University of Veterinary and Pharmaceutical Sciences Brno

Basics of Statistics

for Students of Veterinary Medicine

Doc. RNDr. Iveta Bedáňová, Ph.D.

Brno, 2019

Contents

Preface

There is no doubt that Statistics is a discipline which is a necessary part of education in every biological and medical (both human and veterinary) science. Its importance results from principle of obtaining, analysing, presentation and interpretation of data in medical sciences. Our knowledge and experience in the area of medicine would be full of many errors and mistakes if we would not have statistics. We would not be able to properly interpret the knowledge and experience we gained in the area of medicine without use of statistical methods. From this point of view the statistics is among basic subjects of veterinary medicine education. Contents of the discipline Statistics for students of veterinary medicine predominantly consists of statistical methods for description of data sets and testing of hypotheses for quantitative and qualitative data with specialization on data and hypotheses used in the area of biology and veterinary medicine.

Statistics has a practical impact especially in the area of research and development in biological and medical sciences. It may also be important in the area of clinical veterinary practice and in the sphere of hygiene and ecology in case of inspections regarding the animal provenance food safety. Statistics education is particularly focused on the ability to solve specific problems of veterinary medicine with the use of biostatistical methods.

This textbook is designed especially for English speaking students of veterinary faculties at the University of Veterinary and Pharmaceutical Sciences Brno but it can also be useful for other users of statistics in the area of biological and medical sciences. It represents a basic collection of data-analysis techniques and it may serve as an introductory textbook of biostatistics, assuming no prior knowledge of statistics and mathematics. The text is intended to be a brief guide to choosing and applying the appropriate statistical tests and methods for analysis of biological and medical data. The textbook is compact enough to be used in the laboratory or in the field while providing sufficient details to give the user some knowledge of the theoretical basis for the methods covered. Because of the limited time that can be devoted to this discipline at the University of Veterinary and Pharmaceutical Sciences Brno, the statistical techniques described in this textbook are sometimes simplified in order to be understandable also to non-mathematically specialised students of veterinary medicine as well as other biological disciplines.

We hope that this textbook, in spite of its concise form, will serve students of veterinary medicine as well as other users of statistics in the sphere of biosciences as a useful and handy tool covering basic statistical techniques and analyses.

Author

Brno, February 2007

Chapter 1
Basic concepts of statistics

1.1 Statistics – Importance and Use in Medicine and Biology

Statistics is the science that allows us to formulate and describe complex data in a short form, easily understood by all professionals. It allows us to compare data (numerical facts resulting from observations in some investigative monitoring) and gives us probabilities of the likelihood of studied events. The term "statistics" is often encountered as a synonym for "data": statistics of sickness rate during the last month (how many patients, number of cured patients), labour statistics (number of workers unemployed, number employed in various occupations), election statistics (number of votes in different regions, parties), etc. Hereafter, this use of the word "statistics" will not appear in this textbook. Instead, "statistics" will be used in its other common manner: to refer to the *analysis and interpretation of data with a view toward objective evaluation of the reliability of the conclusions based on the data*.

Statistics are predominantly needed in more ***probabilistic*** and ***less predictive*** sciences such as biology and applied biology (medicine). In a predictive science such as physics, to find out how fast a 300 g stone will reach the ground if dropped from a height of 30 m, one has only apply the data in the appropriate formula to obtain an accurate answer. In art, on the other hand, the evaluation of a given piece depends to a great extend on subjective criteria. Medicine falls somewhere in between. There are numerous and complex physical/chemical events occurring simultaneously which cannot be evaluated separately. For instance, if one wants to determine the time of induction, or return, of a given reflex of the specific tendon, the issue is more complicated than it appears initially. In this case we are dealing with transmission of electrical potential difference across many nerves and transmission to muscles, making relevant calculations more tedious and specific data less well known. Furthermore, the specific functions are affected by several other components of internal milieu such as the level of hormones. To complicate matters, this will represent just one out of many concomitant functions of a total inhomogeneous system, living body. It is, therefore, easy to appreciate why biological sciences in general, and applied biology such as medicine, in particular, are probabilistic in nature. As a result, a good grasp of statistics is essential for one to be effective in this field.

Statistics applied to biological problems is simply called **biostatistics** or, sometimes, *biometry* (the latter term literally meaning "biological measurements"). As biological entities are counted or measured, it becomes apparent that some objective methods are necessary to aid the investigator in presenting and analysing research data. Although the field of statistics has roots extending back hundreds of years, its development began in earnest in the late nineteenth century, and a major impetus from early in this development has been the need to examine biological and medical data.

Nowadays the statistical methods are common and more and more important in all biological and medical sciences. Biostatistics is a necessary part of every biology and medicine (both human and veterinary) education. When dealing with living organisms, we must always keep in mind, that every individual is unique and there is a high level of insecurity regarding its reactions. Therefore all data obtained from biological objects may be very different and variable. This results from vast genetic variability in living organisms and also from other aspects (ambient environment, adaptability, etc.).

This large **variability of biological data** causes problems and difficulties in monitoring, measurements and data acquisition in animals and other living organisms. These problems can partially be solved by means of statistics, because only statistical methods are able to take into account this great variability of biological data, evaluate them and give correct inferences concerning studied biological objects. Statistics handles variability in two ways. First it provides precise ways to *describe and measure the extent* of variability in our measured data. Secondly it provides us with methods for using those measures of variability to *determine a probability* of the correctness of any conclusions we draw from our data.

Before data can be analysed, they must be collected, and statistical considerations can aid in the design of experiments and in the setting up of hypotheses to be tested. Many biologists attempt the analysis of their research data only to find that too few data were collected to enable reliable conclusions to be drawn, or that much extra effort was expended in collecting data that cannot be of ready aid in the analysis of the experiment. Thus, knowledge of basis statistical principles and procedures is important even before an experiment is begun.

Once the data have been obtained, we may organize and summarize them in such a way as to arrive at their orderly and informative presentation. Such procedures are often termed ***descriptive statistics***. For example, tabulation might be made of the heights of all students of the Faculty of veterinary medicine, indicating an average height for each sex, or for each age. However, it might be desired to make some generalizations from these data. We might, for example, wish to make reasonable estimate of the heights of all students in the university. Or we might wish to conclude whether the males in the university are on the average taller than the females. The ability to make such generalized conclusions, inferring characteristics of the whole from characteristics of its parts, lies within the realm of *inferential statistics*.

Summary: Use of Biostatistics in the Area of Veterinary Medicine:

1. Research – analysis of data measured during experiments, e.g. in the course of examination of effects of new drugs, remedies, feed mixtures, medical treatments, methods, etc. We can confirm or reject hypotheses that are investigated in experiments by means of special statistical methods (statistical hypotheses testing - inferential statistics).

2. Clinical practice – evaluation and generalization of observation results in clinical practice – i.e. monitoring and comparison of disease incidence in different groups of animals, in regions, time periods, etc. (descriptive statistics). We can for example compare the sickness rate in animals observed in a stable to the statistical sickness rate (known in the whole population from long-term monitoring). Statistical methods (inferential statistics) help us to decide, whether the possible increase in sickness rate in the stable is only random, or whether it is caused by some ambient causes (e.g. bad treatment, feeding stuff or conditions in the stable).

1.2 Types of Biological Data

In biological and medical sciences, we analyse biological properties of living organisms that are described on the basis of selected biological characters. These biological characters can be measured usually by some means. Their values differ from one entity to another – therefore they are called *variables* in statistics. Variables describe studied biological characters (properties of living organisms usually) and they can quantify (more or less) these biological properties. Different kinds of variables may be encountered by biologists, and it is desirable to be able to distinguish among them. Variables can be quantitative or qualitative. **Quantitative** variables record the *amount* of something (ordinal data and numerical data); **qualitative** variables describe the *category* to which the data ca be assigned and are therefore sometimes referred to as categorical data.

Exactness of those biological variables can differ in their values – according to the exactness we can distinguish between 3 types of biological data in statistics:

A. Nominal Data (Categorical)

Sometimes the variable under study is classified by some quality it possesses rather than by a numerical measurement. In such cases the variable may be called an *attribute*, and we are said to be using a *nominal scale* of measurement (categories, classes). On a nominal scale ("nominal" is from the Latin word for "name"), animals might be classified as male or female, as left- or right-handed, as ill or healthy, with or without horns, as vaccinated or not vaccinated, as alive or dead, etc. Such variables describe only some quality in a living organism that is not measurable – there are no values (it is the *lowest level of quantification* in variables). We can only distinguish between 2 possibilities (situations): the quality is either present or not present in each individual observed. Sometimes data from an ordinal or numerical scale of measurement (see below) may be recorded in nominal-scale categories. For example, heights may be recorded as tall or short, or performance on an examination as pass or fail, etc.

B. Ordinal Data (Rank-Order Data)

They are represented by an ordering (up or down) of observations based on subjective scale given by an evaluator (experimenter). These data are arranged into an ascending or descending row and may be a record only of the fact that one individual has lower intensity of a studied biological character than the other (with no indication of how much more). Differences between various degrees on the scale are different and dependent on a measure given by the evaluator (e.g. classification using grades in school, points in breeding competitions, marks for animal behaviour in experiment, etc.). Thus, we are dealing with relative differences rather than with quantitative differences. Such data that consists of an ordering or ranking of measurements are said to be on an ordinal scale of measurement ("ordinal" being from the Latin word for "order"). One may speak of one biological entity being shorter, darker, faster, or more active than other; the sizes of five cell types might be labelled 1, 2, 3, 4, and 5, to denote their magnitudes relative to each other; or success in learning may be recorded as A, B, C, D, E, or points gained in breeder competitions, etc.

C. Numerical Data

They are represented by exact numeric values obtained by means of some *objective measurement* (meter, thermometer, scale, measuring device etc.). Differences between various degrees on the scale are uniform – the numerical scale consists of the same intervals. There is the

highest level of quantification in statistical data – they are most often used for statistical evaluation. Numerical variables allow us to record an amount for each observation and to compare the magnitude of differences between them. They can be either *continuous* or *discrete*.

- ***Discrete Data (discontinuous)***
  Variables that can take only specific available values – most often integer numbers. For example the number of bacterial colonies on a Petri dish can only be a positive integer value; there can be 24 colonies, but never 24.5 colonies or -24 colonies. Similarly also number of laid eggs, puppies in a litter, animals in a stable, patients, cells, etc.
- ***Continuous Data***
  These variables can take on any conceivable value in our infinite spectrum of real numbers - within any observed real range (height, length, weight, volume, body temperature, concentration of enzyme, etc.)

Different categories of statistical data have their own specific statistical method used for their examination. These methods are differently exact according to the exactness of data category. Statistical methods used for numerical or ordinal data are more exact and generally they are not applicable to nominal data (since nominal data contain only little information for exact methods). It is possible reversely: the less exact methods intended for nominal (or ordinal) data are useful also for numerical data. In this case we can purposely use these not very precise methods e.g. for preliminary analyses that must be performed quickly.

Sometimes the distinction between different types of data is not very obvious, e.g. categories fall into a natural order and it is not reasonable to distinguish between ranked and categorical data. Values of heights in students are continuous data. Of course, they may also be ranked: smallest, next smallest, ……., highest. If the height is categorized into three groups, <160, 161-180, >181, the values are still ordered, but we have lost a lot of information. There are so many ties that analysis methods will be sensitive only to three ranks, one for each category. Although we could analyse them as categories A, B, and C, the methods treating them as ranks first, second, and third are still "stronger" methods. When rendering data into categories, one should note whether the categories fall into a natural order. If they do, treat them as ranks. For example, categories of ethnic groups do not fall into a natural order, but the pain categories severe, moderate, small, and absent do.

Note that we can always change data from higher to lower level of quantification, that is, continuous to discrete to ranked to categorical, but not the other way. Thus, it is always preferable to record data as high up on this sequence as possible; it can always be dropped lower.

1.3 Statistical Sets

Basic to statistical analysis is the desire to draw conclusions about a group of measurements of a variable being studied. Statistical variables in biological and medical sciences may be studied in groups (statistical sets) of living individuals – animals, plants, cells, bacteria, etc. (in general *entities* or *items*). We can distinguish between 2 types of statistical sets in statistics:

- **The Population** (the Universe) : **N = ∞** (number of members)

By the word population, we denote the entire set of subject about whom we want information. Thus, the population means also "all items" (individuals) that could show studied variable. If we were to take our measurements on all individuals in the population, descriptive statistics would give us exact information about the population. Populations are often very large sets; number of individuals in the population is considered to be "endless" (for statistical purposes and calculations). In practice, the number of members in the population can be literally "endless" – especially from the time viewpoint: e.g. body weights in all cattle in CR, dogs in Europe, etc. – number of individuals is not fixed (it fluctuates since new members are born and others die).

Occasionally populations of interest may be relatively small, such as the ages of men who have travelled to the moon or the heights of women who have swum the English Channel. If the population under study is very small, it might be practical to obtain all the measurements in the population. Generally, however, populations of interest are so large as to render the obtaining of all the measurements unfeasible (it's time-consuming, expensive, etc.). We are not able to obtain all possible measurements from the population in practice – for example, we could not reasonably expect to determine the body weight of each dog in Europe. What can be done in such cases is to obtain a subset of all measurements in the population. This subset of measurements comprises a *sample*, and from the characteristics of samples conclusions can be drawn about the characteristics of the populations from which the samples come.

Often one samples a population that does not physically exist. Suppose an experiment is performed in which a food supplement is administered to thirty piglets, and the sample data consist of the growth rates of these thirty animals. Then the population about which conclusions might be drawn is the growth rates of all the piglets that conceivably might have been administered the same food supplement under identical conditions. Such a population is said to be "imaginary" and is also referred to as "hypothetical" or "potential".

- **The Sample** (the Subset) : **n** (number of members)

By the word sample, we denote definite number (marked as *n*) of individuals selected from population. We perform measurements in this sample in practice calculate descriptive statistics from the sample and use them to estimate the true characteristics of the population. The definite (often quite small) number of members in samples measured implies some inaccuracies in examinations and calculations performed on the basis of these small subsets in comparison to the whole population.

To reach the most valid conclusions about a population, the sample must be a ***representative*** subset of the population the sample must fully represent the population in its characteristics. For example, male dogs tend to be heavier than females because they tend to be bigger. We could be led into making wrong decisions on the basis of weight if we generalized about dogs from a sample containing only males. We would say this sample is *biased*. To avoid bias, our sample should contain the same proportion of males as the dog population contains male dogs.

Samples from populations can be obtained in a number of ways; however, to reach valid conclusions about populations by induction from samples, statistical procedures typically assume that the samples are obtained in a ***random*** fashion. To sample a population randomly requires that each member of the population has an equal and independent chance of being selected. That is, not only must each measurement in the population have an equal chance of being chosen as a member

of the sample, but the selection of any member of the population must in no way influence the selection of any other member.

It is sometimes possible to assign each member of a population a unique number and to draw a sample by choosing a set of such numbers at random. This is equivalent to having all members of a population in a hat and drawing a sample from them while blindfolded. We may not to do a subjective choice generating a random sample; we can use e.g. table of random digits from statistical literature (E.g. Zar: Biostatistical Analyses), drawing lots for registration numbers of animals in a stable, etc.

Another requirement for a dependable generalization about certain characteristics is an appropriate size of sample. The pattern of sample values (as well as sample descriptive characteristics) gets closer in nature to the pattern of population values as the sample size gets closer to the population size. It means that the bigger is our sample, the better, but there are practicable limits in practice - not enough time, money, etc. Thus, we must do compromises often in practice; in general the sample size above 30 members is considered to be an appropriate sample size to give us results of calculations, which are comparable to population. However, samples that consist e.g. only of 10 individuals may be sufficient in some cases in practice.

1.4 Characteristics of Variables

Data obtained through measurements in a random sample represent a ***random variable*** (discrete or continuous), that can be described by means of some specific terms used in statistics:

Variant Sequence – a listing (rank) of all the observed values (variants) measured in a sample, that are arranged up (in an ascending row) or down (descending row).

E.g.: 2, 3, 4, 4, 5, 5, 5, 6, 6, 7, 7, 8 (discrete data – number of pups in a litter).

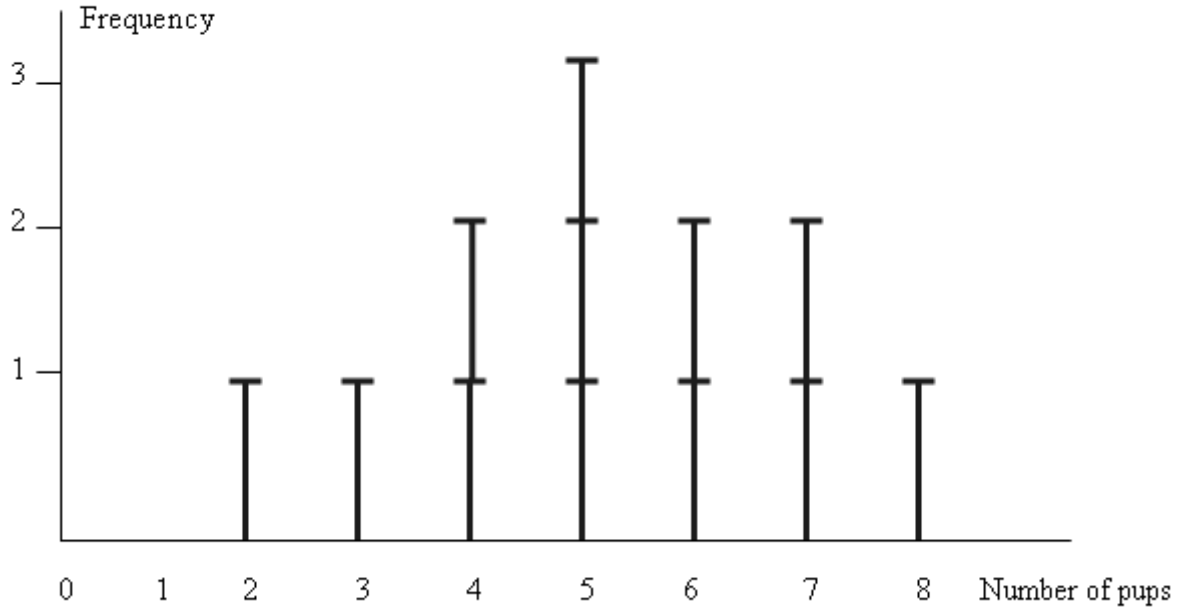
Note that some values are repeated in the row – this repetition (multiplication) of certain value is called frequency of this variant.

