(N_C - 1)} \qquad (25)$$

where x_{jk} are all the elements in the array, $M_{j.}$ are the row means, $M_{.k}$ are the column means, M is the global array mean, N_C is the number of columns, and N_R the number of rows. The degrees of freedom for the numerator and denominator are now defined as

$$df_{numerator} = N_R - 1 = N_G - 1$$

Brand A	1.87	3.88	2.68	1.19	0.93	0.38	2.69	1.8	0.39	1.62
Brand B	2.48	1.71	3.05	1.58	1.7	3	0.47	2.11	2.18	2.22
Brand C	2.29	1.49	2.52	1.26	3.71	2.14	2.33	2.79	2.61	0.29

Table 14. Signal to noise ratio for 10 patients and for three brand of EEG system.

$$df_{denominator} = (N_R - 1)(N_C - 1)$$

Using the same example as shown in Table 14, and now assuming that the data samples are paired (EEG systems were tested with the same patients), we can compute the inter-subject variance V_{errort}=1.04, and

$$F = \frac{0.44}{1.04} = 0.42$$

The degrees of freedom for the numerator is $df_{numerator}$=N_G-*1=2*. The degrees of freedom for the denominator is $df_{denominator}$=(N_R-1)(N_C-1)=(3-1)(10-1)=18. For a test at 5% significance, reading F_{crit}=3.55 in Table 13, we may conclude that there is no significant difference (since F<3.55) in terms of signal to noise ratio between the three sets of EEG equipments.

Note that one could argue that instead of using ANOVA analysis, we could perform t-tests between each pair of samples. Although this is possible, the ANOVA test is more sensitive than a series of paired t-tests because it processes all data samples simultaneously.

Two-way ANOVA for two-factor experiments

This type of test is being used for experiments with two factors or two attributes. In the example above, to test the reliability of the EEG equipment, the clinician might want to perform three experimental protocols and measure the signal to noise ratio in each of these protocols. The two factors are now the three sets of EEG equipment and the three protocols as shown in Table 15.

In each of the cells of Table 15, the clinician recorded nine values. In the case of only one value per cell, the analysis would be similar to the one-way ANOVA (row and column data may be analyzed separately using one-way ANOVAs for matched samples). However, if several values are recorded for each cell (several subjects for instance), one must use the repeated measures two-way ANOVA test. This test is especially interesting because it is possible to test for interaction between variables. Hypothesis H_0 would be that there is no significant relationship between brands and type of protocol and Hypothesis H_1 would be that there is indeed such a relationship. Running a repeated measures two-way ANOVA test under any software will return 3 p-values: the first value is

	Protocol 1	Protocol 2	Protocol 3
Brand A	6 8 8 10 8 2 10 6 2	1 0 2 2 0 1 1 3 3	1 4 4 4 2 2 3 5 0
Brand B	2 2 10 6 10 10 6 2 6	4 9 8 5 5 9 3 7 7	3 3 6 5 4 2 3 5 6
Brand C	6 0 4 6 4 4 4 8 8	5 3 2 1 1 1 3 3 2	6 6 8 6 0 10 . 4 8 6

Table 15. Example of table for a 2-factor experiment.

for significant differences between rows; the second value is for significant differences between columns; the last *p*-value is for the interaction between columns and rows. In the case of Table 15, the *p*-value for the columns (protocol) is 0.0004 indicating a significant difference between protocols. As observed in Table 15, the values for the first protocol are indeed higher than the values for other protocols. The *p*-value for the different rows (device brand) is not significant (p=0.22). The *p*-value for the interaction between brand and protocol is 0.0006. In fact, it appears that the device of brand *B* returns higher values for protocol 2 than other brands, and that the device of brand *C* returns higher values for protocol 3 than other brands.

Experimental design and ANOVA in its many variations is perhaps the most important statistical methodology for experimenters, and the literature is immense. Extreme care should be taken when choosing an ANOVA test. For instance, there are different ways to treat multifactor ANOVAs analytically when the number of observations is unequal among the treatment combinations (called unbalanced designs). A non-technical discussion is the classic Planning of Experiments by D. R. Cox [6]. Other general introductions are [1, 7-12].

Regression and Correlation

Regressions and correlations aim at determining relationship between variables. We may wish to determine if there is a significant correlation between independent and dependent variables, the independent variable being set by the experimenter, and the dependent variable being measured. For example, to test the reliability of a device, an experimenter may change the temperature of the room where the device is being tested (independent variable), and see if this change affects measures returned by the tested device (dependant variable). Regression and correlation can also be used to estimate the relationship between two (or more) dependent variables.

The first step in determining the relation between two variables is to plot values of one variable versus values of the other variable. This is usually called a scatterplot (Fig. 5). From the scatterplot it is often possible to visualize a smooth curve that approximates the data. If it is a straight line, then the least squares regression method may be used. Otherwise, other curve fitting procedures may be used. It is sometimes useful to plot scatterplots of transformed variables (for instance, log transformation of values the in first variable versus values of the second variable).

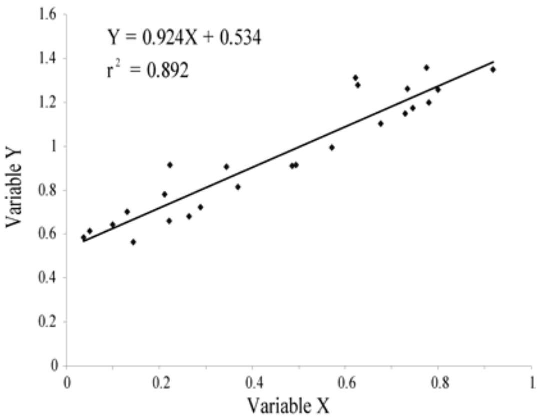

Fig. 5. A typical scatterplot with the least-square line drawn through the data points. The r^2 value as well as the best fit equation is indicated on the diagram. The *t*-value is equal to 9.24 and indicate a significant relationship between *X* and *Y* (at p=0.05, for 22 degrees of freedom, t_{crit}=2.07).