Frequency of the Variant – how many times each value (variant) is observed (repeated) in the sample (e.g.: frequency of values 2, value 3, and value 8 is 1 (these values occur only once in the row), frequency of value 5 is 3 (value 5 is repeated 3 times in the row).

Frequency Distribution – a graphically presented distribution of all the observed frequencies in the sample (among the various values – variants). It is the way the data are distributed along the scale (or axis) of the variable of interest.

The frequency distribution is a concept through which most of the essential elements of biological and medical statistics can be assessed. The graphical presentation of frequency distribution can have various forms that may be slightly different in discrete data and continuous data. Bar charts with separate strokes are most often used for presentation of frequency distribution of discrete variables (see Fig.1.1 Frequency distribution for number of pups in a litter) and histograms (column charts) are common in continuous variables.

Fig. 1.1 Frequency Distribution – Discrete Data (Bar Graph)



When measuring continuous data we create classes i.e. equivalent intervals (categories) of data to simplify the situation. Number of classes should be appropriate according to the sample size:

Up to 100 items in the sample: we usually create 6 - 9 classes,

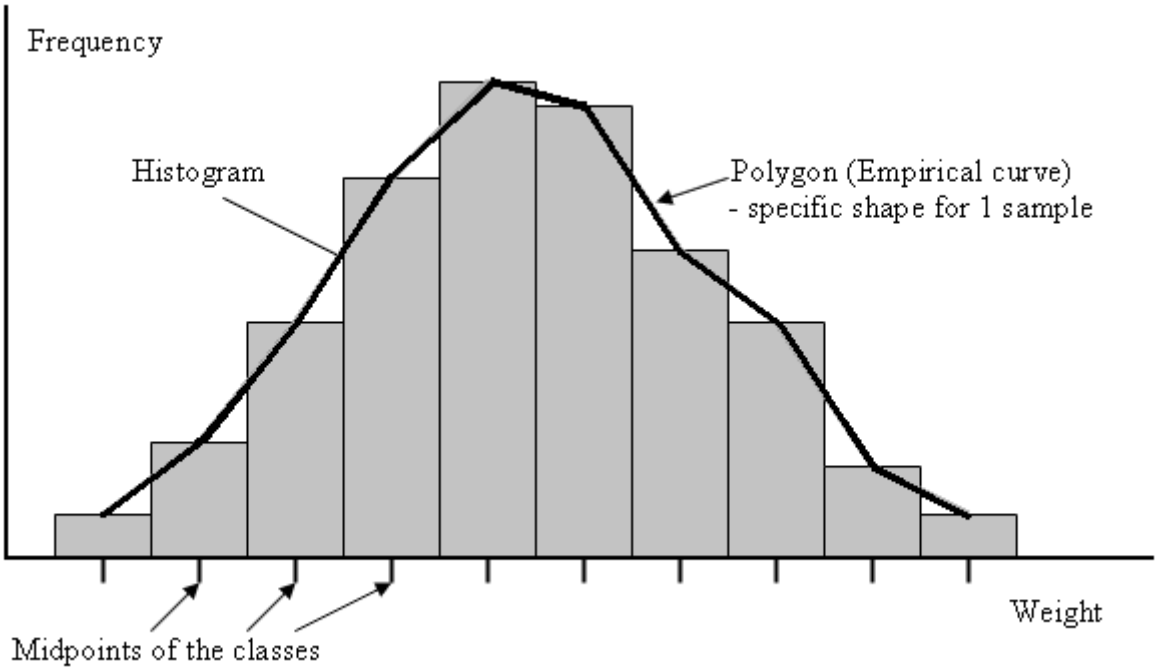
Up to 500 items in the sample: we usually create 10 -15 classes,

Above 500 items in the sample: we usually create 16 – 20 classes.

Then a certain number of values fall into each defined interval (class). All the data in this interval get the same value – so called *midpoint* (mean value) *of the class*. These values replace the original values measured in all individuals in the sample monitored. In this way, we are able to obtain a *frequency of the class* i.e. number of items (individuals) in the appropriate interval to draw a chart of frequency distribution for continuous data.

Fig.1.2 represents the frequency distribution of continuous data (e.g. body weights).

Fig. 1.2 Frequency Distribution – Continuous Data (Histogram)



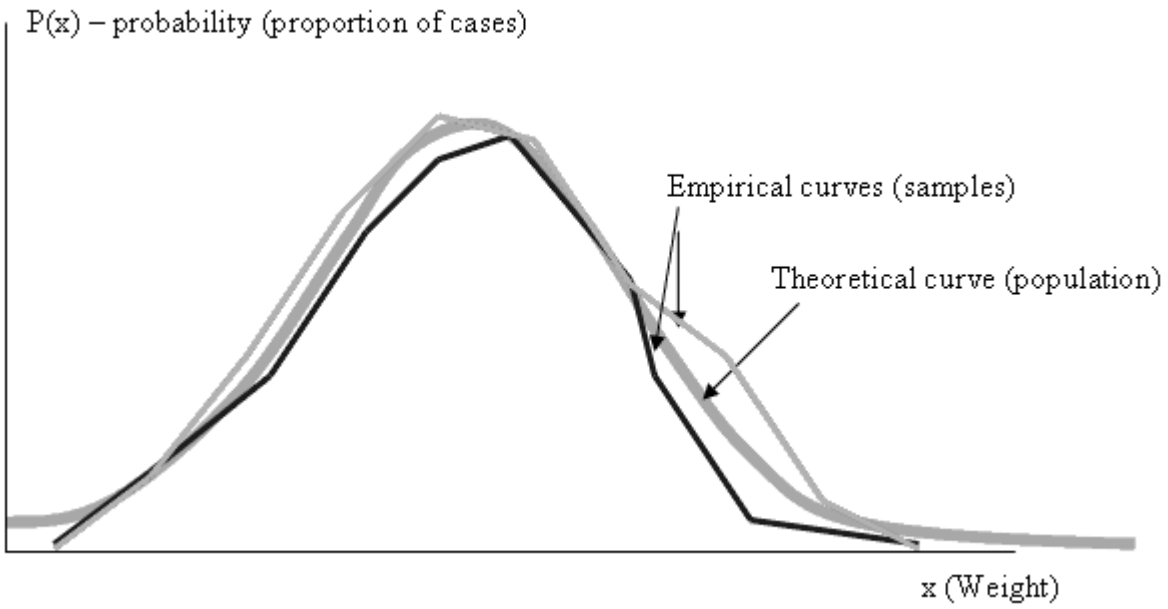
As you can see from the chart above, it is possible to use one more type of graphical presentation of frequency distribution in the sample - a polygon. Polygon is represented by a broken line that joins tops of columns in the midpoints of the classes. The shape of the polygon is specific for the unique sample that was used for our measurements i.e. the shape of polygon varies from sample to sample. Therefore we can use the term "empirical curve" for polygon as well (from Latin word empiria = experience).

Most of biological variables (both discrete and continuous) possess the characteristic property – frequencies in the middle of the sample (around the mean value) are the highest and frequencies of extremely small and large values in the sample are the lowest.

1.4.1 Probability distribution

When repeatedly measuring the same variable in various sample(s) selected from one population we get different shape(s) of empirical curve(s) – this results from genetic variability of individuals in the sample. Nobody (empirical curve) will have the same shape. Empirical curves for different samples (obtained from one population) are located along only one theoretical curve (continuous) that describes **probability distribution** of the variable in the **population**. In practice we can't measure all data for constructing of theoretical curve – its shape is only hypothetical (theoretically). However we can estimate its shape on the basis of empirical curves of samples selected from the population (Fig.1.3).

Fig. 1.3 Frequency (Probability) Distribution



The frequency distribution for the whole population is a *statistical distribution* that determines the probability of occurrence of values in studied variable; therefore we use the term probability P(x) instead of frequency on the axis y. The term frequency used with samples represents an absolute scale (that is possible only in samples - they have definite number of entities), whereas the term probability represents a relative scale (proportion of cases) that is necessary to be used in populations, where the number of entities is infinite.

1.4.2 Shapes of Probability Distributions

The probability distribution gives us the fundamental piece of statistical information of studied variables. It contains all the basic information we need for our statistical methods. From distribution, we can learn what is typical or characteristic of variable (e.g. where the majority of values are located in population, how much the values are variable etc.). Various biological characters can have different shapes of probability distribution.

Most variables in biological sciences follow normal (Gaussian) probability distribution that is symmetrical – most values in population are located around the mean value (Fig. 1.4). But some

variables in biology and medicine can behave in another way – they can follow different probability distributions, which have other shapes of their curves: asymmetrical - up to extreme (Fig. 1.5, 1.6) or non-normal, irregular (Fig.1.7):

Fig. 1.4 Normal (Gaussian) Distribution - symmetric bell curve

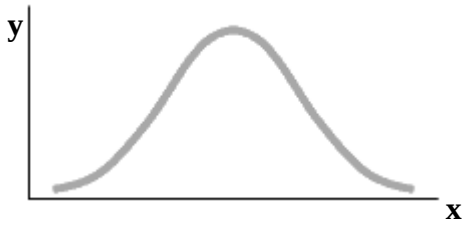


Fig. 1.5 Asymmetric (right-skewed, left-skewed) Distribution

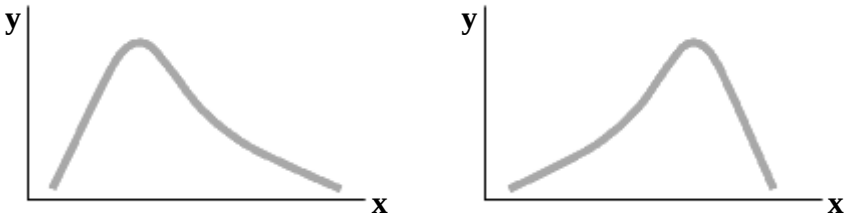


Fig. 1.6 Extreme (decreasing, increasing)

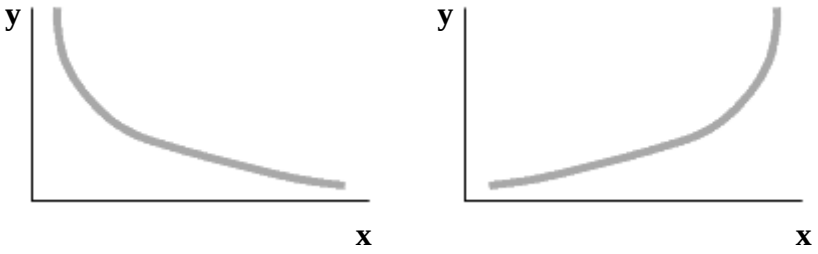
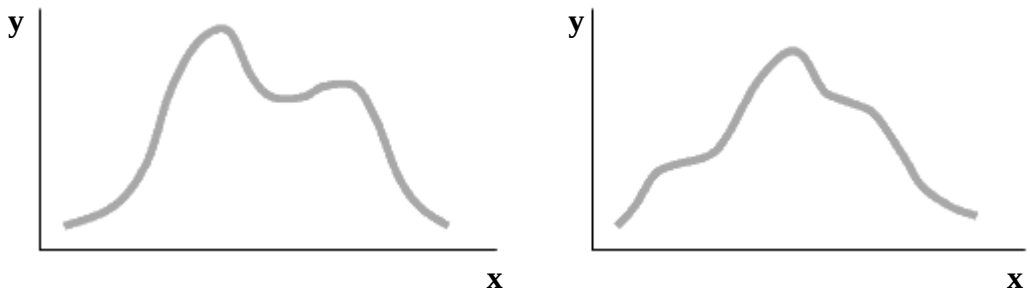


Fig. 1.7 Non-normal (unknown, irregular, 2 and more peaks)



1.4.3 Portions of Distribution

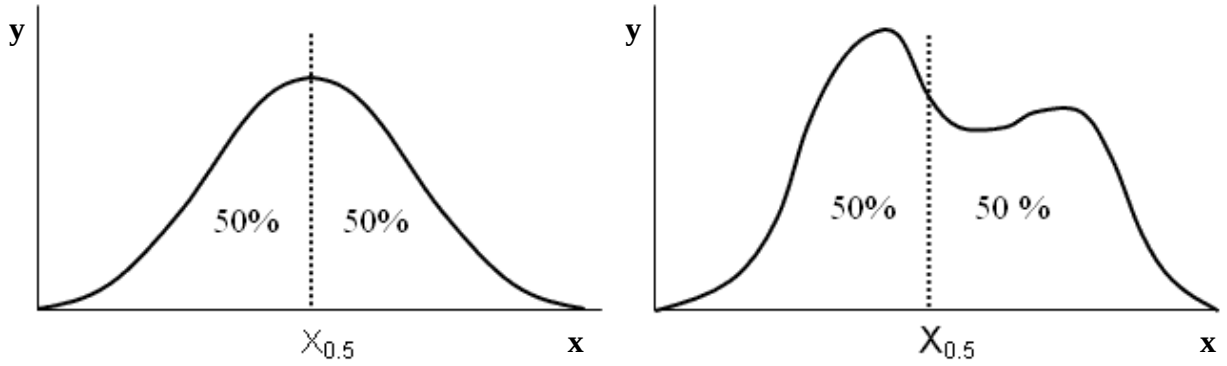
For each distribution we can define measures (**quantiles**) that divide a group of data - population (displayed as the area under the curve of distribution) into 2 parts (**portions**):

- Values which are smaller than quantile,
- Values which are larger than quantile.

There are specific quantiles used for description of distributions in statistics:

50% quantile – $x_{0.5}$ (called **median**) divides a group of data into 2 equal halves (Fig. 1.8):

Fig. 1.8 Median (50% quantile) in normal and non-normal distribution



Quartiles – divide a group of data into four equal parts,

Deciles – divide a group of data into 10 equal parts,

Percentiles – divide a group of data into 100 equal parts.

Important quantiles and their corresponding proportions of the most common distribution curves are tabulated in statistical tables and used as **critical values** in statistical hypotheses testing (see Chapter 5) or as **coefficients** in calculations (see Chapter 4: confidence intervals of statistical parameters – e.g. mean value μ, standard deviation σ).

Chapter 2
Descriptive Characteristics of Statistical Sets

The aim of statistical data evaluation: to get an **image of monitored biological** characters in the whole population on the basis of data samples. At first we usually classify observed sample data according to the measured values, arrange the variant sequences and draw graphs of frequency distributions. These arranged data us give basic information about the sample and offer source material for further statistical methods of data evaluation for monitored biological characters.

The deeper analysis follows, when we try to resume data information into one or several numbers by means of specific exactly defined ***parameters (statistical characteristics)***. We can't really determine the exact values of these parameters at the level of the whole population, therefore we select a sample (or several samples) from the studied population and we calculate the so-called ***statistics*** from this sample data. It serves as estimation of the exact population parameters.

Several measures help to describe or characterize a population. For example, generally a preponderance of measurements occurs somewhere around the middle of the range of a population of measurements. Thus, some indication of a population "average" would express a useful bit of descriptive information. Such information is called a *measure of central tendency*, and several such measures (e.g. the mean and the median) will be discussed below. It is also important to describe how dispersed the measurements are around the "average". That is, we can ask whether there is a wide spread of values in the population or whether the values are rather concentrated around the middle. Such a descriptive property is called a *measure of dispersion* (e.g. the range, the standard deviation, the variance etc.).

A quantity such as a measure of central tendency or a measure of dispersion is called a parameter when it describes or characterizes a population, and we shall be very interested in discussing parameters and drawing conclusions about them when studying a biological character in the population. However, one seldom has data for entire populations, but nearly always has to rely on samples to arrive at conclusions about populations. Thus, as mentioned above, one rarely is able to calculate the true exact parameters. However, by random sampling of populations, parameters can be estimated very well by means of special statistical methods (see the chapter Estimation of population parameters). Due to the statistical methods, we can determine so-called confidence intervals for population parameters or to calculate the estimates of population parameters called statistics (on the basis of sample data). It is statistical convention to represent the true population parameters by Greek letters and sample statistics by Latin letters.

Among the most often used measures of central tendency belong: the mean, the median, and the mode. Among the most often used measures of dispersion and variability of statistical sets belong: the range, the variance, the standard deviation, the coefficient of variability, and the standard error of mean.

2.1 Measures of Central Tendency

In samples, as well as in populations, one generally finds a preponderance of values somewhere around the middle of the range of observed values. The description of this concentration near the middle is an *average*, or a *measure of central tendency* to the statistician. It is also termed a *measure of location*, for it indicates where, along the measurement scale, the sample or population is located. Various measures of central tendency are useful parameters, in that they describe a property of populations. The characteristics of the most often used parameters and the sample statistics that are good estimates of them are described below.