The method of least squares computes the best linear regression between two variables. Specifically, for two variables *X* and *Y*, the data consist of *n* pairs $(x_1, y_1),\ldots, (x_n, y_n)$. For all values of *X* and *Y*, we wish to find the parameter *a* and *b* such that

$$Y = aX + b \tag{26}$$

Assuming the jittering of points along the straight line is normally distributed, parameters *a* and *b* may be obtained using the formula

$$a = \frac{N\sum x_i y_i - (\sum x_i)(\sum y_i)}{N\sum x_i^{\,2} - (\sum x_i)^2} \tag{27}$$

$$b = \frac{1}{N}\left(\sum y_i - a\sum x_i\right) \tag{28}$$

To draw the linear regression line, y_i^{est} values may be calculated using Equation (26) for all values of *X*. A sample-based measure of the strength of the linear association between the *X* and *Y* variables is the sample correlation coefficient (also known as the Pearson correlation coefficient) defined by

$$r = \pm\sqrt{\frac{explained\_variation}{total\_variation}} = \pm\sqrt{\frac{\Sigma(y_i^{est} - M_Y)^2}{\Sigma(y_i - M_Y)^2}} \tag{29}$$

r may also be expressed using the original variables *X* and *Y*.

$$r = \frac{\mathrm{cov}(X,Y)}{SD_X.SD_Y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - M_X)(y_i - M_Y)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - M_X)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - M_Y)^2}} \tag{30}$$

where M_X and M_Y (SD_X and SD_Y) are the mean (the standard deviation) for *X* and *Y* respectively and *cov(X,Y)* is the covariance between *X* and *Y* (the numerator on the right of Equation (30) is equal to *cov(X,Y)* and the denominator is equal to SD_X*SD_Y). Necessarily $-1 \le r \le 1$. Positive (respectively negative) values of *r* indicate that large values (respectively small values) of *X* are associated with large values (respectively small values) of *Y*. Values of *r* near 0 indicate little or no linear association. Interpretation must be done with care because there are many reasons for the presence or absence of a correlation. Also, comparing *r* values may be misleading as a value of r = 0.6 does not mean that the linear relationship is twice as strong as r = 0.3. On the other hand, r^2, called the sample coefficient of determination, represents the proportion of the total variation in the sample values of *Y* that can be “explained” by a linear relationship as in Equation (26). Thus $r^2 = (0.3)^2 = 0.09$ versus $r^2 = (0.6)^2 = 0.36$ indicates a 9% versus 36% accountability for total variability by the proposed linear relationship.

To test if the linear correlation between the two variables is significant, different tests may be used. The null hypothesis H_0 states that there is no relationship between the two variables. A *t*-test (with degrees of freedom equal to *N-2*) may be used if the expected population correlation coefficient between variable X and Y is 0 and if we expect the correlation coefficient to be normally distributed when random samples of X and Y are drawn. The variable t is defined as

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{31}$$

More details for determining if correlation coefficients are significant or to compare between correlation coefficients may be found in Spiegel and Stepens [2].

As shown in Fig. 5, most regression computer packages will output scatterplots, and correlation coefficients. Residuals plots (not shown here) indicate if the distribution of distance between estimated and actual values of Y. A histogram of these residuals should be normally distributed (computing the parameter a, b, and the coefficient of correlation r requires that these residuals are normally distributed with mean 0 and a constant standard deviation irrespective of the X values).

A comprehensive presentation of regression methods for linear and non-linear regression is given in [2, 7, 13].

Non-parametric testing

Elementary tests mentioned in the previous section require that the distribution of values in the population be normally distributed. In practice, this assumption may not hold so statisticians have devised tests that are less dependent of population distribution. Non-parametric or distribution-free statistical methods generally are not concerned with inferences about parameters of distributions and assume little or no knowledge about the distributions of the underlying populations. Their primary advantage is that they are subjected to less restrictive assumptions than their parametric counterparts. Moreover, the data need not be quantitative (data values may indicate ranks on an ordinal scale). However, a disadvantage of nonparametric methods is that that they may not utilize all the information in a sample, consequently requiring a larger sample than the parametric version to attain the same Type II error (see error types).

The χ^2 goodness-of-fit tests previously mentioned is an example of a non-parametric test. Other non-parametric tests make various hypotheses for medians (or means of a symmetric distribution) and differences in location and/or variability of two populations. There are also tests for randomness, independence, and association among random variables. Relatively elementary texts that give a fairly broad and complete coverage of non-parametric methods are [14, 15].

Patient	1	2	3	4	5	6	7	8	9	10
Pacemaker off	0.15	0.32	0.25	1.1	0.82	0.83	0.94	0.42	0.48	0.21
Pacemaker on	0.12	0.19	0.28	0.56	0.37	0.52	0.24	0.73	0.81	0.13
difference	0.03	0.13	-0.03	0.54	0.45	0.31	0.70	-0.51	-0.43	0.08
abs difference	0.03	0.13	0.03	0.54	0.45	0.31	0.70	0.51	0.43	0.08
Rank of abs difference	1.5	4	1.5	9	7	5	10	8	6	3
Signed rank	1.5	4	-1.5	9	7	5	10	-9	-6	3

Table 16. Heart rate variability for 10 patients while their pacemaker is switched on or off, and calculation of signed rank for Wilcoxon test. The sum of the values in last row is 23.

Compare sample distribution to a hypothetical distribution

As for binomial and discrete data, a χ^2 goodness-of-fit test may be performed. For continuous data, a χ^2 goodness-of-fit test may be used on the frequency distribution (histogram) of the data compared to a hypothetical distribution.