2.1.1 The Arithmetic Mean *(average - AVG)*

Notation: μ (Population), $\bar{x}$ (Sample)

The most widely used measure of central tendency is the arithmetic mean, usually referred to simply as the mean, which is the measure most commonly called an "average" (the term "average" is used predominantly for sample statistic, the term "mean" is used for population exact parameter most often).

Each measurement in a population may be referred to as x_i value. The subscript *i* might be any integer value up through *N*, the total number of values *X* in the population.

The calculation of the population mean μ (the theoretical exact parameter):

$$\mu = \frac{\sum_{i=1}^{n} x_i}{N}$$

The most efficient and unbiased estimate of the population mean μ, is the sample mean, denoted as $\bar{x}$ (read as "x bar"). Whereas the size of the population (which we generally do not know) is denoted as N, the size of a sample is indicated by *n* (definite number of members in a specific sample used for measurements).

The calculation of the sample average $\bar{x}$:

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

Properties of the mean:

- Mean is affected by extreme values in the set (when changing one value x_i arithmetic mean change as well). The average is the correct measure of the central tendency of a sample only if the sample is homogenous enough in its values (*it should be used in homogenous regular distributions (Gaussian) only*). Otherwise, especially in small samples, the average can be

biased through possible extreme values in the sample, and does not represent the correct measure of centre in sample (in irregular distributions as well).

- It has the same units of measurement as the individual observations.
- $\sum(x_i - \bar{x}) = 0$ - the sum of all deviations from the mean will be always 0.

2.1.2 The Median

Notation: $\tilde{\mu}$ (Population), $\tilde{x}$ (Sample)

The median is typically defined as the middle measurement in an ordered set of data (ordered in an ascending or descending row). That is, there are just as many values larger than the median as there are smaller ones. The sample median is the best estimate of the population median. In a symmetrical distribution the sample is also an unbiased estimate of μ, but it is not as efficient a statistic as $\bar{x}$, and should not be used as a substitute for $\bar{x}$**.** If the frequency distribution is asymmetrical, the median is poor estimate of the mean.

The median of a sample of data may be found by first arranging the measurements in order of magnitude. Then the middle value of this row is the median.

In larger samples we can find out, which datum in a ranked sample data is median by means of the calculation of **rank** of this figure: $\frac{n+1}{2}$ (it can be applied for centre of any row of n values in math generally).

- *If the sample size (n) is odd* $\Rightarrow$ there is only 1 middle value (rank will be an integer) and indicates, which datum in ordered sample is the median.
- *If n is even* $\Rightarrow$ rank of the median is a half-integer and it indicates that there are two middle values, and the median is a midpoint (mean) between them.

Properties of median:

- Median is not affected by extreme values in the sample;
- Median = 50% quantile (divides the sample data into 2 halves: values that are smaller then median and values that are larger than median);
- It may be used in irregular (asymmetric) distributions – in this case median is better characteristic for the middle of the set than the average.

2.1.3 The Mode

Notation: $\hat{\mu}$ (Population), $\hat{x}$ (Sample)

The mode is commonly defined as the most frequently occurring measurement in a set of data (the value with the highest frequency). Mode always indicates the top of the distribution curve. A distribution with two modes (two tops) is said to be *bimodal* and may indicate a combination of two distributions with different modes (e.g. heights of men and women). The sample mode is the

best estimate of the population mode. When we sample a symmetrical unimodal population, the mode is an unbiased estimate of the mean and median as well as, but it is relatively inefficient and should not be so used.

The mode is a somewhat simple measure of the central tendency and it is not often used in biological and medical research, although it is often interesting to report the number of modes detected in a population, if there are more than one.

Properties of the mode:

- Mode is not affected by extreme values in the sample.
- It is not a very exact measure of the middle of the set.

2.2. Measures of Variability (dispersion)

Mean value indicates only the centre of sets but does not indicate the dispersion of values around the centre (how much the values are scattered). For this purpose we use measures **of variability** that describe this **dispersion** of values around the centre of the set, and determine also the reliability of mean value of the set - reliability will be larger in samples that have similar (homogenous) values and no extreme values. Measures of variability of a population are exact parameters of the population, and the sample measures of variability that estimate them are statistics.

2. 2. 1 The Range

The difference between the highest and the lowest value in the set of data is termed the range. If sample measurements are arranged in increasing order of magnitude, as if the median were about to be determined, then sample range R is calculated:

$$\mathbf{R = x_{max} - x_{min}}$$

Properties of the range:

- It is dependent on 2 extreme values of data
- It is a relatively rough measure of variability – it does not take into account any measurements between the highest and lowest value.

Furthermore, as it is unlikely that the sample will contain both the highest and the lowest values in the population, the sample range usually underestimates the population range. Nonetheless, it is considered useful by some to present the sample range as an estimate (although a poor one) of the population range. Whenever the range is specified in reporting data, however, it is usually a good practice to report another measure of variability as well.

It is evident that the range conveys no information about how clustered about the middle of the distribution the measurements are. As the mean is so useful a measure of central tendency, one might express dispersion in terms of deviations from the mean.

If we want to express variability in terms of deviations from the mean, there will be a difficulty: as the sum of all deviations from the mean, i.e., $\sum(x_i - \bar{x})$, will always equal zero, summation would be useless as a measure of dispersion and variability. Summing the absolute values of the deviations from the mean results in a quantity that is an expression of dispersion about the mean. Dividing this quantity by *n* yields a measure known as the *mean deviation* (or the *mean absolute deviation*) of the sample. This measure has the same units as do the data, but it is not very often used as a measure of dispersion and variability in practice.

Another method of eliminating the signs of the deviations from the mean is to square the deviations. The sum of squares of the deviations from the mean is called the sum of squares, abbreviated SS, and is defined as follows:

$$population\ SS = \sum (x_i - \mu)^2$$

$$sample\ SS = \sum (x_i - \bar{x})^2$$

By means of term "SS" we can define other measures of variability in the set of data:

2.2.2 The Variance

Notation: σ^2 **(Population),** s^2 **(Sample)**

The variance is defined as the mean sum of squares about the mean value of data. Sometimes this measure is called also mean square – short for *mean squared deviation*.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$ Population variance

The best estimate of the population variance, σ^2, is the sample variance, s^2:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$ "Estimated" variance (used for samples)

If, in equation of population variance, we replace μ by $\bar{x}$ and *N* by *n*, the result is a quantity that is a biased estimate of σ^2 and this would be not a correct measure of variability for sample data (especially in the case of small sample sizes). The dividing of the sample sum of squares by ***n-1*** (called ***degrees of freedom***, abbreviated DF or df) rather than by *n*, yields an unbiased estimate, and it is equation for "Estimated" variance that should be used to calculate the sample variance.

If all measurements are equal, then there is no variability and $s^2 = 0$. And, s^2 becomes increasingly large as the amount of variability, or dispersion, increases. Because s^2 is a mean sum of squares, it can never be a negative quantity.

The variance expresses the same type of information as does the mean deviation, but it has certain very important properties relative to probability and hypothesis testing that make it distinctly superior. Thus, the mean deviation is very seldom encountered in biostatistical analysis.

The calculation of s^2 can be tedious for large samples, but it can be facilitated by the use of the equality:

$$s^2 = \frac{\sum x_i^{\,2} - \frac{\left(\sum x_i\right)^2}{n}}{n-1}$$

This formula is much simpler to work with in practice, therefore is often referred to as a "working formula", or "machine formula" for sample variance. There are, in fact, two major advantages in calculating s^2 by this equation rather than by equation of previous (original) formula for "estimated variance". First, fewer computational steps are involved, a fact that decreases chance of error. On many calculators the summed quantities, Σx_i and Σx_i^2, can both be obtained with only one pass through the data, whereas the original formula for "estimated variance" requires one pass through the data to calculate $\bar{x}$ and at least one more pass to calculate and sum the squares of the deviations, $x_i - \bar{x}$. Second, there may be a good deal of rounding error in calculating each deviation $x_i - \bar{x}$, a situation that leads to decreased accuracy in computation, but which is avoided by the use of the latter formula above.

Variance has the *square units* as the original measurements. If measurements are in grams, their variance will be in grams squared, or if the measurements are in cubic centimetres, their variance will be in terms of cubic centimetres squared, even though such squared units have no physical interpretation.

2.2.3 The Standard Deviation (SD)

Notation: σ (Population), *s* (Sample)

The standard deviation is the positive *square-root of the variance*; therefore, it has the same units as the original measurements.

Thus, the formula for a **population SD** is:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \mu\right)^2}{N}} \qquad \text{or} \qquad \sigma = \sqrt{\frac{\sum x_i^{\,2} - \frac{\left(\sum x_i\right)^2}{N}}{N}}$$

and for a **sample SD**:

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}} \qquad \text{or} \qquad s = \sqrt{\frac{\sum x_i^{\,2} - \frac{\left(\sum x_i\right)^2}{n}}{n-1}}$$

Remember that the standard deviation is, by definition, always a nonnegative quantity. The standard deviation and the mean shall be reported with the same number of decimal places.

2.2.4 The Coefficient of Variability *(„Relative SD")*

The coefficient of variability is a relative measure of variability (expressed in % most often). It is not dependent on measurement units like the standard deviation, therefore it can be used for comparison of the variability in data sets with different magnitude of their units (e.g. body weight in mice and cows).

Calculation formulas for sample coefficient of variability:

$$V = \frac{s}{\bar{x}} \qquad \text{or} \qquad V = \frac{s * 100}{\bar{x}} \ [\%]$$

The coefficient of variability expresses sample variability relative to the mean of the sample; because s and $\bar{x}$ have identical units, V has no units at all, a fact emphasizing that it is a relative measure, divorced from the actual magnitude or units of measurements of data.

The coefficient of variability of a sample, namely V, is an estimate of the coefficient of variability of the population from which the sample came (i.e., an estimate of σ/μ).

2.2.5 The Standard Error of the Mean (SEM, SE)

Notation: $\sigma_{\bar{x}}$ (Population), $s_{\bar{x}}$ (Sample)

The population standard error of the mean is the theoretical standard deviation of *all* sample means of size *n* that could be drawn from a population. The sample standard error of the mean can be used as a measure of the precision with which the sample mean $\bar{x}$ estimates the true population mean μ.

We don't know what the true mean value in the population is. We can only estimate it by means of the sample average. But we don't know how precise our calculation is and what's the difference between our calculated sample AVG and the true population mean μ. SEM may serve as a measure of precision of the calculated sample mean.

Value of the standard error of the mean depends on both the population variance (σ^2) and the sample size (*n*):

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Since in general we don't know the population variance, the best estimate for the population standard error of the mean in practice (**sample standard error of the mean**) is calculated as:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

The sample standard error of the mean is useful for construction of the **confidence interval** for the mean. The true mean value of population will lie within the interval sample average AVG ± SEM (see Chapter 4 for details).

Chapter 3
Distributions Commonly Used in Statistics

(Continuous Data)

In Chapter 1, we saw how frequency distributions arise from sample data and that the population distribution, arising from sampling the entire population, becomes the probability distribution. This probability distribution is used in the process of making statistical inferences about population characteristics on the basis of sample information. There are, of course, endless types of probability distributions possible. However, luckily, the great majority of statistical methods for continuous data use only several probability distributions.

The probability distributions most often used in statistics for continuous data are the normal, the non-normal, Student's *t*, Pearson's χ^2 (chi-square), and Fisher-Snedecor's *F* distribution. Rank–order methods depend on distributions of ranks rather than continuous data, but several of them use the normal or chi-square. Categorical data depend mostly on the chi-square, with larger samples transformed to normal. We need to become familiar with these commonly used distributions to understand most of the methods given in this text. The following paragraphs describe these distributions and some of their properties needed to use and interpret statistical methods.

Some of these distributions are useful with the population variables and some of distributions are useful with the sample variables.

3.1 Distributions for Population

3.1.1 Gaussian Normal Distribution

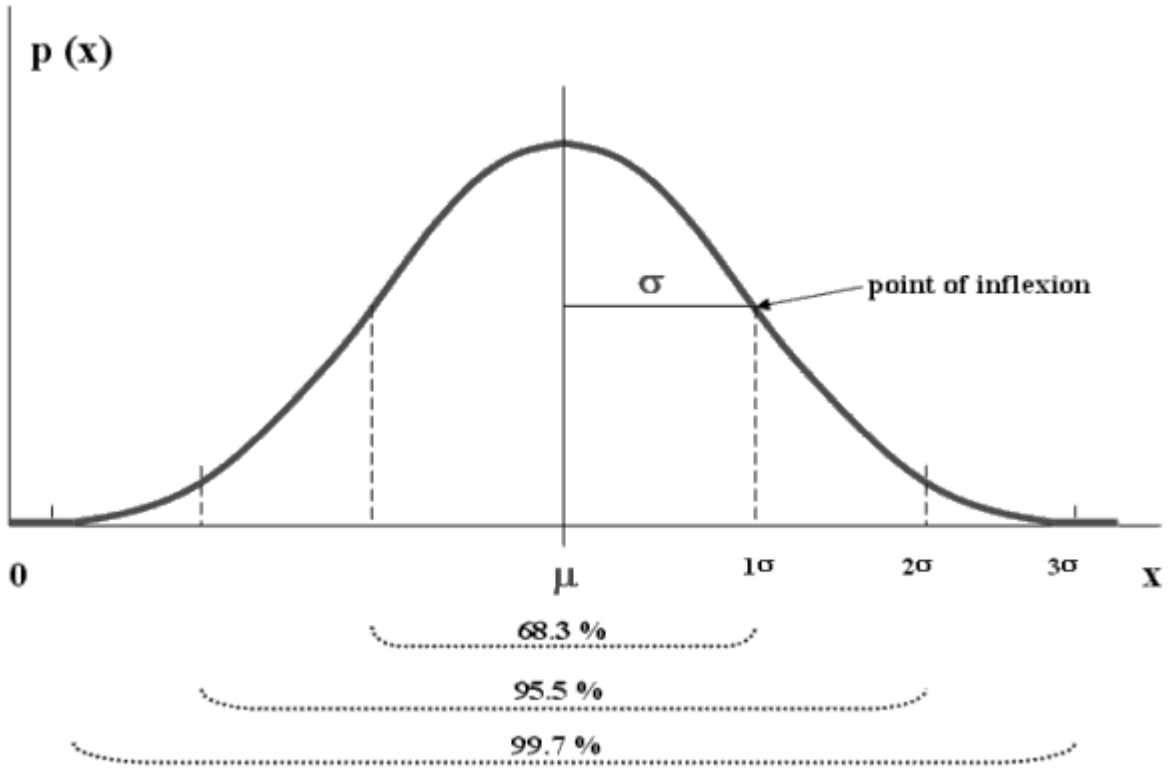
The normal distribution, called Gaussian by some users, is the perfect case of the famous symmetrical bell curve. In biology, a great many of data sets naturally follow Gaussian normal distribution, at least approximately.

The graphical description of the Gaussian normal distribution is in the Figure 3.1.

Axis x: values of monitored biological character,

Axis y: probability of occurrence of these values in population.

Fig. 3.1 Graph of Gaussian Normal Distribution



The curve is bell-shaped and symmetrical; majority of values is located around the mean (centre of symmetry) with progressively fewer observations toward the extreme values. In extremes, the curve is not terminated – the curve theoretically gets near the axis x in infinities (both + and - infinity).

The shape of the normal curve is fully described by means of **two parameters - μ and σ**:

μ (Mean value) – "Parameter of location" – it describes the centre of symmetry and also the location of the curve on axis x.

σ (Standard deviation) – "Parameter of dispersion (variability)". It describes the spread of the curve in the inflexion point (where the flexure changes from convex to concave). Spread of the curve determines the variability of biological character monitored in the population.

The whole area under the curve represents all individuals in the population (100%); then:

Within the range $\mu \pm 1\sigma$: there are 68.3 % of all values (individuals) in the population,
Within the range $\mu \pm 2\sigma$: there are 95.5 % of all values (individuals) in the population,
Within the range $\mu \pm 3\sigma$: there are 99.7 % of all values (individuals) in the population.

The occurrence of remaining values (0.3%) in both extreme ends of axis x is so highly improbable, that such extreme values are considered as an error of measurement in terms of statistics.