Sign test and Wilcoxon test for paired samples

As for binomial and discrete data, a sign test allows the comparison of paired samples (see the beginning of the section for a definition of paired and unpaired samples). A sign test simply involves pair-wise comparisons of measures between the two sample data sets (see sign test for binomial and discrete data). A variation of this test is called the Wilcoxon test which takes into account the signed rank of the difference between each pair (instead of using all the signs). This is best illustrated using an example. To test if a pacemaker device has any effect on heart rate variability (defined as the standard deviation of heart beat intervals in seconds), 10 patients' heart rate variability are measured while the pacemaker was either switched on or off (Table 16).

The Wilcoxon test begins by taking the difference in heart rate variability between the two conditions for each patient (4th row of Table 16). If a difference is equal to 0 it is eliminated from further consideration, since it provides no useful information. The second step consists of taking the absolutes of the differences which is accomplished simply by removing all the positive and negative signs (5th row of Table 16), then ranking these absolute differences from lowest to highest, with tied ranks included where appropriate. Tied rank means that if two values are equal they are first ordered randomly and then assigned their average rank (see the 1st and 3rd columns of the 6th row in Table 16). Finally re-attach to each rank the positive or negative sign that was removed from the difference in the transition from row 4 to row 5, and sum up these values. In our case W=23 and the number of values used in this sum is 10 (degrees of freedom).

If we were to draw repeatedly two sets of sample values from the same distribution (which verify hypothesis H_0 that the two samples belong to the same distribution) and calculate W values, we would realize that the distribution (histogram) of W values is close to normal. In fact, we may define

$$z = \frac{W}{SD_W} \tag{32}$$

where z is normally distributed with mean 0 and variance 1, and SD_W is the standard deviation of W, which can be shown to be equal to

$$SD_W = \sqrt{\frac{N(N+1)(2N+1)}{6}} \tag{33}$$

For N=10 values, SD_W=19.6, so z=23/19.6=1.17. As mentioned earlier, the t-distribution is equal to the normal distribution for infinite degrees of freedom. Looking in the last row of the t-table (Table 10), for a significant threshold at

p=0.05 (two-tailed), we obtain z_{crit}=1.64. Since z<1.64, we cannot reject hypothesis H_0. Although it seems that heart rate variability is higher when the pacemaker is switched on, the difference did not reach significance.

Mann-Whiney U test for unpaired samples

The Mann Whitney U test is similar to the Wilcoxon test. Once more we will illustrate this test using an example. To compare sensitivity of two hearing aids, the minimum sound a patient can hear using each brand is measured (in dB) and reported in Table 17. 10 different patients tested each prosthetic device (unpaired samples).

Brand A	0.1	-1	4.1	2.4	-2.3	3.8	0.9	1.4	0.4	1.2
Brand B	2.7	3.1	5.2	2.1	4.7	1.5	-1.2	3.7	2.8	3.1

Table 17. Patient maximal sensitivity (in dB) for two brands of hearing aids.

Brand A	4	3	18	11	1	17	6	8	5	7
Brand B	12	14.5	20	10	19	9	2	16	13	14.5

Table 18. Rank of measures for table 17.

To perform a Mann Whitney test, first combine all values in an array and assign a rank from 1 to 20 to all these values, assigning tied ranks where appropriate (see Wilcoxon test). The rank for each value is indicated in Table 18.

Then sum up the ranks for each brand. R_A=80 is the sum for brand A and R_B=130 is the sum for brand B. A significant difference between the two rank sums implies a significant difference between the two samples. Calculate the U statistic to test the difference between the ranks:

$$U = N_A N_B + \frac{N_A(N_A+1)}{2} - R_A \tag{34}$$

Note that the formula above is symmetrical with respect to A and B. In the hearing aid example, N_A=10 and N_B=10, so

$$U = 10*10 + \frac{10(10+1)}{2} - 80 = 75$$

There is no Table for U values. Instead, as for the Wilcoxon test, the Table for z values is used because of a property of the U distribution. When calculating the U value repeatedly on samples known not to be statistically different (for instance two data samples drawn from the responses of the same device), then it can be shown that the repeated U values (*U1, U2, U3, …*) have a Gaussian distribution with mean M_U and standard deviation SD_U defined as:

$$M_U = \frac{N_A N_B}{2} \tag{35}$$

$$SD_U = \sqrt{\frac{N_A N_B (N_A + N_B + 1)}{12}} \tag{36}$$

This means that the U distribution can be normalized and that

$$z = \frac{U - M_U}{SD_U} \tag{37}$$

is normally distributed with mean 0 and variance 1.

In the example above M_U=10*10/2=50 and SD_U=13.2, so z=3.78. Looking up the last row of the t-table (Table 10) for a significance level of 5%, we read z_{crit}=1.64. Since z>1.64, we can reject hypothesis H_0 and conclude that one hearing aid performs better than the other one. Looking at the mean or median for each brand, or for this simple example simply at Table 17, brand A clearly allows patients to hear sounds of smaller amplitudes than brand B. Note that the calculations above are usually not necessary since most statistical software will return the value of U along with its significance level.

Kruskal-Wallis test for unmatched samples

The Kruskal-Wallis H test is a generalization of the Mann Whitney U test to more than two samples (for instance three brands *A, B,* and *C* of sample sizes N_A, N_B, N_C,… with the total number of samples equal to N). As for the Mann Whitney test, values from all distributions are sorted and once the sum of the rank for each sample is calculated R_A, R_B, R_C, … the value of H is given by

$$H = \frac{12}{N(N+1)}\left(\frac{R_A}{N_A} + \frac{R_B}{N_B} + \frac{R_C}{N_B} + ...\right) - 3(N+1) \tag{38}$$

It can be shown that, after collecting repeated measures of H from several samples from the same population (verifying the hypothesis H_0 that they originate from the same population), the histogram of H values is very close to a χ^2 distribution with degrees of freedom equal to the number of groups minus one (so the χ^2 table may be used for H). Thus, to use the Kruskal-Wallis test, first calculate H, then compute the degrees of freedom (number of groups minus one), and look up the χ^2 critical value in Table 7. If the calculated H value is larger than the critical value, reject hypothesis H_0.

Friedman test for matched samples

Suppose we wish to determine if three spectroscopy machines *A, B,* and *C* returns the same hematocrit density (density of blood cells in a blood sample). We test the three machines using 20 blood samples (the same blood sample is used for all machines). Since preliminary analysis shows that the readings are not normally distributed we have to use a non-parametric test. To do so, for each blood sample, we rank the machines (from 1 to 3) and compute the total rank for each machine T_A, T_B, and T_C. T_{all} being the sum of all the ranks, the squares deviate SS is equal to

$$SS = \frac{(T_A)^2 + (T_B)^2 + (T_C)^2}{N_G} - \frac{(T_{all})^2}{N_G N} \tag{39}$$

where N_G is the number of groups and N is the number of samples in each group. As for the Kruskal-Wallis test, we may use the χ^2 distribution with degrees of freedom equal to df=N_G—1. In the Friedman test, we simply refer to this value as χ^2

$$\chi^2 = \frac{SS}{N_G(N_G+1)/12} \tag{40}$$

If the calculated χ^2 value is larger than the critical value for the specified degrees of freedom, we reject hypothesis H_0.

The spearman's rank correlation test

Rank methods may also be used to determine the correlation between two variables. Instead of using exact variable values, their ranks may be used. For two sample *A* and *B* of the same size, corresponding to two variables *X* and *Y* (for instance lifespans and prices of a family of devices), rank each sample value from 1 to *N* separately for *A* and *B*. Then calculate the difference D_1, D_2, D_3, … between the sorted rank for *A* and *B* and compute

$$r_S = 1 - \frac{6\left((D_1)^2 + (D_2)^2 + (D_3)^2 + ...\right)}{N(N^2-1)} \tag{41}$$

If r_S is close to 0, there is no correlation between the two variables whereas if it is close to 1 or -1, there is a strong correlation between the two variables. To test if r_S is significantly different from 0, the same *t*-test as for the Pearson correlation coefficient may be used (replacing *r* by r_s and using the same degrees of freedom *df*=*N*-2).

Resampling methods

Resampling methods help provide confidence intervals for parameters in situations where these are difficult or impossible to derive analytically. Resampling methods also help perform statistical inference without assuming a known probability distribution for the data. The bootstrap method consists of drawing random sub-samples and the randomization method consists of shuffling data samples.

Bootstrap method

The bootstrap method is the most recently developed method to estimate errors and other statistics. It is not primarily aimed at performing inference although it may be used to do so, since it provides confidence intervals for the measure of interest. The term "bootstrap" derives from the phrase "to pull oneself up by one's bootstrap" (Adventures of Baron Munchausen, by Rudolph Erich Raspe). Suppose we have a data sample and an estimator (e.g., mean). The basic idea involves sampling with replacement to produce random samples of size *N* from the original data sample (of size larger than *N*). Each of these samples is known as a bootstrap sample and provides an estimate of the parameter of interest. Repeating the sampling a large number of times provides information on the variability of the estimator and help define confidence limits. There are *N* to the power of *N*, N^N, possible samples, called the ideal bootstrap samples. It is important to emphasis that sub-samples are drawn with replacement: for instance, for an empirical distribution composed of 2 values (5 and 8), the bootstrap samples are (5,8), (5,5), (8,8), and (8,5) (note that there are 2^2=4 of them). Getting all ideal bootstrap samples becomes unrealistic as *N* becomes larger, so the Monte-Carlo approach (which consists of random draws) is used. The sampling is said to be balanced if each sample value is drawn the same number of times. For each bootstrap sample, let's suppose that the mean is calculated. The standard deviation of the bootstrap distribution for the mean correspond to the standard error (Equation (19)) and may be used in parametrical *t*-test to compute the *t* value (Equation 20) and perform inference testing (assuming normality of the distribution of course). However this mixture of bootstrap and parametric *t*-test is relatively unconventional and it is better to estimate the bootstrap distribution of *t* value as explained below.

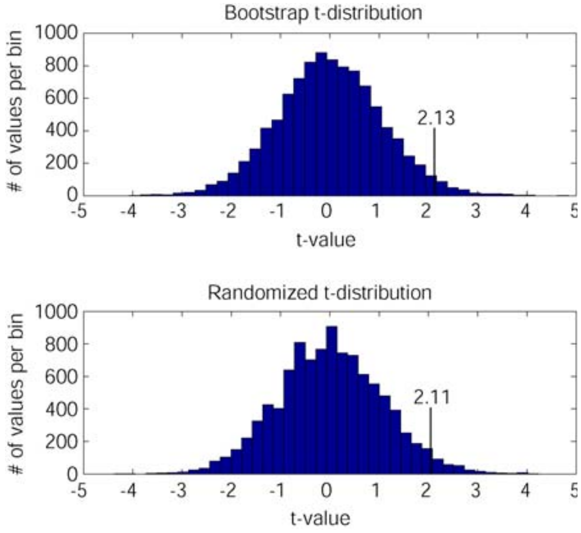

Fig. 6. Bootstrap t-distribution for Table 12 (top) and randomized t-distribution for Table 12 (bottom). Since the actual t value obtained from the original samples in Table 12 (*t*=2.17) belong to the rightmost 2.5% value in both the bootstrap and the randomized distribution (the 2.5% limit being indicated by a vertical line), it may be considered significant at 5%.