3.1.2 Standard Normal Distribution

The normal (Gaussian) variable (*X*) can be standardized, transforming it to *Z* by subtracting the mean and dividing by the standard deviation, e.g.,

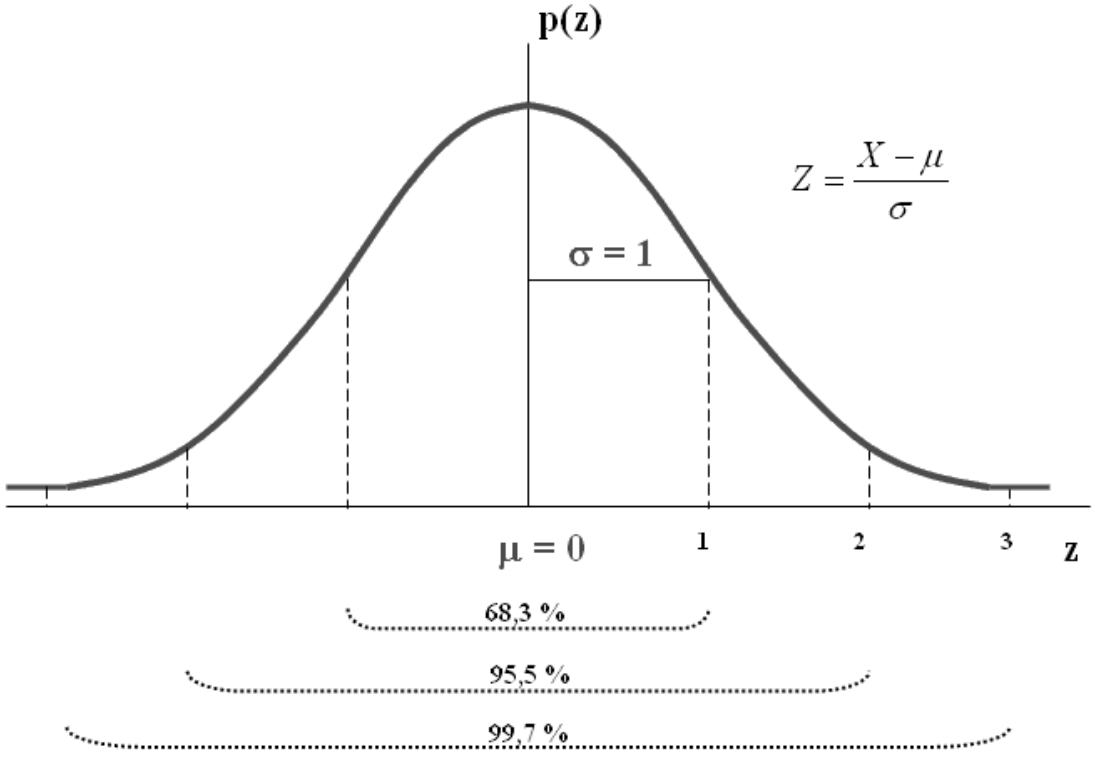
$$Z = \frac{X - \mu}{\sigma}$$

The normal distribution then becomes the *standard normal*, which has mean in the value 0 and standard deviation always equal to 1. Units of the new standardized variable *Z* on the axis x express the number of standard deviations away from the mean (zero value), positive for above the mean and negative for below the mean. It means that standardized variable Z represents a dimensionless quantity, it is a relative measure, divorced from the actual magnitude or units of measurements of data.

This transformation usually is made in practice as the probability tables available usually are of the standard normal curve. Table Appendix 1, in the back of the textbook, contains selected value of *z* with four areas that often are used: (a) the area under the curve in the positive tail for given z, i.e., one-tailed α; (b) the area under all except that tail, i.e., 1- α; (c) the areas combined for both positive and negative tails, i.e., two-tailed α; and (d) the area under all except the two tails, i.e., 1- α. (See Statistical tables Appendix 1)

The graphical description of the Standard normal distribution is in the Figure 3.2.
Axis x: values of standardized variable *Z*,
Axis y: probability of values *Z*.

Fig. 3.2 Graph of Standard Normal Distribution



3.1.3 Non-normal Distribution

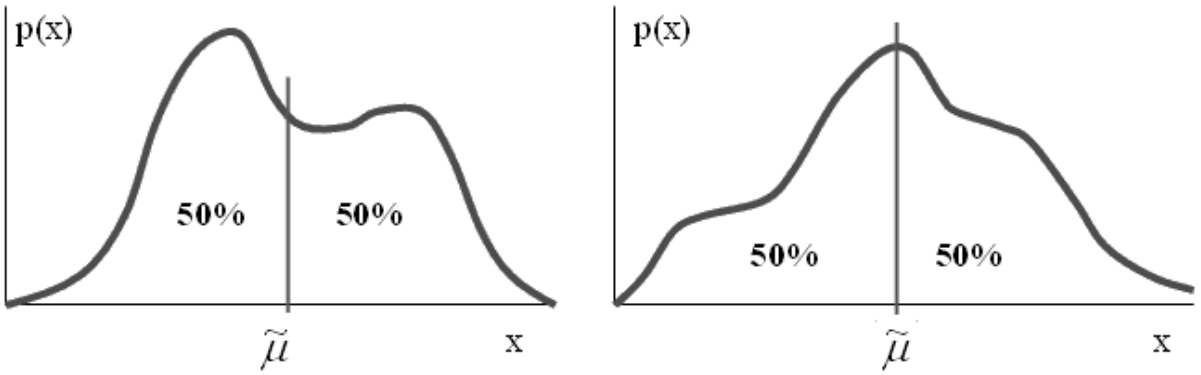
Some of variables monitored in biological and medical sciences don't follow Gaussian normal distribution – then they usually have variously irregular shape of the probability distribution curve, often asymmetrical or with 2 and more peaks. Such curves are most often called "non-normal" or "unknown"; because for their irregularity it is not possible to describe the shape of the curve in a very exact way.

The example of graphical description of the non-normal distribution is in the Figure 3.3.

Axis x: values of monitored biological character,

Axis y: probability of occurrence of these values in population.

Fig. 3.3 Graph of Non-normal Distribution



The curves of the non-normal probability distribution have the shapes that can be variously irregular; therefore it is impossible to use some exact parameters that would determine centre and spread of data (like it was possible in Gaussian normal distribution). Only one descriptive characteristics, the median, is usually used in such non-normal distributions. The median is considered as the centre of such irregular curve. Since the median is defined as 50% quantile, it divides the whole area under the curve into 2 equal halves regardless of the shape of the probability distribution. It is not possible to determine the spread of the curve (variability of monitored character) for its irregularity.

3.2 Distributions for Samples

3.2.1 *t*-distribution (Student's)

This distribution is defined for description of a theoretical variable *t* that is calculated from mean and standard deviation in a sample by means of various formulas used in different situations in statistics (first of all in statistical hypotheses testing). Common form of this formula for calculation of variable *t* is $t = \frac{X - \mu}{s}$. This variable is used in testing for differences between 2 means of statistical sets particularly.

(t-distribution was published in 1908 by English chemist W. S. Gosset under the pseudonym "Student", because the policy of his employer, Guinness Brewery, forbade the publication.)

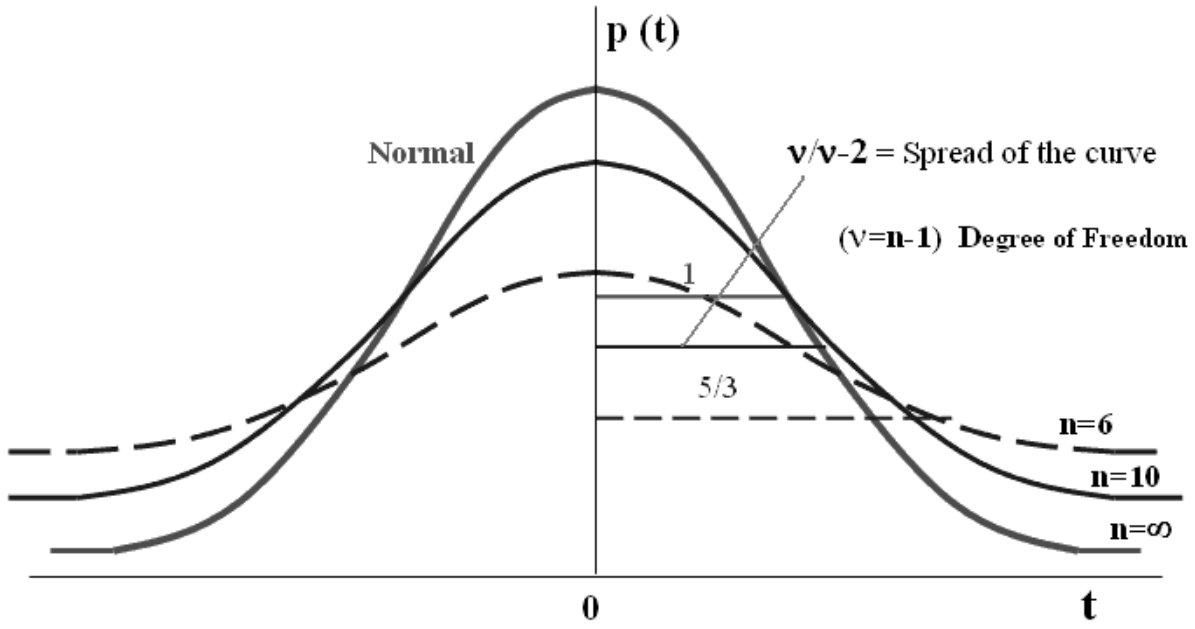
The *t* looks like the standard normal curve for variable $Z = \frac{X - \mu}{\sigma}$, however, it is a little fatter because it uses *s* instead of accurate σ. Whereas the normal is a single distribution, *t* is a family of curves. In the figure 3.4, two *t* distributions are superposed on a standard normal distribution. The particular member of the *t* family depends on the sample size or, more exactly, on the degrees of freedom (see below).

The graphical description of the *t*-distribution is in the Figure 3.4.

Axis x: values of variable *t*,

Axis y: probability of values *t*.

Fig. 3.4 Graph of *t*-Distribution



t-distribution reflects an error of samples (when compared to the population sampled) that is evident in all statistical calculations performed on the basis of such samples. This error of the sample is caused by small numbers of members in the sample, and generally we can say that the smaller is the sample, the more erroneous are calculations performed on the basis of this sample.

The shape of *t*-distribution curve is similar to the standard Normal distribution (bell-shaped, symmetry above 0 value), but the spread of the curve is specific for different samples according to

the sample size – or more exactly according to the **Degrees of Freedom** (DF, ν) of the sample monitored: **ν = n-1.**

It is obvious from the graph of *t*-distribution figured above that:

- The smaller is the sample size, the broader and lower is the curve,

- The larger is the sample size, the narrower and higher is the curve.

In case of the endless expansion of a sample in the extreme: $n = \infty$ the curve joins the Normal distribution that describes all population (such sample will have no error in statistical calculations in comparison with population).

In small samples (that have large error in comparison with the population) the shape of the curve is also much different from the shape of Normal distribution (used for population).

We can define the exact **spread of the curve** for *t*-distribution by ratio: **ν/ν-2.**

Values of *t*-distribution are tabulated in statistical tables (See the statistical tables - Appendix 2: Critical values for Student's *t*-distribution) and they can be used in statistical calculations as e.g.:

- **Critical values** in testing for difference between two means (see Chapter 6: Student *t*-test),
- **Coefficients** in calculations of confidence intervals for mean values (see Estimation of population parameters).

We search the critical values in the tables of *t*-distribution according to the degree of freedom calculated for our sample (ν = n- 1) and also according to an error α chosen to **specify the exactness** of our calculations in statistics - for **biological data** is commonly used **α = 5% or 1%** (when we need more precise calculations).

3.2.2 Chi-square (χ^2) Distribution (Pearson's)

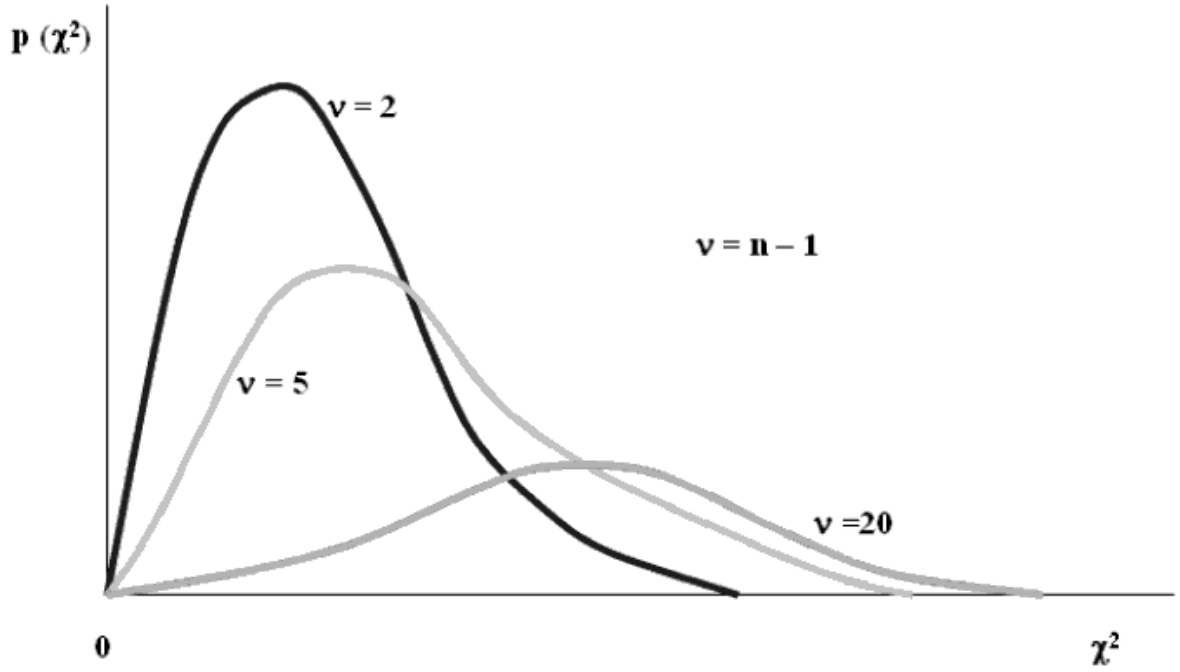
Chi-square distribution is defined for description of a theoretical variable χ^2 (in practice, we calculate it from data frequencies in samples) and we use it for calculations in testing for differences between frequencies in samples - e.g. when we need to compare sickness rate in different groups of animals (see Chapter 9 Categorical Data).

The graphical description of the chi-square distribution is in the Figure 3.5.

Axis x: values of variable χ^2,

Axis y: probability of values χ^2.

Fig. 3.5 Graph of Chi-square (χ^2) distribution.



Chi-square distribution has an asymmetrical curve – right skewed (it rises from 0 rapidly to a mode and then tails off slowly in a skew to the right) and it has different shapes for different sample sizes (ν = n-1 determines the shape).

It is obvious from the graph of χ^2 distribution figured above that:

- the smaller the sample size is the higher and more asymmetrical is the curve shape
- the larger the sample size is the lower and more symmetrical is the curve shape

Values of chi-square distribution are tabulated in statistical tables (See tables of the χ^2 distribution: Appendix 3, 4) and they can be used in statistical calculations e.g.:

- **Critical values** in testing for difference between frequencies in samples (See χ^2-test),
- **Coefficients** used in calculations of confidence intervals for standard deviation (see Estimation of population parameters).

3.2.3 *F*-distribution (Fisher-Snedecor's)

F-distribution is defined for description of a theoretical variable *F* and we use it in calculations in statistics e.g. in testing for differences between 2 variances in two groups of data. In practice, the theoretical variable *F* is calculated in the so called *F*-test, when we test the two sample variances s_1^2 and s_2^2 to determine whether, in the populations being sampled, one is greater in

probability. To do this, we use their ratio, dividing the bigger by the smaller. The probability distribution for this ratio is called *F*, named (by George Snedecor) after Sir Ronald Fisher, the greatest producer of practical theories in the field of statistics.

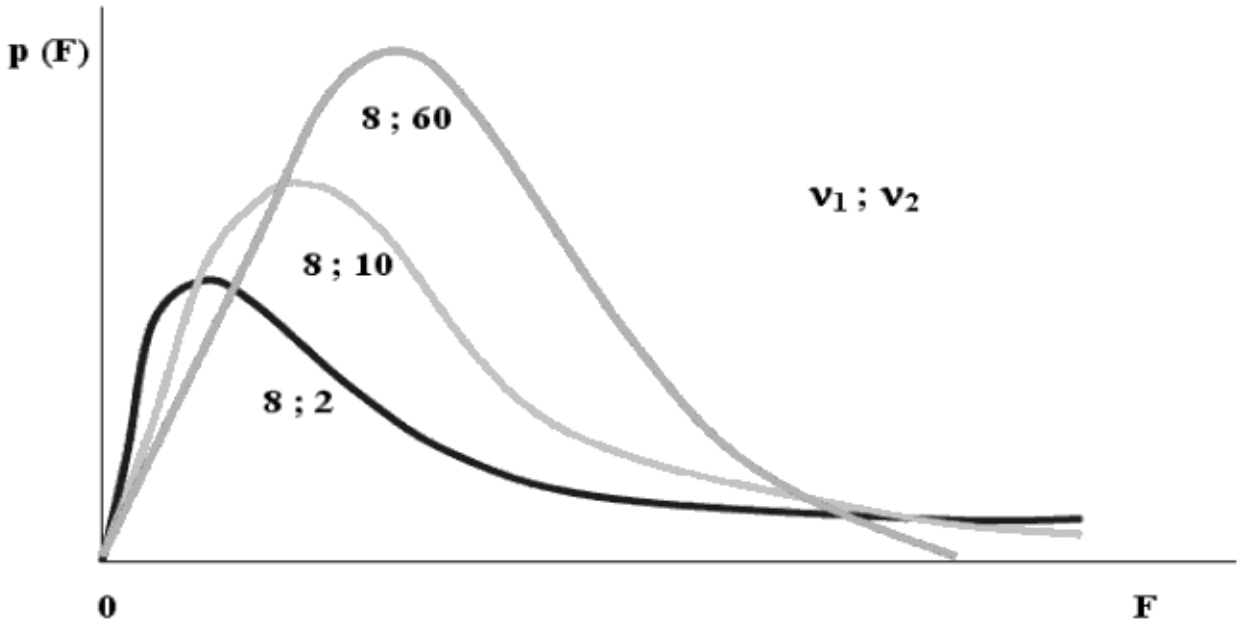
The *F*-distribution looks very much like the chi-square distribution (indeed it is the ratio of two independent chi-square-distributed variables), as can be seen in figure 3.6. *F* distribution has one more complication than chi-square: because it involves two samples, the degrees of freedom for each sample must be used. Then the shape of the *F*-distribution curve is determined by the degree of freedom of 2 samples (ν_1 and ν_2) that are used in testing (e.g. when we want to test differences between variability of monitored biological character in two groups of animals: see Chapter 6 *F* test for details).