To perform a statistical inference test using bootstrap, we first have to state a null hypothesis H_0. Null hypotheses for resampling tests are usually vague because there may be many reasons (based on the shape of the distribution) why two samples may differ (whereas when performing a parametric *t*-test, the non-null hypothesis states clearly that the means are non-equal). Moreover, bootstrap statistics use the implicit assumption that data samples are representative of the underlying population and in fact do as if the data samples were the population itself. Therefore it is not possible to draw direct conclusions about the underlying population either.

In the case of the heart rate study of Table 12 for instance, where comparing a measure (i.e. heart rate) for patients suffering from heart failure (sample *A*) and control subjects (sample *B*), the null hypothesis would be "patient suffering from heart failure have abnormal heart rate". One way to test this hypothesis is to perform a bootstrap *t*-test. Two bootstrap samples are first drawn from the pooled distribution of *A* and *B*: sample *A'* and *B'* of the same size as *A* and *B* respectively. The *t*-value is then computed using the two bootstrap samples as in Equation (16). The operation is repeated *m* times to obtain the distribution of *t*-values for the null hypothesis. Note that, even if we are computing a *t*-value, we do not assume normality for the data samples since the distribution of *t* values for the null hypothesis is estimated using bootstrap samples. The actual *t*-value is calculated for the original data samples *A* and *B* and tested against the bootstrap *t*-distribution. If it lies in the lower 2.5% or upper 2.5% tails, then the bootstrap test may be considered to be significant at the 5% level of significance. In Fig. 6 (top), 10,000 bootstrap *t* values have been accumulated for the two samples in Table 12. Since the original *t* value for Table 12 is equal to 2.17 (see the *t*-test section) and since it lies in the upper 2.5% of the bootstrap *t*-value distribution, we may conclude that the data support the hypothesis that heart rate is affected in patients suffering from heart failure at the 5% significance level.

There are other ways to test for significance using bootstrap, such as the bootstrap-percentile method, or the bootstrap-bca method (see [16] for a comprehensive reference). In general, it should be remembered that bootstrap methods are designed primarily for estimating characteristics of data samples, not for performing inference tests. Resampling methods specializing in statistical inference are called randomization methods and are describe below.

Randomization methods

For the purpose of performing paired or unpaired comparisons, randomization methods consist of random permutations of data. Randomization methods are also often called permutation methods or surrogate methods. Specifying the null H_0 hypothesis is the same as for the bootstrap and involves a vague formulation about the result of the experiment, such as “patient suffering from heart failure have abnormal heart rate” or “the drug treatment does not have an effect on blood pressure”.

Randomizing the data is straightforward. Using the same example as for the bootstrap distribution with two unpaired samples *A* and *B* of sizes N_A and N_B, a randomization method consists of pooling the data of *A* and *B* together (into *C*), then randomly drawing from *C* (without replacement) two groups *A'* and *B'* which have the same size as *A* and *B* respectively [17]. Then we compute the estimator (e.g., *t* value) for each randomized pair of samples. We repeat this procedure a large number of times to obtain the distribution of the estimator (e.g., *t* value) for the null hypothesis. Significance is assessed as for the bootstrap *t*-test. For example, in Fig. 6 (bottom), 10,000 randomized *t* values have been accumulated for the two samples in Table 12. Note that irregularities in the distribution are due to the fact that we are randomizing a relatively small number of values. As for the bootstrap, since the original *t* value (t=2.17) lies in the upper 2.5% of the randomized *t* values, we may conclude that the data support the hypothesis that heart rate is affected in patients suffering from heart failure at the 5% significance level. It is reassuring to notice that the upper 5% significance threshold *t* value for the bootstrap (t_{crit}=2.13), the randomized (t_{crit}=2.11), and the normal distribution (t_{crit}=2.12) are all similar.

For paired comparisons, the principle is slightly different since we are now randomizing not the sample values but the pairs. For instance, for the data of Table 11, half of the pairs are selected randomly then shuffled (the value for the first device is now attributed to the second device and vice-versa) and the paired *t* test value is recalculated (Equation (17)). This procedure is repeated a large number of times. To assess significance, as in the previous paragraph, the original *t* value computed using the non-randomized samples is compared against the distribution of randomized *t* value.

This procedure may be generalized to compare an arbitrary number of samples. For instance, to compare several unpaired sample, data sample values may be randomized among groups and one-way ANOVA values may be calculated repeatedly. The ANOVA value for the non-randomized groups is then compared against this ANOVA randomized distribution. Web reference [18] provides a clear introduction to resampling methods.

MULTIVARIATE METHODS

We previously discussed probability distributions involving one variable, but in many situations there are two or perhaps many interdependent variables, for example, height, weight, daily caloric intake, genetic strain, etc. Data samples involving several variables are called multivariate. Many multivariate analytical methods involve inference for the parameters (means, variances, and correlation coefficients) based on multivariate normal distribution. One such method is known as discriminant analysis and is concerned with the problem of distinguishing between two or more populations on the basis of observations of a multivariate nature. Principal components analysis, cluster and factor analysis seek to determine relatively few out of possibly many variables that will serve to “explain” the variability or the interrelationships in the variables. Principal component analysis (PCA) would specifically make each successive component account for as much as possible of the remaining variability uncorrelated with previously determined components. In Fig. 7, data points from two variables are represented. Coordinates of data points on the abscissa axis correspond to values of the first variable and coordinates on the ordinate axis correspond to values of the second variable. PCA is able to find a first principal axis (labeled one) that accounts for most of the variance of the data. The second principal axis (labeled two) has to be perpendicular to the first principal axis and accounts for the remaining of the variance.