The graphical description of the *F*-distribution is in the Figure 3.6.

Axis x: values of variable *F*,

Axis y: probability of values *F*.

Fig. 3.6 Graph of *F*-distribution.



As it follows from the figure, the curve is asymmetrical, it rises from 0 value, shortly runs up and then tails off slowly in a skew to the right toward higher values on the axis x. The shape of the curve is changing according to sample sizes (more exactly according to ν_1 and ν_2).

It is obvious from the graph of F-distribution figured above that:

- The smaller the sample size is the lower and more asymmetrical is the curve shape,
- The larger the sample size is the higher and more symmetrical is the curve shape.

Values of *F*-distribution are tabulated in statistical tables (See tables of the F-distribution: Appendix 5) and they are most often used as **critical values** in testing for differences between 2 sample variances (see F-test).

Chapter 4

Estimation of population parameters

(Confidence Intervals)

Having obtained a random sample from a population of interest, we are ready to use information from that sample to estimate the characteristics of the underlying population. If you are willing to assume that the sample was drawn from a normal distribution, summarize data with the sample mean and sample standard deviation, the best estimates of the population are mean and population standard deviation, because these two parameters completely define the normal distribution. When there is evidence that the population under study does not follow a normal distribution, summarize data with the median as the only descriptive characteristics used for the non-normal distribution.

Although sample statistics are the best estimates of true population parameters, they are still only estimates. Therefore, it is appropriate to determine confidence intervals for true population parameters that allow us to express the precision of the estimates based on the sample data. A confidence interval is an interval about an estimate, based on its probability distribution, that expresses the confidence, or probability, that that interval contains the true population parameter being estimated.

4.1 Normal Distribution – Estimation of μ and σ

Populations with Gaussian normal distribution are fully described by means of mean value and standard deviation (SD); true exact parameters can't be calculated in practice, so we use imprecise sample statistics as estimations of true parameters. The mean value **μ** of the population is estimated by means of the sample average $\bar{x}$ (AVG); and standard deviation **σ** is estimated by means of the sample ***s***.

Although sample average $\bar{x}$ is the best estimate of population μ, and sample standard deviation *s* is the best estimate of population **σ,** they are still only estimates. Therefore, it is useful to calculate also the confidence intervals for μ and σ that allow us to express the precision of the estimates.

4.1.1 Confidence Interval for the Mean Value μ

The calculation of confidence interval for the mean consists of determination of confidence limits L_1 (lower) and L_2 (upper). Limits will be symmetrical around the sample average $\bar{x}$, true mean value will lie within the interval restricted by the limits L_1, L_2.

For determination of the limits L1, L2 we need to know the **standard error of the mean** $s_{\bar{x}}$ **(SEM, SE)** = a measure of the precision with which a sample average $\bar{x}$ estimates the true population mean μ.

If we try to estimate the true mean value μ by means of several sample averages – we will see that every AVG is slightly different (caused by variability of individuals in samples) – but all of them will estimate the same true μ. The question is: what AVG is the best one? We need some measure to specify its precision = SEM. This statistic quantifies the certainty with which the mean computed from a random sample estimates the true mean of the population from which the sample was drawn.

Calculation formula for standard error of the mean:

$$SEM = s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

SEM is dependent on standard deviation (directly) and sample size (indirectly).

- If the sample size increases -> SEM decreases

  (Precision with which we estimate the true mean increases),
- If the sample is more variable (SD increases) -> SEM increases

  (Precision with which we estimate the true mean decreases).

SEM is used for calculation of confidence interval:

$$L_{1,2} = \bar{x} \mp s_{\bar{x}} . t_{\alpha,\nu}$$

$\bar{x}$ - ***sample mean***

$s_{\bar{x}}$ - ***standard error of the mean***

$t_{\alpha,\nu}$ - ***confidence coefficient*** *= critical value of t-distribution (Appendix 2. Critical values for Student's t-distribution) - determined according to the selected error α and DF: $\nu = n-1$.*

In the course of calculation we can determine a specific precision of the calculation by selecting the error α**.** For biological data this **error α = 0.05 or 0.01** is usually used. When referring to the selected α in the calculation of confidence interval, we call the quantity 1-α (namely, 1 - 0.05 = 0.95 or 1 – 0.01 = 0.99) the **confidence level** (95% or 99% confidence level).

Example:

Calculate confidence intervals for the mean of body weights of piglets at the 95% and 99% confidence level.

Body weights (in kg) of the sample of 25 piglets:
x_i : 25.8, 24.6, 26.1, 22.9, 25.1, 27.3, 24.0, 24.5, 23.9, 26.2, 24.3, 24.6, 23.3, 25.5, 28.1, 24.8, 23.5, 26.3, 25.4, 25.5, 23.9, 27.0, 24.8, 22.9, 25.4.

Method:

1) Calculation of the sample mean and SEM:

Mean: $\bar{x} = 25.0\ kg$

SEM: $s_{\bar{x}} = 0.27\ kg$

2) Calculation of the confidence intervals:

At the 95% confidence level: $\nu=24$ $t_{0.05,\ 24} = 2.064$

$$\bar{x} \mp s_{\bar{x}} . t_{(0.05,\ 24)} = 25.0 \mp 0.27\ .\ 2.064 = 25.0 \mp 0.56\ kg$$

At the 99% confidence level: $\nu=24$ $t_{0.01,\ 24} = 2.797$

$$\bar{x} \mp s_{\bar{x}} . t_{(0.01,\ 24)} = 25.0 \mp 0.27\ .\ 2.797 = 25.0 \mp 0.76\ kg$$

3) Conclusion:

The true mean value for population of body weights in piglets lies within the confidence intervals: $25.0 \mp 0.56\ kg$ (at the 95% confidence level),
$25.0 \mp 0.76\ kg$ (at the 95% confidence level).

4.1.2 Confidence Interval for the SD (σ)

The calculation of confidence interval for the population standard deviation consists of determination of confidence limits L_1 (lower) and L_2 (upper).

Confidence limits L_1, L_2 for determination of confidence interval above estimate of standard deviation (s) must be calculated separately, since the interval is not symmetrical; that is, the distance from L_1 to s is not the same as the distance from s to L_2.

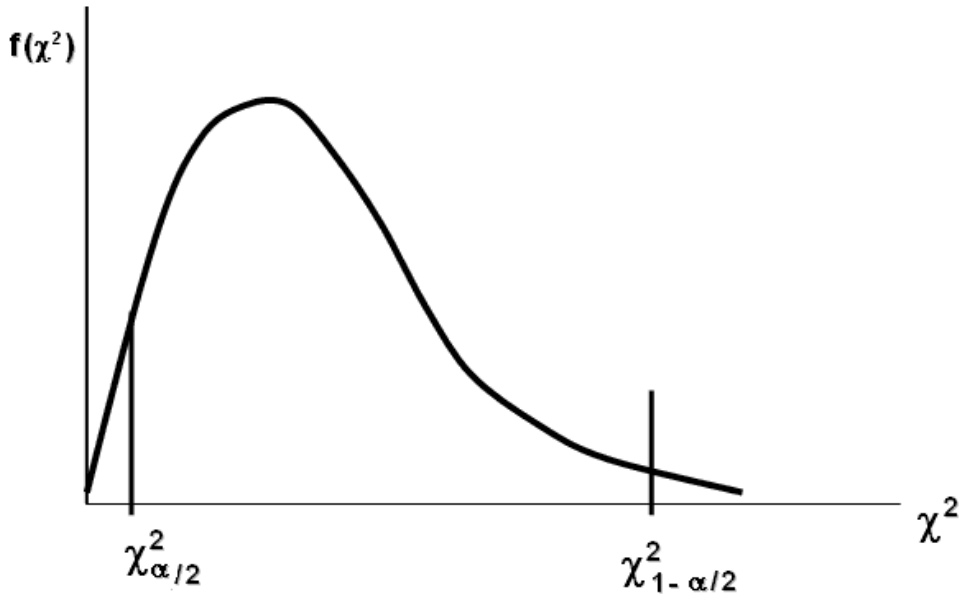
$$L_1 = \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2(\nu)}}} \qquad L_2 = \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2(\nu)}}}$$

s^2 – sample variance
n – sample size
Confidence coefficients *(=critical values of χ^2distribution):*
Quantile $\chi^2_{\alpha/2}$ *– left tail value $\Rightarrow$ upper limit of the interval*
Quantile $\chi^2_{1-\alpha/2}$ *– right tail value $\Rightarrow$ lower limit of the interval*

(See Appendix 3 and 4: Critical values for χ^2 distribution, Right tail, Left tail)

Appropriate quantiles of χ^2 distribution used in the calculation of limits L_1 and L_2 for confidence interval of standard deviation are presented in the Figure 4.1.

Fig. 4.1 Left tail value and right tail value of χ^2-distribution (quantiles $\chi^2_{\alpha/2}$, $\chi^2_{1-\alpha/2}$)



Example:

Calculate confidence intervals for the standard deviation of body temperatures of twenty five intertidal crabs placed in air at 24.3°C at the 95% confidence level.

Body temperatures (measured in °C): 25.8, 24,6, 26.1, 22.9, 25.1, 27.3, 24.0, 24.5, 23.9, 26.2, 24.3, 24.6, 23.3, 25.5, 28.1, 24.8, 23.5, 26.3, 25.4, 25.5, 23.9, 27.0, 24.8, 22.9, 25.4.

Method:

1) Calculation of sample variance:

   $s^2 = 1.80\ (°C)^2$

2) Calculation of the confidence intervals:

   At the 95% confidence level: $\nu=24$ $\chi^2_{0.975,\,24} = 12.40$, $\chi^2_{0.025,\,24} = 39.36$

   (See Appendix 3 and 4: Critical values for χ^2 distribution, Right tail, Left tail)

$$L_1 = \sqrt{\frac{(25-1).1.80}{39.36}} = \sqrt{\frac{43.20}{39.36}} = \sqrt{1.10} = 1.05\,°C$$

$$L_2 = \sqrt{\frac{(25-1).1.80}{12.40}} = \sqrt{\frac{43.20}{12.40}} = \sqrt{3.48} = 1.87°C$$

3) Conclusion:

The true standard deviation for population of body temperatures of intertidal crabs placed in air at 24.3°C lies within the confidence interval that is restricted by limits: L_1 = 1.05°C and L_2 = 1.87°C (calculated at the 95% confidence level).

(Note that the confidence limits are not symmetrical around s; that is, the distance from L_1 to s is not the same as the distance from s to L_2).

4.2 Non-normal Distribution – Estimation of the Median

When the population under study does not follow the Gaussian normal distribution, then the only one descriptive characteristic, the median, can be used for definition of such non-normal distribution. For irregularity of the distribution curve, it is not possible to determine the spread of the distribution, i.e. the variability of monitored data set.

The true exact median of the population under study can't be calculated in practice, so we use imprecise sample median as estimation of the true parameter. The population median $\tilde{\mu}$ is estimated by means of the sample median $\tilde{x}$.

Although sample median $\tilde{x}$ is the best estimate of population $\tilde{\mu}$, it is still only estimate. Therefore, it is useful to calculate also the confidence intervals for $\tilde{\mu}$ that allows us to express the precision of the estimate.

4.2.1 Confidence Interval for the Median

The calculation of confidence interval for the population median $\tilde{\mu}$ consists of determination of confidence limits L_1 (lower) and L_2 (upper).

Confidence limits L_1, L_2 are values derived from the statistical tables (see Table 1 below).

According to the sample size **n** and selected α we find in statistical tables **ranks for L_1, L_2**.

Then we replace these ranks with the actual values from the variant sequence (arranged order of measured data, ascending or descending).

***Example**:*

Calculate the confidence interval for the true population median of body weights (kg) in the sample of 14 dogs of a particular breed:

Measured values (in kg): 14.1, 16.4, 16.8, 14.3, 12.3, 14.9, 15.3, 12.8, 15.6, 13.5, 16.0, 16.2, 17.1, 17.0

Method:

1) According to the sample size and selected α we find in statistical tables ranks for L_1, L_2:

   n = 14
   α = 0.05

Tab. 4.1 Ranks for Confidence Limits for the Median (Part of the table, α = 0.05)

n	Lower Limit	Upper Limit
8	1	8
9	2	8
10	2	9
11	2	10
12	3	10
13	3	11
14	3	12
⋮	⋮	⋮
100	40	61

2) We arrange the measured values in an ascending row (variant sequence).
3) We replace found ranks (3 and 12) with the actual values from the variant sequence of measured values to determine confidence limits L_1 and L_2 (See Table 2):

Tab. 4.2 Sample data (n = 14) with marked confidence limits for the median

x_1	12.3
x_2	12.8
x_3	**13.5**
x_4	14.1
x_5	14.3
x_6	14.9
x_7	15.3
x_8	15.6
x_9	16.0
x_{10}	16.2
x_{11}	16.4
x_{12}	**16.8**
x_{13}	17.0
x_{14}	17.1

3) Conclusion:

Confidence interval for the true median of body weights in dogs of the particular breed will lie within limits: L_1 = 13.5 and L_2 = 16.8 (calculated at the 95% confidence level).

Chapter 5

Statistical Hypotheses Testing

5.1 Statistical Hypothesis

Testing of statistical hypotheses is one of the most important parts of statistics in regard to the practical use of biostatistics (first of all it can be used for evaluation of experimental data) and it helps us to make conclusions from experiments performed with animals or some individuals in general. The major goal of statistical analyses is to draw conclusions regarding the whole population by examining a sample (or more samples) from that population – on the basis of this sample data we can decide on acceptance of some hypothesis regarding the population that we are interested in. A very common example of this is the desire to draw conclusions about one or more population means, as the most important statistical characteristics, or conclusions regarding variability of two populations, sometimes also conclusions regarding distribution of variables monitored in the population.

We begin by making a concise statement about the population – a specific hypothesis.

Hypothesis can be any statement about a population characteristic: its distribution or parameters (mean, SD).

For example: A population matches Gaussian normal distribution

2 populations have the same mean

2 populations have the same variance

This hypothesis is called a ***null hypothesis*** (H_0) – it expresses the concept of "no difference".
For example:

H_0: μ = const. (e.g. the population mean is equal to certain value known about the studied population – e.g. physiological values of some biochemical indices)

$\mu_1 = \mu_2$ (2 populations have the same mean value)

$\sigma_1^2 = \sigma_2^2$ (2 populations have the same variance)

If it is concluded (through a statistical test – see below) that it is likely that a null hypothesis is false, then an ***alternate hypothesis*** (abbreviated H_A) is assumed to be true. H_A denies H_0, so for examples above:

H_A: $\mu \neq$ const.

$\mu_1 \neq \mu_2$

$\sigma_1^2 \neq \sigma_2^2$

One states a null hypothesis and an alternate hypothesis for each statistical test performed, and all possible outcomes are accounted for this pair of hypotheses.

It must be emphasized that statistical hypotheses are to be stated *before* data are collected to test them. A statement of hypotheses after examination of data can devalue a statistical test. One may, however, legitimately formulate hypotheses *after* inspecting data if a new set of data is then collected with which to test the hypotheses.

The use in practice: ***The experimental data evaluation.***

E.g.: we need to find out if a vitamin supplement in food causes the increase of body weight in piglets.

We set up an ***Experiment****:*

*Group1 of piglets (**Test** sample) gets the vitamin supplement in food*

*Group2 of piglets (**Control** sample) gets standard food*

After some period we measure the body weight in both groups of animals and we can find out e.g. that test sample has a mean $\bar{x}_1$ *which is higher than the mean of the control group:* $\bar{x}_2$*. We have to decide (through a statistical test), whether the difference between the sample means is only random (caused by variability of animals) – or whether it is big enough to conclude that population means are different as well. It would mean that the difference was caused by our experimental activity (we can say that this experimental activity is generally effective).*

In this case we can ***reject the null hypothesis*** *(*$\mu_1 = \mu_2$*) and it means that the alternate hypothesis is true: (*$\mu_1 \neq \mu_2$*).*

Conclusion *in practice (for this particular experiment): the statement that "**the increase of body weight is caused by the vitamin supplement"** is generally true (the increase of the body weight is not a random effect).*

Probability and significance

To draw conclusions from experimental data we need first to set arbitrary critical thresholds of probability (P-values). The occurrence of an event whose estimated probability is less than a critical threshold is regarded as a ***statistically significant*** outcome. The usual thresholds of probabilities (P-values) chosen for biological and medical data are P = 0.05, i.e. significant; P = 0.01, i.e. highly significant. The procedure for deciding if an outcome is significant is called a ***statistical test***.

5.2 Statistical Tests

The statistical test is used as a decision rule about the acceptance (or rejecting) of the null hypothesis verified in an experiment. The objective of a statistical test is to obtain (from experimental data) a single number called a ***test statistic*** (calculated variable, whose probability distribution is known).

Different variables are used as the test statistics in various statistical tests e.g.:

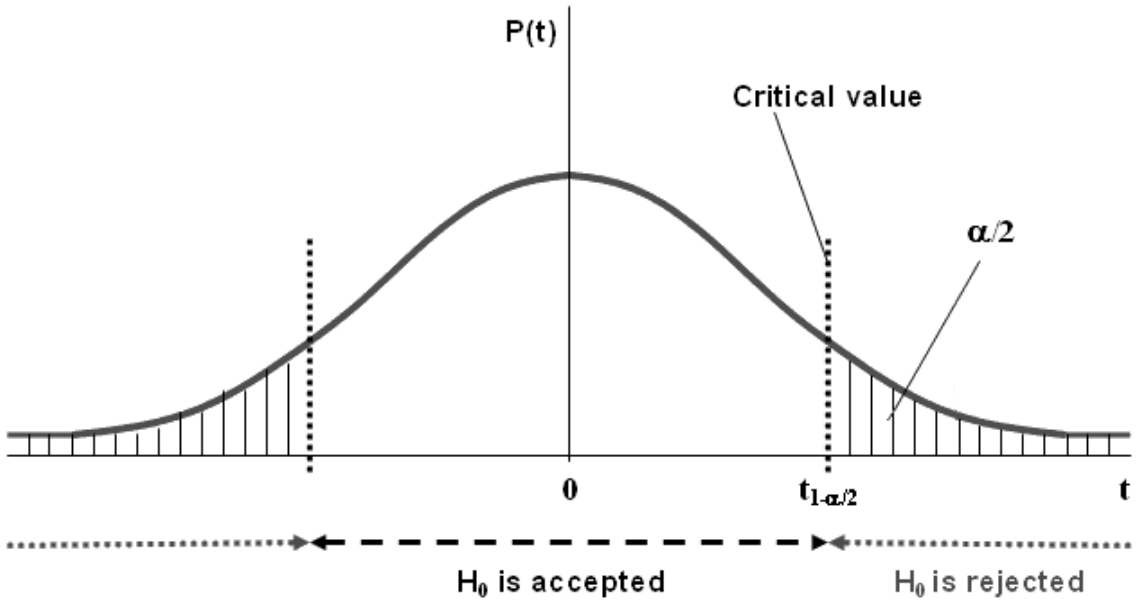
t – Testing for difference between 2 means (t-test),

F – Testing for difference between 2 variances (F-test),

χ^2 – Testing for difference between 2 frequencies (χ^2-test).

The procedure of each of the statistical tests consists in a test statistic calculation – then we determine if the calculated value of the test statistic exceeds some ***critical value*** of the test statistic. When the calculated test statistic (in absolute value) exceeds the critical value, then the null hypothesis is rejected. Otherwise, the null hypothesis is accepted. Fig. 5.1 shows an example of critical value for the test statistic *t* (*t* distribution).

Fig. 5.1 Critical value for the test statistic *t*

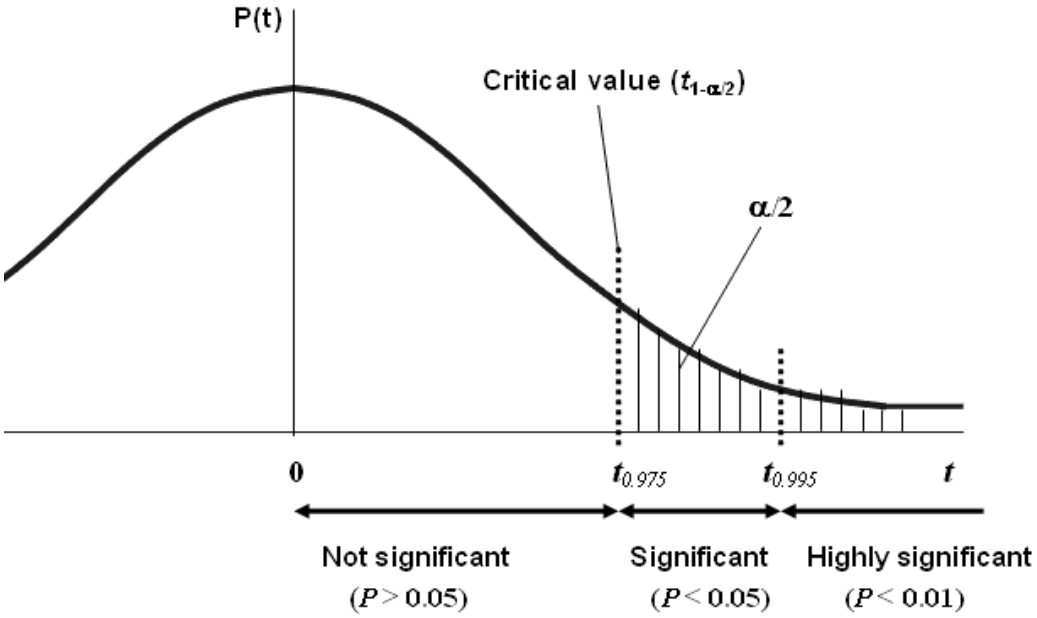


t – Test statistic *t* (for details, see Chapter 6: *t*-test)
P(t) – Probability of *t*-values

This critical value is associated with a particular probability threshold (*P*-value) that is used as a criterion for rejection of H_0 in the test and that is called the ***significance level***, denoted by **α**. In fact, the critical value is usually the 1-α/2 quantile of the appropriate distribution used as the test statistic. As explained above, a probability of 5% or less is commonly used as the criterion for rejection of H_0 in biological and medical data testing. It means that the calculated test statistic (in absolute value) has to exceed the critical value at the $\alpha = 0.05$ level of significance to obtain a statistically significant outcome of the test (denoted as $P < 0.05$ usually). When the calculated test statistic exceeds the critical value at the $\alpha = 0.01$ level of significance, we obtain a statistically highly significant outcome of the test (denoted as $P < 0.01$ usually). In the case, when the calculated

test statistic does not exceed the critical value at the α = 0.05 level of significance, we obtain a statistically not significant outcome of the test (denoted as $P > 0.05$ usually). Following Figure 5.2 demonstrates the aforesaid possible outcomes of the test statistic *t*.

Fig. 5.2 Test statistic *t* – possible outcomes of the test



t – Test statistic *t* (for details, see Chapter 6: t-test)
P(t) – Probability of t-values

Types of Errors in Hypotheses Testing

It is very important to realize that a true null hypothesis occasionally will be rejected, which of course means that we have committed an error in drawing a conclusion about the sampled population. Moreover, this error will be committed with a probability of α. That is, if H_0 is in fact a true statement about a statistical population, it will be concluded (erroneously) to be false 5% of the time. The rejection of a null hypothesis when it is in fact true is what is known as a *Type I error* ("Type 1 error", also called an alpha error or an "error of the first kind"). On the other hand, if H_0 is in fact false, a statistical test will sometimes not detect this fact, and we shall thus reach an erroneous conclusion by not rejecting H_0. The probability of committing this error, of not rejecting the null hypothesis when it is in fact false, is represented by β. This error is referred to as a *Type II error* ("Type 2 error", also called a beta error, or an "error of the second kind"). The *power* of a statistical test is defined as 1- β; i.e., power is the probability of rejecting the null hypothesis when it is in fact false and should be rejected.

Whereas the probability of committing a Type I error is α, the specified significance level, the probability of committing a Type II error is β, a value that generally we neither specify nor

know. What we do know is that for a given sample size, *n*, the value of α is related inversely to the value of β. That is, lower probabilities of committing a Type I error are associated with higher probabilities of committing a Type II error. Both types of error may be reduced simultaneously by increasing *n*. Thus, for a given α, larger samples will result in statistical testing with greater power (1 - β). Table 5.1 summarizes these two types of statistical errors.

Tab. 5.1 The Two Types of errors in Hypotheses Testing

DECISION / REALITY	**REJECTING H_0**	**NOT REJECTING H_0**
H_0 IS TRUE	Type I error **α**	**NO ERROR** **1- α**
H_0 IS FALSE	**NO ERROR** **1- β** (power of test)	Type II error **β**

Since, for a given *n*, one cannot minimize both of types of errors, it is appropriate to ask what the acceptance combination of the two might be. In terms of veterinary medicine: "*Not to treat an ill animal (statistically evaluated as a healthy one – Type I error) is a more serious mistake than to treat a healthy animal statistically evaluated as the ill one (Type II error)*". Therefore, statistical tests used in medicine are set up to achieve a minimal Type I. error α. By experience, and hence by convention, an α of 0.05 is usually considered to be a "small enough" chance of committing a Type I error, while not being so small as to result in "too large a chance" of a Type II error. But there is nothing sacred about the 0.05 level. Although it is the most widely used significance level, researchers may decide for themselves whether it is more important to minimize one type of error or the other. In some situations, for example, a 5% chance of an incorrect rejection of H_0 may be felt to be unacceptably high, so the 1% level of significance is sometimes employed.

It is necessary, of course, to state the significance level used when reporting the results of a statistical test. Indeed, rather than simply stating whether the null hypothesis is rejected, it is good practice to state also the test statistic itself and the best estimate of its exact probability (calculated by means of a statistical software). In this way, readers of the research results may draw their own conclusions, even if their choice of significance level is different from author's.

Bear in mind, however, that the choice of α is to be made before even seeing the data. Otherwise there is a great risk of having the choice influenced by examination of the data, introducing bias instead of objectivity into proceedings. The best practice generally is to decide on the null hypothesis, alternate hypothesis, and significance level before commencing with data collection.

As we already know, it is conventional to refer to rejection of H_0 at the 5% significance level as denoting a "significant" difference (e.g. between compared population means) and rejection at the 1% level as indicating a "highly significant difference". As the significance level selected is

somewhat arbitrary, if test results are very near that level (e.g. between 0.04 and 0.06 if $\alpha = 0.05$ is used), then it may be wiser to repeat the analysis with additional data than to declare emphatically that the null hypothesis is or is not a reasonable statement about the sampled population.

5.3 Classifications of Statistical Tests for Different Types of Data

It is important to choose the appropriate statistical test for a specific type of data. This is not always straightforward and sometimes, more than one test can be used indeed. We have noted in the chapter 1 that there are three types of data: categorical, rank-order, and numerical. The categorical and rank-order data are discrete by their nature; numerical data may be continuous or discrete. Each type of data requires its own **form of statistical testing**.

A) Categorical Form of Testing

To compare two variables using categorical data, we compare counts (frequencies) in two samples – e.g. number of ill animals in a stable, number of vaccinated dogs, number of dead born piglets etc. We form two-way tables of counts with one variable representing rows and the other variable representing columns. We test the proposition that knowledge of the counts in one variable's categories tells us something about the counts in the other variable's categories, i.e. that the two variables are not independent. Analyses of such dependences in categorical data, as well as analyses of differences between counts in categorical data are performed by means of ***Chi-square tests*** (See Chapter 9: Categorical data, Contingency tables).

B) Rank-Order Form of Testing

To compare two groups that are in rank order, we attach ranks to the data combined over the two groups and then add the rank values for each group separately, forming rank sums. If the group rankings are not much different, the ranking from the two groups will be interleaved and the rank sums will not be much different. If one group has most of its members preceding the other in rank, one rank sum will be larger and the other small. Probabilities of rank sums have been tabulated, so that the associated *P*-value can be looked up in the table and the decision about the group difference made.

Rank-order form of statistical testing is represented by ***Non-parametric tests*** that are used for testing of hypotheses with rank-order data, discrete numerical data and numerical continuous data those are not assumed to come from a normally distributed population. (See Chapter 7: Non-parametric tests).

***Non-parametric Tests – a summary*:**

- They are used for data sets following the *non-normal distribution* especially,
- *Hypotheses concerning common characteristics* of statistical sets are tested in the non-parametric tests (e.g. two sets have the same shape of distribution),
- Calculations in these tests are based on *ranks* of measured values.

.

C) Continuous Form of Testing

Whether a difference between means exists most often is the focus in comparing two groups with data in continuous form. Our first inclination is to look at the difference between means. However, this difference depends on the scale. The offset distance of a broken femur appears larger if measured in centimetres than in inches. The distance must be standardized into units of data variability. We divide the distance between means by a measure of variability and achieve a statistic *t* (if the population variability is estimated by small samples: see *t*-test). The risk of concluding a difference when there is none (the *P*-value) is looked up in a table and the decision about the group difference is made.

Continuous form of statistical testing is represented by ***Parametric tests*** that may be used for testing of hypotheses with numerical data those are assumed to come from a normally distributed population (See Chapter 6: Parametric tests).

Parametric Tests – a summary:

- They are used in testing for differences between data sets that follow *Gaussian normal distribution*,
- *Hypotheses concerning parameters* (μ, σ) of this distribution are tested,
- Calculations in these tests are based on *sample statistics* ($\bar{x}$, s).

Chapter 6

Parametric Tests

Among the most commonly employed biostatistical procedures is the comparison of two samples to infer whether differences exist between the two populations sampled. In parametric tests, we consider hypotheses concerning population parameters μ (mean value) and σ^2 (variance) of Gaussian normal distribution.

As the mean is the most important characteristic of a population, the basic question asked in parametric test most often is whether two samples have the same mean or whether a sample mean is the same as a population mean. Questions concerning two variances (or standard deviations) are also considered in parametric tests in some instances. The question is answered by testing the null hypothesis that the means (or variances) are equal and then accepting or rejecting this hypothesis.

Tests of means and variances were developed under the assumption that the sample was drawn from a normal distribution. Whereas usually not truly normal, a distribution that is roughly normal in shape is adequate for a valid test. That is because the test is moderately *robust*. Robustness is an important concept. A robust test is one that is affected little by deviations from underlying assumptions. If a small-to-moderate sample is too far from normal in shape, the calculation of error probabilities, based on the assumed distribution, will lead to erroneous decisions; then non-parametric (rank-order) methods should be used preferably (see chapter Non-parametric tests). In particular, tests of means are only moderately robust and they are especially sensitive to outliers (extremely high or low values), whereas tests of variance are much more robust.

Student's *t*-test (used in testing for difference between two means) and Snedecor's *F*-test (used in testing for difference between two variances) belong to the group of parametric tests.

6.1 *F*-test (Variance ratio Test)

We can decide by means of this *F*-test whether some treatment (activity used in an experiment) influences the **variability** (variance - σ^2) of some biological character studied in a population. A null hypothesis $\mathbf{H_0: \sigma_1^2 = \sigma_2^2}$ is verified by examining sample variances – s_1^2 and s_2^2.

We select 2 samples from the population monitored:

Sample 1 (n_1 individuals)

Sample 2 (n_2 individuals)

We apply a tested treatment (e.g. a new medical preparation) to one of these samples, the second sample (without any treatment) will serve as a control group for comparison.

Method:

1) We calculate sample **variances** s_1^2 and s_2^2:

$$s_1^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n_1}}{n_1 - 1} \qquad s_2^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n_2}}{n_2 - 1}$$

2) We calculate a **test statistic**: $F = \frac{higher\ variance\ (out\ of\ s_1^2, s_2^2)}{lower\ variance\ (out\ of\ s_1^2, s_2^2)}$

3) We specify **degree of freedom** of Numerator (of the *F* ratio) and Denominator (of the *F* ratio), appropriate to the size of sample1 and sample2:

$DF_{Numerator}$: $\nu_N = n_{1(2)} - 1$ (for the higher out of s_1^2, s_2^2)

$DF_{Denominator}$: $\nu_D = n_{1(2)} - 1$ (for the lower out of s_1^2, s_2^2)

4) We find out **critical value** $\boldsymbol{F_{(\alpha, \nu N, \nu D)}}$ in statistical tables for *F*-distribution (Appendix 5) according to the chosen error α (0.05), and degree of freedom (*DF*) for numerator and denominator of our ratio for the test statistic *F*.

5) We will compare the calculated test statistic *F* with table critical value $F_{crit.}$ to make a **conclusion** about population variances and about an effect of the treatment used in the experiment on variability of studied biological character:

If calculated $F > F_{(crit.)} \Rightarrow$ We **reject H_0**, then alternate hypothesis is true: H_A: $\sigma_1^2 \neq \sigma_2^2$.

i.e. There is a **significant difference** between variances – it means that variability of 2 populations sampled is <u>*not equal*</u> (at the α level).

Conclusion: the treatment used in experiment has influenced the variability of the studied biological character.

If calculated $F \leq F_{(crit.)} \Rightarrow$ **H_0 is true:** $\sigma_1^2 = \sigma_2^2$.

i.e. There is an **insignificant difference** between variances – it means that variability of 2 populations sampled <u>*is equal*</u> (at the α level).

Conclusion: the treatment used in experiment has not influenced the variability of the studied biological character.

Example:

The effect of a new veterinary preparation on AST (aspartate aminotransferase) level in blood serum in dairy cows has been monitored. In 10 dairy cows (control group), to which the preparation was not applied, the following AST activities in blood serum have been found (in $\mu mol.l^{-1}$):

0.409, 0.345, 0.392, 0.377, 0.398, 0.381, 0.400, 0.405, 0.302, 0.337

In 10 dairy cows (test group), to which the preparation was applied. the following AST activities in blood serum have been found ($\mu moll^{-1}$):

0.341, 0.302, 0.504, 0.452, 0.309, 0.375, 0.479, 0.423, 0.311, 0.333

Does the preparation influence the variance of AST activity in blood serum of dairy cows?

Method:

1) We calculate sample variances s_1^2 and s_2^2:

Control: $s_1^2 = 0.00125$

Preparation: $s_2^2 = 0.00575$

2) We calculate a test statistic F:

$$F = \frac{s_2^{\,2}}{s_1^{\,2}} = \frac{0.00575}{0.00125} = 4.618$$

3) We specify **degree of freedom** of the numerator and denominator of the ratio:

$\nu_N = n_2 - 1 = 9$ (for $s_2^{\,2}$)
$\nu_D = n_1 - 1 = 9$ (for $s_1^{\,2}$)

4) Critical value $F_{crit\,(0.05,\,9,\,9)} = 4.026$

5) $F > F_{(crit.)}$. $\Rightarrow$ statistically significant difference was found between variances (An alternate hypothesis H_A: $\sigma_1^2 \neq \sigma_2^2$ is true) at the level $\alpha = 5\%$.