Recent progresses in signal processing and information theory have seen the development of blind source separation methods, which attempt to find a coordinate frame onto which the data projections have minimal overlap. For example, if two sources of sounds (e.g., a conversation and a CD player) are recorded simultaneously in the same room on two microphones, the sound signal from the two sources are mixed on both microphones. Coordinates of data points in Fig. 7 could represent the signal recorded from the two microphones. Separating the two sound sources from the microphone signal is called blind source separation. Independent component analysis (ICA) is a family of linear blind source separation methods. The core mathematical concept of ICA is to minimize the mutual information among the data projections. PCA components are orthogonal as shown in Fig. 7, which is usually not a realistic assumption for bio-physical data. To find biologically plausible sources, PCA must be followed by an axis rotation procedure, and ICA can be viewed as a powerful rotation method. ICA seeks to find axes for which the projection of data is maximally non-normal (i.e., contains the maximum amount of information). It uses the property of the central limit theorem in statistics that states that any linear mixture of two or more source activities is more normal that the original source activities, so, by finding axes that maximize non-normality, source separation may be achieved. As can be seen in Fig. 7, ICA is free to adapt to the actual projection patterns of source generators, if their activity time courses are (near) independent of one another. Performing ICA decompositions is most appropriate when sources are linearly mixed in the recorded signal, without differential time delays.

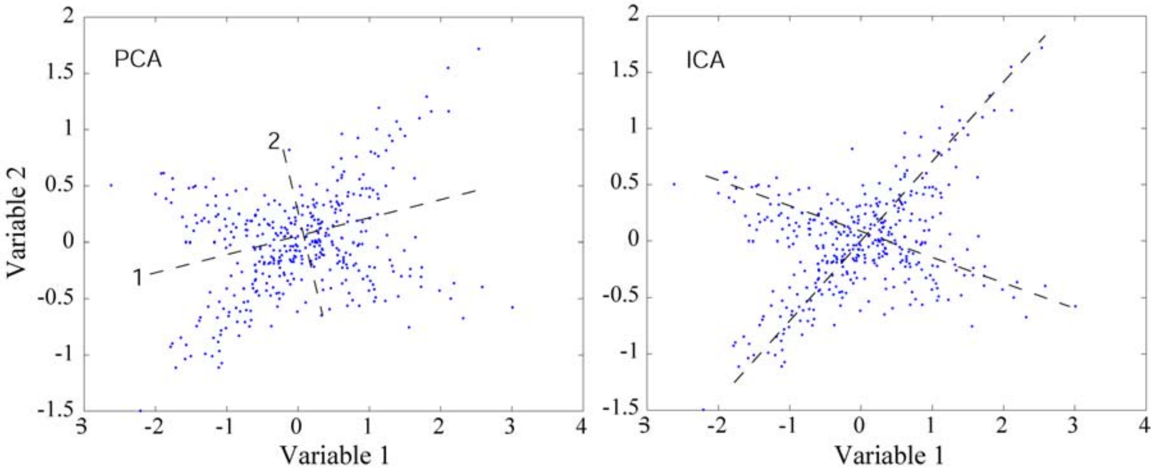

Fig. 7. Illustration of PCA and ICA algorithms. PCA finds axis with maximum variance. By contrast in ICA, the projection of data point on ICA axis is maximally independent.

ICA is being applied to various biomedical signal processing problems that include performing speech and noise separation [19], decomposing functional resonance imaging data [20], and separating brain area activities and artifacts mixed in electro-encephalography scalp sensors [21].

Texts that give broad coverage of multivariate analysis are [22-24].

CLINICAL TRIALS

A clinical trial is not a method per se but is a term applied to any form of planned experiments that involves human patients. The purpose of a clinical trial is to evaluate and verify the efficacy and safety of a new treatment or sets of treatments for a given medical condition. Although most of the analytical methods employed for clinical trials are the same as in other contexts, there is a special effort to avoid bias, which leads to some unique designs. Another distinguishing characteristic of clinical trials is the constraint imposed by studying living patients and the often difficult ethical considerations that must be addressed.

Double blind: The usual method to avoid bias in experimental designs is the random allocation of experimental subjects to treatments, but this will generally not suffice in clinical trials. A major potential source of bias is when subjects or evaluators in a trial know which treatment (e.g., placebo or active) is being received. In double-blind trials, neither the subject nor the evaluators are aware of which treatment is being received. Sometimes ethical or practical considerations make double-blinding infeasible, and sometimes partial blinding, for example, independent “blinded” evaluators only, may be sufficient to reduce bias in treatment comparison.

Within patient studies versus across patient studies: Most clinical trials are conducted as parallel studies in which two or more treatments are evaluated concurrently in separate groups of patients. As many researchers remain reluctant to assign patients randomly to new or standard treatments, current patients on the new treatment may be compared with data external to the study containing patients who had received standard treatments. Such an approach invites severe bias, since there is no assurance that treatment and control groups do not differ with respect to some factors other than the treatment itself. In crossover studies, each patient receives in succession two or more treatments. When feasible, such “within-patient” studies require smaller sample sizes than “between-patient” studies to achieve the same level of significance.

Lifetime variables: Some clinical studies are conducted as life data analysis and survival studies, and require specific statistical tools. In such studies, a variable represents the time to the occurrence of some event of interest, and is called a lifetime variable. In the engineering context, a life test consists of monitoring the operation of a sample of devices and to observe causes of and times to failure for all or some of the devices. In the clinical context, a survival study may involve observing cause of death (and time from entry to the study until death occurs) for some potentially fatal or, in the case of animal studies, induced disease. Alternatively, the event of interest may be time to relapse or time to remission for some diseases or conditions. The purpose of life tests or survival studies is to estimate or to compare lifetime or survival between different treatment groups.

Statistical test for lifetime variables: Since a lifetime variable must be positive (number of remissions for instance), the normal distribution is not usually a suitable model. The normal-based methods of multiple regression and analysis of variance cannot be used in the usual manner and in general requisite mathematical and computational methods are much less tractable than normal-based methods. Consequently, a non-parametric, partially parametric, or non-normal distributional analytic approach is taken. Data is usually visualized using Kaplan-Meier survival curves where censored patients (patients that have left the study) are explicitly indicated on the curve. Comparing between unpaired groups usually involve a log-rank test or a Mantel-Haenszel test. Conditional proportional hazards regression may be used to compare between two or more paired groups. Finally Cox proportional hazard regression may be used to compare between more than two unpaired groups and perform regression analysis.

Censoring: As mentioned above, a further complicating factor for survival studies is censoring. Under censoring, exact lifetimes are known only for a portion of the experimental units, the remainder known only to exceed certain censoring times. Censoring is usually a practical necessity and must be preplanned. For example, a life test on a random sample of 100 devices that has median time-to-failure of 2500 h will likely take over a year to complete if the tests were to continue until all devices fail. Instead, the test might be terminated at some predetermined time, for example, 1000 h, or immediately upon achieving some predetermined number of failures, for example, 30. These are called Type I and Type II censoring, respectively, and are the simplest to deal with. A distinguishing characteristic of survival studies involving human patients is that censoring times are often random. For example, suppose patients with a certain cancer are undergoing different chemotherapy treatments. Patients may enter the study in a random manner and patients may survive the termination time of the study or may die due to causes unrelated to the cancer. There are probability models that incorporate these data and lead to appropriate statistical inferential techniques. For example, some techniques assess the effectiveness of different treatments by comparing estimated mean survival times with the effect of unrelated causes of death removed. We shall not discuss any further methods used for dealing with censoring. It is sufficient to say that the special problems of statistical inference in the presence of censoring necessitate the use of large sample approximations and computer-aided numerical solutions. Some of these methods incorporate strong assumptions which users should be aware of.

Extensive treatment of methods for censoring and the analysis of survival data is given in [25-27]. Non-technical discussions of clinical trials and the special statistical treatments they require are given by Pocock [28] and Shapiro [29].

STATISTICAL COMPUTING AND SOFTWARE

Standardized computer programs aiming at performing a variety of statistical analyses were developed through the 1960s at several universities and became widely available in the 1970s. There is now a large number of them and the one to use will depend on the users expertise in statistics and field of research. For infrequent usage on small data samples and testing of simple hypothesis (χ^2, *t*-test, ANOVA), MS Excel which is usually already installed on many computer desktops may be sufficient. Note the availability of extra statistical functions when one selects the “Analysis Toolpack” add-in (installed but inactive by default). However MS Excel is not a statistical software per se, so to go beyond exploratory analysis stages it is better to rely on professional statistical software.

The best known and most comprehensive of these, all now under privately managed companies, are the Statistical Package for the Social Sciences (www.spss.com), the Statistical Analysis System (www.sas.com), and JMP (www.jmp.com). SPSS, as its name suggests, was developed primarily for use by social scientists and is relatively easy to learn by individuals with limited statistical and computer backgrounds. SPSS graphical interface is organized as tabular spreadsheets similar to MS Excel. The programs comprising SPSS, their output format, and the examples in the manuals retain a social science flavor. SAS has evolved into a widely utilized and extremely flexible package that is generally regarded to be more statistically sophisticated and complete than SPSS. JMP, also developed by the SAS institute, is a user-friendly graphical interface that sequentially guides the user through all stages of the experimental design and data analysis.

Apart from the graphical packages mentioned above, most other statistical softwares rely on command line calls, where users call functions from a prompt (note that most of these softwares also include menus). The free R software (www.r-project.org) offers powerful functions contributed by leading statisticians in the world. Because it is an open source project, it is used by many scientists and its extensive libraries are probably the place to look for rare statistical procedures. The Biomedical Programs (BMDP) (www.statsol.ie) contains a large variety of elementary and advanced statistical procedures. The programs are widely applicable, but some are particularly appropriate in biomedical contexts such as repeated measures ANOVA designs (see ANOVA). The S-plus software is also very popular (www.insightful.com) and very similar to R. It is based on the S language developed at AT-T. Finally, a widely used package in academia as well as in industry is a package called MINITAB (www.minitab.com) which is one of the most user friendly command line software.

There are many smaller, less comprehensive statistical analyses packages available for computers. These range from packages that perform elementary, mostly descriptive analyses, to some that are rather sophisticated. For bootstrap and surrogate statistics, SAS software is preferred among graphical software, although it is possible to program bootstrap and surrogate data routines in SPSS. The R software contains the majority of such user-contributed routines and S-Plus also contains a few of them. Finally MATLAB (www.mathworks.com), an interpreted language widely used in engineering, also has a large number of user-contributed bootstrap and surrogate statistics routines available.

Caution against the ignorant use of computerized statistical analyses cannot be over-emphasized. In planning studies, the methods of analysis and the constraint they impose on experimental designs should be taken into consideration in advance. If not, much work and data collection efforts could be wasted. Worse still, misleading and even meaningless results are often given undeserved weight merely because they represent the voluminous output of computer programs. How often do we hear that “a computer analysis shows….”, but such programs can be totally inappropriate. For example, the mathematical methods underlying repeated-measures ANOVA incorporate restrictive assumptions on the normality of the data and the experimental design for appropriate randomization of events. Although, these considerations are often ignored, researchers should systematically assess the degree to which test-related assumptions are satisfied. These facts notwithstanding, computer-aided data management and analysis can be of great benefit if used properly and wisely.

BIBLIOGRAPHY

This list is not meant to be comprehensive. For the naïve reader, a basic introduction to statistics with a plethora of exercises is given in the Schaum’s outline series on statistics [2]. For the non-naïve reader in statistics, a more technical yet still accessible reference is [30]. Other texts dealing with general statistical methods, particularly regression and analysis of variance are [31, 32]. Comprehensive web references are [18, 33, 34].