Conclusion: As difference between variances of control and test groups is statistically significant ($P < 0.05$), the preparation tested influences the variance of the activity of AST in blood serum in dairy cows.

6.2 *t*-test (Student's)

Student's *t*-test is used for testing for difference between **2 population means μ** (in general). However, there are several different variants of *t*-test in practice according to the data sets that are available for comparison (see below). The *t*-test is the most common statistical procedure in the medical and biological literature; you can expect it to appear in more than half the papers you read in the general medical literature. In addition to being used to compare two group of means, it is widely applied incorrectly to compare multiple groups, by doing all the pairwise comparisons, for example, by comparing more than one intervention (treatment) with a control condition.

Student's *t*-test is especially useful for testing for significant differences between results obtained under two experimental conditions (treatments). First, we hypothesise that the means of our two populations are not different (**null hypothesis,** e.g. $\mathbf{H_0: \mu_1 = \mu_2}$). We then determine the probability (our *P*-value) that the difference in our samples´ means could have arisen by chance. This *P*-value is thus a measure of the compatibility between our experimental observations and our null hypothesis. A low *P*-value), say <0.05, is typically regarded as statistical evidence to reject the null hypothesis and conclude that there is significant difference in the result obtained from the two experimental conditions (treatments).

The null hypothesis states that the mean of the population from which the sample is drawn is not different from a theoretical mean or from the population mean of another sample drawn from the same population. We also must choose the alternate hypothesis, which will select a **two-tailed** or **one-tailed test**. We should decide this before seeing the data so that our choice will not be influenced by the outcome. The two-tailed *t*-test is used to test against the alternative hypothesis that $\mu_1 \neq \mu_2$. It is sometimes the case that *before* the data are collected there is only one reasonable way in which the 2 means could differ, and the alternative hypothesis would be for example $\mu_1 > \mu_2$ (or $\mu_1 < \mu_2$). It is the appropriate to carry out a one-tailed *t*-test. We often *expect* the result to lie toward one tail, but expectation is not enough. If we are sure the other tail is impossible, such as for physical or psychological reasons, we unquestionably use a one-tailed test. Surgery to sever adhesions and return motion to joint frozen by long casting will allow only a positive increase in angle of motion; a negative angle physically is not possible. An one-tailed test is appropriate.

There are cases in which an outcome in either tail is possible, but a one-tailed test is appropriate. When making a decision about a medical treatment, i.e., whether we will alter treatment depending on the outcome of the test, the possibility requirement applies to the alteration in treatment, not the physical outcome. If we will alter treatment only for significance in the positive tail and it will in no way be altered for significance in the negative tail, a one-tailed test is appropriate. However, very often we also need to test an alternate hypothesis that there is any difference (whichever difference: toward positive or negative tail) between treatments used in experiment – in these cases the two-tailed test is appropriate.

The difference between one-tailed and two-tailed tests is also reflected in the size of the critical value; generally we can say that critical values used in one-tailed tests are lower than critical values used in two-tailed tests (at the same α level of significance). In the table of critical values of *t*-distribution (Appendix 2), we can see that one-tailed critical values (marked as α(1)) at the specific α level are the same as two-tailed critical values (marked as α(2)) at the double α level of significance for given ν.

The procedure of Student's *t*-test consists in the calculation of a **test statistic *t*** that results from the estimation of parameters μ and σ in samples: $\bar{x}$ and ***s***. Calculated test statistic is compared with the tabulated critical value $t_{\alpha,\nu}$ that we can find out in tables of *t*-distribution (Appendix 2) according to the chosen error α (our probability level for acceptance of significant difference it is typically set at 0.05 by most researchers in the biological and medical sciences) and ν (DF - degree of freedom calculated by means of a specific formula for each of the variants of *t*-test).

If calculated $t > t_{\alpha,\nu}$ => we reject the null hypothesis, it means that there is a significant difference between means of populations sampled at the α level of significance. We accept the alternate hypothesis that our two experimental groups (typically one sample with the treatment and the second one – control without treatment) produced statistically significant results (i.e. samples was not drawn from the same population).

If calculated $t \leq t_{\alpha,\nu}$ => we accept the null hypothesis, it means that there is an insignificant difference between means of populations sampled at the α level of significance, i.e. our two experimental groups (typically one sample with the treatment and the second one – control without treatment) produced insignificant results (i.e. samples was drawn from the same population).

6.2.1 Population vs. Sample Comparison (One-sample *t*-test)

This variant of *t*-test is used for evaluation of data in experiments, where a population parameter μ is known. It may be e.g. physiological value of a biochemical indicator – this value is considered as a constant. Then in the experiment, we verify a null hypothesis whether the test sample (under a treatment) comes from a population with this known parameter μ **(H_0: μ = const.).** An alternate hypothesis is H_A: μ ≠ const.

Method:

1) We calculate sample mean and variance

2) We calculate a **test statistic**:

$$t = \frac{|\bar{x} - \mu|}{\sqrt{\frac{s^2}{n}}}$$

$\bar{x}$ -sample mean, μ- population mean, s – sample SD, n–number of items in sample

3) We specify **degree of freedom** for the test: ν = n-1

4) We compare calculated *t* with the tabulated critical value $t_{(\alpha,\nu)}$, where ν = n-1 and α = 0.05 (or 0.01).

- If $t > t_{(\alpha,\nu)}$ ⇒ we **reject H_0: μ = const.** There is a statistically **significant** difference between tested means at the α = 0.05 level ($P<0.05$)

  or **highly significant** difference at the α = 0.01 level ($P<0.01$).

It means that the treatment has been *effective* – it caused a change of the mean in comparison with the known population mean = const., i.e. the tested sample comes from another population with $\mu \neq$const.

- If $t \leq t_{(\alpha,\nu)} \Rightarrow$ we **accept H_0: μ = const.** (i.e. H_0 is true). There is a statistically **insignificant** difference between tested means at the specific α level (P>0.05).

  It means that "the treatment has been *ineffective*" – it did not cause a change of the mean in comparison with the known population mean = const., i.e. the tested sample comes from the another population with $\mu \neq$const.

Example:

In a population of dairy cows the mean value of glucose in blood serum is μ = 3.1 mmol.l^{-1}. After applying an energy preparation glucose level in serum in 10 cows selected at random was measured:
3.1, 2.7, 3.3, 3.1, 3.1, 3.2, 3.0, 2.8, 2.9, 2.7.

Does the preparation influence the glucose level in serum?

Method:

1) We calculate sample statistics:

$\bar{x} = 3.0 \qquad \nu = 9$

$s = 0.21$

$s^2 = 0.044$

2) Mean value known for the whole population: μ = 3.1 mmoll^{-1}.

3) We calculate test statistic t:

$$t = \frac{|\bar{x} - \mu|}{\sqrt{\frac{s^2}{n}}} = \frac{|3.0 - 3.1|}{\sqrt{\frac{0.21^2}{10}}} = 1.58$$

4) Critical value found in statistical tables of the t-distribution: $t_{crit.(0.05;\ 9)} = 2.262$

5) We compare the calculated test statistic with the critical value:

$t < t_{crit.} \Rightarrow$ there is a statistically <u>insignificant difference</u> between means (P>0.05).

(H_0 is not rejected; the sample comes from the population with μ=3.1).

6) *Conclusion:*

Preparation used in experiment is <u>not effective</u> (to change the glucose level in dairy cows).

6.2.2 Samples comparison (Two-sample *t*-test)

This variant of *t*-test is used for evaluation of data in experiments, where a population parameter μ is not known. We compare data of 2 samples that comes either from one group of subjects measured twice (typically before and after treatment – "paired experiment", "dependent samples") or from two different random groups of subjects (typically treated test group and untreated control group) – "unpaired experiment".

A) Paired *t*-test (paired experiment, dependent samples)

Data evaluated in this variant of *t*-test come from paired subjects; it means that the same subjects are submitted to two measurements (both test and control treatments are performed in one group of subjects). Such a situation represents for instance the weight before and after 2 weeks since the beginning of a new diuretic medication. In that case the outcome we are interested in, i.e. the weight, is evaluated before and after the medication of the diuretic using the same individual. Therefore we get "matched" data from these repeated measurements.

Note that paired experiments are, in principle, more robust than unpaired experiments. For example, if each individual taking a diuretic loses 2 kg in 2 weeks then you can feel comfortable that this represents an effect of the diuretic. In contrast, in two random groups, if one receives placebo (control) and the other the diuretic, a 2 kg weight reduction in the diuretic-treated group is less strong evidence of effectiveness of the medication. This is because the variation of weight in each group is larger than 2 kg, so it is unclear whether the difference between the two groups is a real effect of the diuretic or simply a random small difference in mean weight between the two groups. It follows that, when possible, it is better to evaluate a given manipulation in the same subject.

The first step in the procedure of paired *t*-test consists in calculation of differences between paired (matched) values: $\Delta x_i = x_{test} - x_{control}$. Then we calculate the sample mean $\bar{x}$ and SD (standard deviation) of the differences Δx_i. We test a hypothesis that population mean μ of the measurements before and after the treatment are equal (i.e. mean value μ of the differences Δx_i between matched measurements is equal to 0). Then the null hypothesis is $\mathbf{H_0: \mu_{differ.}=0}$ and an alternate hypothesis is $H_A: \mu_{differ.} \neq 0$).

Method:

1) We calculate differences between paired values, mean and standard deviation of the differences.

2) We calculate **test statistic for paired *t*-test**:

$$t = \frac{|\bar{x}|}{\sqrt{\frac{s^2}{n}}}$$

$\bar{x}$ - mean of differences between paired values, s^2 – variance of differences, n – number of pairs

3) We specify **degree of freedom for the test:** $\nu = n-1$

4) We compare the calculated *t* with the tabulated critical value $t_{(\alpha,\nu)}$, where $\nu = n-1$ and $\alpha = 0.05$ (or 0.01):

- If $t > t_{(\alpha,\nu)} \Rightarrow$ **H_0 is rejected**, i.e. difference between means is statistically **significant** (at $\alpha = 0.05$)

  or **highly significant** (at $\alpha = 0.01$)

  It means that the treatment has been effective: mean μ after the treatment is different from mean μ before the treatment.

- If $t \leq t_{(\alpha,\nu)} \Rightarrow$ **H_0: $\mu_{differ.} = 0$ is true,** i.e. difference between means of values before and after the treatment is statistically **insignificant** (at the specific α level).

  It means that the treatment has not been effective: mean μ after the treatment is the same as the mean μ before the treatment.

Example:

Determine a weight change of twelve rats after being subjected to a regimen of forced exercise. Each weight change (in g) is the weight after exercise minus the weight before: 0.2, -0.5, -1.3, -1.6, -0.7, 0.4, -0.1, 0.0, -0.6, -1.1, -1.2, -0.8. Does the exercise cause any significant change in rat weight?

Method:

1) We calculate sample statistics:

$\bar{x} = -0.6g \qquad \nu = 11$

$s^2 = 0.40g^2$

2) We calculate test statistic for the paired t test:

$$t = \frac{|\bar{x}|}{\sqrt{\frac{s^2}{n}}} = \frac{|-0.6|}{\sqrt{\frac{0.40}{12}}} = 3.39$$

3) Critical values found in statistical tables of the t-distribution: $t_{crit.(0.05;\,11)} = 2.201$

$t_{crit.(0.01;\,11)} = 3.106$

4) We compare the calculated test statistic with critical value:

$t > t_{crit.(0.01)} \Rightarrow H_0$ is rejected, there is a statistically <u>highly significant difference</u> between mean before and after exercise ($P<0.01$).

5) *Conclusion:*

The exercise causes a <u>highly significant weight loss</u> in rats.

B) Unpaired *t*-test (unpaired experiment, independent samples)

Data evaluated in this variant of *t*-test come from two independent groups if individuals - we deal with "unmatched" data. Typically there is one test group, to which we apply some tested treatment, and one control group without any treatment. We can also compare two groups with different treatments in experiment, when we are interested in the possibility, whether there is any difference between effects of these two treatments.

Unpaired *t*-test is then used to determine whether the means of two independent samples are different enough to conclude that they were drawn from *different populations*. We test the null hypothesis, whether a population mean value μ_1 in the test group (treated) is the same as the population mean value μ_2 in the control group: **H_0: $\mu_1 = \mu_2$.** A two-tailed alternate hypothesis is H_A: $\mu_1 \neq \mu_2$.

The populations sampled can have different variability – this variability affects the calculation of *t*-test. Therefore at first we have to determine the **difference between variances** (through *F*-test) to specify what type of calculation formula we need to use for the following *t*-test. Therefore the first step of the procedure of unpaired *t*-test consists in calculation of sample statistics (estimated mean, standard deviation and variance for both samples compared):

Sample 1 (n_1): we calculate $\bar{x}_1$, s^2_1

Sample 2 (n_2): we calculate $\bar{x}_2$, s^2_2

In the following step, we determine by means of F-test whether there is any difference between population variances:

F-test:

$$F = \frac{higher\,(s_1^{\,2}, s_2^{\,2})}{lower\,(s_1^{\,2}, s_2^{\,2})}$$

Degree of freedom: for numerator $\nu_N = n - 1$

for denominator $\nu_D = n - 1$

We compare the calculated F with the tabulated critical value of F-distribution that we find out according to the chosen α and degrees of freedom: ν_N (DF of numerator) and ν_D (DF of denominator).

According to the F-test result:

- If $F \leq F_{\alpha(\nu N,\ \nu D)} \Rightarrow$ populations compared have the same variability ($\sigma_1^2 = \sigma_2^2$), we use the following formula for the unpaired *t*-test:

a) $\sigma_1^2 = \sigma_2^2$:

Test statistic:
$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\frac{(n_1 - 1) * s_1^{\,2} + (n_2 - 1) * s_2^{\,2}}{n_1 + n_2 - 2} * \frac{n_1 + n_2}{n_1 * n_2}}}$$
DF: $\nu = n_1 + n_2 - 2$

In a special case of equal sizes of samples compared:

For $n_1 = n_2 = n$: $t = \dfrac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\dfrac{s_1^2 + s_2^2}{n}}}$ DF: $\nu = (n-1)\ .2$

- If $F > F_{\alpha(\nu N,\ \nu D)} \Rightarrow$ populations compared have different variances ($\sigma_1^2 \neq \sigma_2^2$), we use the following formula for the unpaired *t*-test:

b) $\sigma_1^2 \neq \sigma_2^2$:

Test statistic: $t = \dfrac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

DF: $\nu = \dfrac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$ (For $n_1, n_2 > 30$: $\nu = \infty$)

(For $n_1 = n_2 = n$: $t = \dfrac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\dfrac{s_1^2 + s_2^2}{n}}}$)

Conclusion:

- If $t > t_{(\alpha,\nu)} \Rightarrow$ statistically **significant** difference between μ_1 and μ_2 (at $\alpha = 0.05$) or

  highly significant difference (at $\alpha = 0.01$)

  i.e. samples was not drawn from the same population (it means that the "experiment has been effective" and caused a change of mean value in the treated group compared to the control: **$\mu_1 \neq \mu_2$).**

  Therefore **H_0: $\mu_1 = \mu_2$ is rejected.**

- If $t \leq t_{(\alpha,\nu)} \Rightarrow$ statistically **insignificant** difference between μ_1 and μ_2 at the specific α. **H_0: $\mu_1 = \mu_2$** is true; i.e. samples was drawn from the same population.

  (it means, that "the treatment has not been effective")

Example:

Determine a drug effect on the change in blood-clotting times (in minutes) in pigs.

Times of individuals treated with drug (T): 9.9, 9.0, 11.1, 9.6, 8.7, 10.4, 9.5.

Times of untreated control individuals (C): 8.8, 8.4, 7.9, 8.7, 9.1, 9.6, 8.7.

Method:

1) We calculate statistics of both samples:

Test sample: $\bar{x}_1 = 9.7\,\text{min}$ $\quad n_1 = 7$

$s_1^{\,2} = 0.67\,\text{min}^{\,2}$ $\quad \nu_1 = 6$

Control sample: $\bar{x}_2 = 8.7\,\text{min}$ $\quad n_2 = 7$

$s_2^{\,2} = 0.28\,\text{min}^{\,2}$ $\quad \nu_2 = 6$

2) We calculate test statistic for the *F* test: $F = \dfrac{0.67}{0.28} = 2.37$

4) Critical value for the *F* test: $F_{\text{crit. }(0.05;6,6)} = 5.820$

$\boldsymbol{F < F_{\text{crit.}} \Rightarrow \sigma_1^2 = \sigma_2^2}$

5) We calculate test statistic for the unpaired *t*-test for equal variances :

$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\dfrac{s_1^{\,2} + s_2^{\,2}}{n}}} = \frac{9.7 - 8.7}{0.26} = 3.83$$

6) Degree of freedom for the *t*-test: $\nu = (n - 1).2 = 12$

7) Critical values for the t test: $t_{\text{crit. }(0.05;12)} = 2.179$

$t_{\text{crit. }(0.01;12)} = 3.055$

8) We compare the calculated test statistic with critical value:

$t > t_{\text{crit. }(0.01)} \Rightarrow$ there is a statistically highly significant difference between tested means ($P<0.01$).

(H_0 is rejected).

9) *Conclusion*:

Drug administration causes a highly significant longer blood-clotting time in pigs.