Statistical books have also been written for specific research topics. For example, see [35] for a beginner's reference in designing biology experiments and [6, 8-10] for more detailed references. As already mentioned, see [28, 29, 36, 37] for clinical trials. Finally, a recent development in statistics is statistical process control which deals with optimizing production and quality in the industry [38].

Acknowledgements: parts of this article were adapted from a previous version of this encyclopedia paper by P. Sullo and L.E. Ostrander from the Rensselaer Polytechnic Institute. The author also wishes to thank anonymous reviewer for their invaluable comments.

REFERENCES

[1] J. L. Gill, Design and Analysis of Experiments in the Animal and Medical Sciences. Ames: Iowa State Univ. Press, 1978.

[2] M. R. Spiegel and L. J. Stepens, Schaum's outlines in Statistics, 3 ed: McGraw Hill, 1999.

[3] C. E. Bonferroni, "Sulle medie multiple di potenze," Bollettino dell'Unione Matematica Italiana, 5 third series, pp. 267–270, 1950.

[4] S. Holm, "A simple sequentially rejective multiple test procedure," Scandanavian Journal of Statistics, vol. 6, pp. 65-70, 1979.

[5] F. Hoppe, Multiple Comparisons, Selection, and Applications in Biometry: A Festschrift in Honor of Charles W. Dunnett: Marcel Dekker, 1992.

[6] D. R. Cox, Planning of Experiments. New York: John Wiley & Sons, 1992.

[7] R. Norman, N. R. Draper, and H. Smith, Applied Regression Analysis, 2nd ed. New York: John Wiley & Sons, 1998.

[8] T. J. Lorenzen and V. L. Anderson, Design of Experiments: A No-Name Approach. New York: Dekker, 1993.

[9] G. E. P. Box, W. G. Hunter, and J. S. Hunter, Statistics for Experimenters. New York: John Wiley & Sons, 1978.

[10] W. G. Cochran and G. M. Cox, Experimental Designs. New York: John Wiley & Sons, 1992.

[11] C. R. Hicks and K. V. Turner, Fundamental Concepts in the Design of Experiments. Oxford University Press, 1999.

[12] B. J. Winer, D. R. Brown, and K. M. Michels, Statistical Principles in Experimental Design, 3 ed. New York: McGraw-Hill, 1991.

[13] J. Neter and W. Wasserman, Applied Linear Statistical Models, 4 ed. McGraw-Hill/Irwin, 1996.

[14] W. J. Conover, Practical Nonparametric Statistics Methods, 3 ed. New York: John Wiley & Sons, 1998.

[15] M. Hollander and D. A. Wolfe, Nonparametric Statistical Methods, 3 ed. New York: John Wiley & Sons, 1999.

[16] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. Chapman and Hall, 1994.

[17] R. Blair and W. Karniski, "An alternative method for significance testing of waveform difference potentials," Psychophysiology, pp. 518-524, 1993.

[18] D. C. Howell, "Resampling Statistics: Randomization and the Bootstrap," 2005. http://www.uvm.edu/~dhowell/StatPages/Resampling/Resampling.html

[19] H.-M. Park, H.-Y. Jung, T.-W. Lee, and S.-Y. Lee, "On Subband-Based Blind Signal Separation for Noisy Speech Recognition," Electronic Letters, vol. 35, pp. 2011-2012, 1999.

[20] J. R. Duann, T. P. Jung, S. Makeig, and T. J. Sejnowski, "fMRLAB: An ICA Toolbox for fMRI Data Analysis," presented at Human Brain Mapping, Sendai, Japan, 2002.

[21] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," J Neurosci Methods, vol. 134, pp. 9-21, 2004.

[22] T. W. Anderson, An Introduction to Multivariate Statistical Analysis. New York: John Wiley & Sons, 2003.

[23] R. Gnanadesikan, Methods for Statistical Data Analysis of Multivariate Observations. New York: John Wiley & Sons, 1997.

[24] J. Stone, Independent Component Analysis : A Tutorial Introduction. The MIT Press, 2004.

[25] J. D. Kalbfleisch and R. L. Prentice, The Statistical Analysis of Failure Time Data, 2 ed. New York: John Wiley & Sons, 2002.

[26] J. F. Lawless, Statistical Models and Methods for Lifetime Data. New York: John Wiley & Sons, 2002.

[27] R. Miller, Survival Analysis. New York: John Wiley & Sons, 1998.

[28] S. J. Pocock, Clinical Trials: A Practical Approach. New York: John Wiley & Sons, 1984.

[29] S. H. Shapiro, Clinical Trials. Dekker: New York, 2004.

[30] W. Hays, Statistics, 5 ed. Wadsworth Publishing, 1994.

[31] A. Afifi, V. A. Clark, and S. May, Computer-Aided Multivariate Analysis, 4 ed. Chapman & Hall/CRC, 2004.

[32] G. W. Snedecor and W. G. Cochran, Statistical Methods, 8 ed. Ames: Iowa State Univ. Press, 1989.

[33] R. Lowry, "Concepts and Applications of Inferential Statistics," 1999. http://faculty.vassar.edu/lowry/webtext.html

[34] J. Wasson, "Statistics in Educational Research - An Internet Based Course." http://www.mnstate.edu/wasson/ed602.htm

[35] A. Cann, Maths from Scratch for Biologists: John Wiley & Sons, 2002.

[36] J. W. Tukey, "Some thoughts on clinical trials, especially problems of multiplicity," Science, pp. 198:679, 1977.

[37] T. C. Chalmers, P. Celano, H. S. Sacks, and J. S. Jr., "Bias in treatment assignment in controlled clinical trials," N. Engl. J. Med., pp. 309:1358, 1983.

[38] R. Amsden, H. Butler, and D. Amsden, SPC Simplified: Practical Steps to Quality. Quality Resources, 1998.