Chapter 7

Non-Parametric Tests

Non-parametric tests belong to special statistical methods that comprise procedures not requiring the estimation of the population parameters (mean and variance) and not stating hypotheses about parameters. As these methods also typically do not make assumptions about the nature of the distribution (e.g., normality) of the sampled populations (although they might assume that the sampled populations have the same dispersion or shape), they are sometimes referred to as *distribution-free tests*. Non-parametric tests are often called "rank tests" as their calculations are based on sum of ranks of values measured in experiment.

Non-parametric tests may be applied in any situation where we would be justified in employing a parametric test, such as two-sample *t* test, as well as in instances when the assumptions of the latter are untenable. However, if either the parametric or non-parametric approach is applicable, then the latter will always be less powerful than the former (difference between tested data sets must be considerable to achieve a statistical significance).

Non-parametric tests are especially employed when dealing with ordinal scale data (data that consist of ranks) and numerical scale data when normality is not assumed, but they may also be employed when dealing with numerical data that follow normal distribution (for preliminary analyses especially, as calculations of non-parametric tests are often more quick and simpler then parametric ones). Non-parametric tests may also be useful in instances when dealing with small sample sizes – then sample frequency distribution is insufficient to tell us whether the assumption of normality may be confirmed.

In non-parametric tests, the hypothesis is verified that data from both samples were drawn from the same population, i.e. that they have the same dispersion or shape of the distribution curve (**null hypothesis**). We then determine (similarly to parametric tests) the probability (*P*-value) that the observed difference in our samples could have arisen by chance. A low *P*-value ($P<0.05$), is often regarded as statistical evidence to reject the null hypothesis and conclude that there is significant difference in the results obtained from our two experimental conditions.

7.1 Mann-Whitney *U*-Test (Rank-Sum Test)

(Two-Sample Rank Testing)

For this test, as for many other non-parametric procedures, the actual measurements are not employed, but we use instead the ranks of the measurements.. The data of both samples compared are arranged into one (mixed) sample and may be ranked either from the highest to lowest or from

the lowest to the highest values. The samples compared can consist of equal or unequal number of the observations. E.g. if data in samples are arranged from the highest to the lowest, then the highest value in either of the two samples compared is given rank 1, the second highest value is assigned rank 2, and so on, with the lowest value being assigned rank N, where

$$N = n_1 + n_2.$$

We calculate the **Mann-Whitney test statistics**:

$$U = n_1 * n_2 + \frac{n_1 * (n_1 + 1)}{2} - R_1,$$

$$U' = n_1 * n_2 + \frac{n_2 * (n_2 + 1)}{2} - R_2,$$

where n_1 and n_2 are the number of observations in samples,

R_1 is the sum of the ranks of the observations in Sample1,

R_2 is the sum of the ranks of the observations in Sample2.

When U is already calculated U′ can be also found more quickly as

$$U' = n_1 * n_2 - U$$

The **larger** of the two calculated test statistics ***U* and *U′*** is compared to the critical value $U_{\alpha, n1, n2}$, found in Appendix 6 (Critical Values for Mann-Whitney *U*-test). This table assumes that $n_1 > n_2$; if $n_2 > n_1$, simply use $U_{\alpha, n2, n1}$ as the critical value.

Then:

If the **larger** from ***U and U′*** > $U_{\alpha, n1, n2}$ => we reject H_0 at the α level of significance (i.e. samples tested were not drawn from the same population, there is a significant difference between populations sampled - they don't have the same shape of the distribution curve). It means that "the treatment used in the experiment was effective" at the α.level of significance.

If the **larger** from ***U and U′*** < $U_{\alpha, n1, n2}$ => we accept H_0 at the α level of significance (i.e. samples tested were drawn from the same population, there is an insignificant difference between populations sampled - they have the same shape of the distribution curve). It means that "the treatment used in the experiment was not effective" at the α.level of significance.

Note that neither parameters nor parameter estimates are employed in the statistical hypotheses or in the calculations of test statistics *U* or *U′*.

We may assign ranks either from large to small data, or from small to large, calling the smallest datum rank 1, the next smallest rank 2, and so on. The value of *U* obtained using one ranking procedure will be the same as the value of *U′* using the other procedure.

Example:

By means of Mann-Whitney test for non-parametric testing find out whether there is some difference between the heights of male and female students.

Method:

1) We arrange data from the highest to the lowest and find out ranks of male and female heights in this arranged (mixed) sample:

193 >**188** > **185** > **183** > **180** > **178** > 175 >173 > **170** > 168 > 165 > 163

Heights of males (cm)	***Heights of females (cm)***	***Ranks of male heights***	***Ranks of Female heights***
193	175	1	7
188	173	2	8
185	168	3	10
183	165	4	11
180	163	5	12
178		6	
170		9	
$n_1 = 7$	$n_2 = 5$	$R_1 = 30$	$R_2 = 48$

2) We calculate test statistics:

$$U = n_1 * n_2 + \frac{n_1 * (n_1 + 1)}{2} - R_1 = 7 * 5 + \frac{7 * 8}{2} - 30 = 33$$

$$U' = n_1 * n_2 - U = 7*5 - 33 = 2$$

3) Critical value for the *U*-test**:** $U_{\alpha,\ 7,5} = 30$

4) We compare the calculated test statistic with critical value:

U > $U_{0.05,\ 7,5}$ => we reject the null hypothesis, i.e. there is statistically significant difference between data sets ($P < 0.05$).

5) *Conclusion:*

There is a statistically significant difference between male and female heights at the significance level $\alpha = 0.05$.

7.2 Wilcoxon Signed-Rank Test

(A non-parametric test for paired samples)

The Wilcoxon paired-sample test is a non-parametric analogue to the paired-sample *t* test, just as the Mann-Whitney test is a non-parametric procedure analogous to the two-sample *t* test. Whenever the paired-sample *t* test is applicable, the Wilcoxon paired-sample test is also applicable. But there are instances when the Wilcoxon paired-sample test is applicable and the parametric paired-sample *t* test is not, as when one can not assume that data are from a normal distribution.

Since the two groups of measurements to be compared are obtained from the same sample of subjects tested twice (before and after treatment), instead of analysing the raw data from each group separately we look only at the *differences* between the pre-treatment and post-treatment values for each subject. By subtracting the first value from the second for each subject and then analysing only these paired differences we exclude all the variation in our data that results from the differing initial values of individual subjects.

The testing procedure involves the calculation of differences, as does the paired-sample *t* test. Then we rank the absolute values of the differences, from low to high, and affixes the sign of each difference to the corresponding rank. If there are some equal differences (in absolute value), then they will get so called "average rank" (e.g. if the first and second difference has the same value, then they both will get the 1.5). If there is some difference that is equal to 0 ($d_i = 0$), then the *i*th pair is omitted from the analysis.

Then we sum the ranks with a plus sign (we shall call this sum W_+) and the ranks with a minus sign (calling this sum W_-). Having calculated either W_+ or W_-, the other can be determined also as:

$$W_- = \frac{n*(n+1)}{2} - W_+$$

or

$$W_+ = \frac{n*(n+1)}{2} - W_-$$

The **smaller** of the two calculated $\boldsymbol{W_+}$ ***and*** $\boldsymbol{W_-}$ is compared to the critical value $W_{\alpha, n}$ from the Tables of critical values for Wilcoxon signed rank test (Appendix 7):

If the smaller from $\boldsymbol{W_+}$ ***and*** $\boldsymbol{W_-} < \boldsymbol{W_{\alpha, n}}$ => we reject H_0, i.e. difference between measurements before and after treatment is statistically significant at the α.level („The treatment used in the experiment was effective").

If the smaller from $\boldsymbol{W_+}$ ***and*** $\boldsymbol{W_-} > \boldsymbol{W_{\alpha, n}}$ => we accept H_0, i.e. difference between measurements before and after treatment is statistically insignificant at the α.level („The treatment used in the experiment was not effective").

Example:

By means of Wilcoxon paired-sample test for non-parametric testing find out whether there is some difference between the lengths of hind- and forelegs in deer.

Method:

1) We calculate differences d_i between paired values:

Deer	Hind leg length (cm)	Foreleg length (cm)	Difference (d_i)
1	142	138	4
2	140	136	4
3	144	147	-3
4	144	139	5
5	142	143	-1
6	146	141	5
7	149	143	6
8	150	145	5
9	142	136	6
10	148	146	2

2) We arrange the absolute values of differences into an ascending row:

$|-1| < 2 < |-3| < 4 = 4 < 5 = 5 = 5 < 6 = 6$

Note that there are some equal values of differences d_i.

3) We determine the ranks of differences and apply appropriate sign to the rank according to the difference. Note that there are several "average ranks" used for equal differences (E.g. there are three differences $d_i = 5$ in the ascending row above, therefore all of them will get the rank 7 instead of original ranks 6, 7, 8):

Deer	Difference (d_i)	Ranks of $\|d_i\|$	Signed ranks of $\|d_i\|$
1	4	4.5	4.5
2	4	4.5	4.5
3	-3	3	-3
4	5	7	7
5	-1	1	-1
6	5	7	7
7	6	9.5	9.5
8	5	7	7
9	6	9.5	9.5
10	2	2	2

4) We calculate sums of plus and minus ranks :

$W_+ = 4.5 + 4.5 + 7 + 7 + 9.5 + 7 + 9.5 + 2 = 51$

$W_- = 3 + 1 = 4$

5) Critical value $W_{0.05, 10} = 8$

6) Since $W_- < W_{0.05, 10}$ => we reject H_0 at the 5% level of significance.

7) *Conclusion:*

Deer hind leg lengths are not the same as the foreleg lengths ($P < 0.05$).

Chapter 8

Relationship Between 2 Data Sets

(Quantitative Data)

2-dimensional statistics is used for evaluation of relation between 2 variables (biological characters) in a data set. Two variables are related if their values correspond to each other in some systematic way. For example, tall people are usually heavier than short people; therefore height and weight are related variables. Two-dimensional statistics helps us to establish the nature of the relations between variables. In particular we wish to know how strong is the relationship we have observed in biological data and how reliable is our observation of a relationship.

We try to qualify and describe the relationship between 2 variables monitored: one being an *independent* and one being a *dependent* variable. The dependent variable can be predicted by the independent variable (if we know values of the independent variable, we can calculate the values of the dependent variable).

8.1 Functional vs. Statistical Relationship

We can distinguish between ***2 basic groups of relations*** between variables in general:

A) Functional Relationship (typical for relations in mathematics and physics):

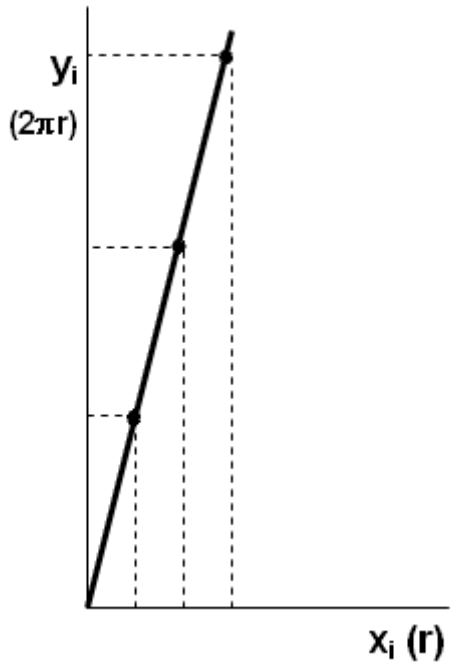
The magnitude of one of the variables (*the dependent variable)* is assumed to be determined by (i.e. is a function of) the magnitude of the second variable *independent variable*, whereas the reverse is not true. Each value of the independent variable (x_i) corresponds to one exact value of the dependent variable (y_i) in this type of relation. Sometimes, the independent variable is called the "predictor", or "regressor", variable and the dependent variable is called the "response", or "criterion", variable.

Such a relation can be described by means of an exact equation (formula): e.g., relation between circle radius (***r***) and its circumference ($y=2\pi r$) or surface ($y=\pi r^2$).

The functional relation is an example of the strictly causal relationship - it is not affected by random. Such a dependent relationship is termed a *regression*; the term *simple regression* refers to the fact that only two variables are being considered.

It is very convenient to use a graph in order to describe this functional relation, using the ordinate (*Y* axis) for the dependent variable (conventionally termed *Y*) and the abscissa (*X* axis) for the independent variable (*X*). An example for the chart of functional relation is shown in the Figure 8.1; it is a linear function in the case of relation between circle radius and its circumference.

Fig. 8.1 Linear function – relation between circle radius (*X*) and its circumference (*Y*)



B. Statistical Relationship (Correlative)

This relationship is typical for relations between data in biology and medicine: there is no exact functional relationship, because most of biological characters are very changeful and unstable; therefore relations result from this variability and they are relative (statistical, correlative) only. These relations in biology and medicine are very complicated – there are many different causes including random effects that we are not able to exclude during our monitoring.

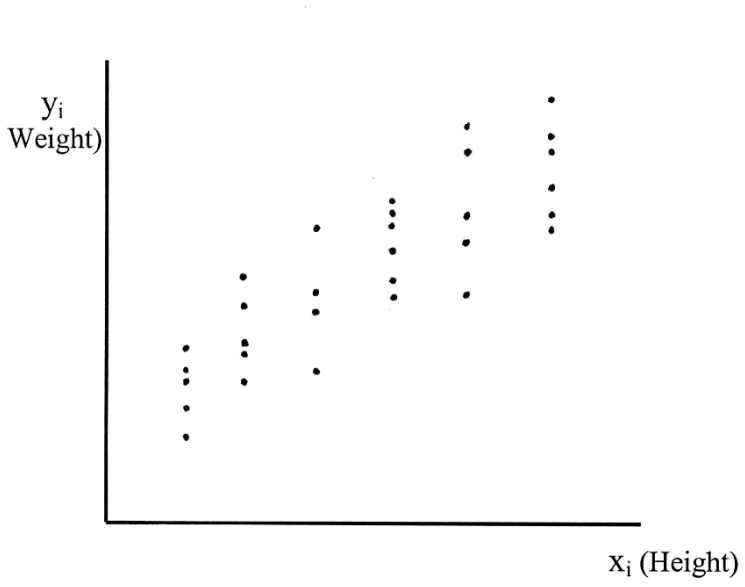
This relation is more or less **free** - the magnitude of one of the variables **probably** changes as the magnitude of the second variable changes. Each value of x_i corresponds to several random values of y_i and also the reverse is possible ("variables are correlated"). In such a case it is not very often reasonable to consider that there is an independent and dependent variable (e.g. fore- and hind leg lengths in animals, human height and weight, arm and leg lengths, etc.). It might be found that an individual with long arms will in general possess long legs, so a relationship may be describable; but there is no justification in stating that the length of the limb is dependent upon the length of the other. In such situations, correlation, rather than regression, analyses are called for, and both variables are theoretically to be random-effects factors.

We use so called correlation chart ("**scatter diagram", "dot plot"**) for the graphical description of such statistical relationship – each point represents a pair of *X* and *Y* values measured in one member of sample under study. One pair of X and Y data may be denoted as (x_1, y_1), another

as (x_2, y_2), another as (x_3, y_3), etc. These corresponding values on the axis x and y for one point in the scattered plot are called "*correlation pairs"* (x_i, y_i).

An example for the scattered diagram of correlative relation is shown in the Figure 8.2; it is a correlation between height and weight in men.

Fig. 8.2 Dot plot for correlative relationship between height and weight in men



If the points in the scattered diagram are clustered in some direction – it means that there is some relation between biological characters monitored; correlation may be positive – "direct relation" (Fig. 8.3) or negative – "inverse relation" (Fig. 8.4).

Fig. 8.3 Positive correlation

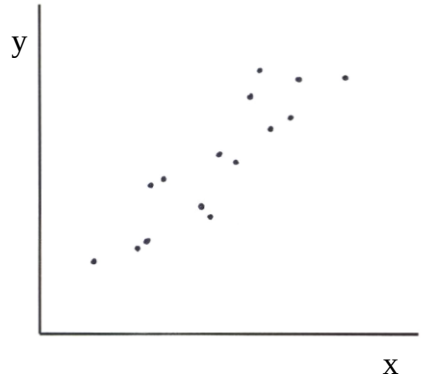
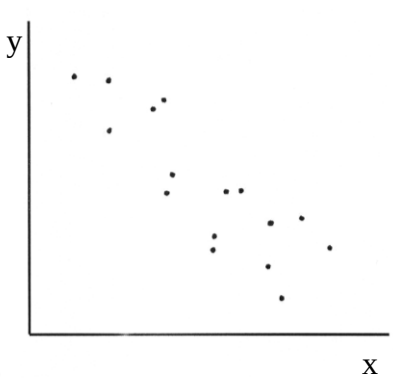
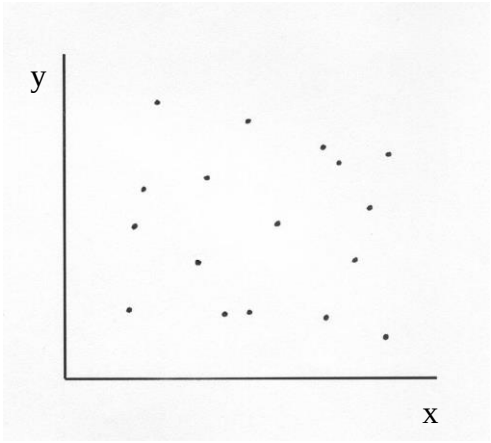


Fig. 8.4 Negative correlation



If the points in the scattered diagram are irregularly scattered in the area – it means that there is no correlation between biological characters monitored (Fig. 8.5).

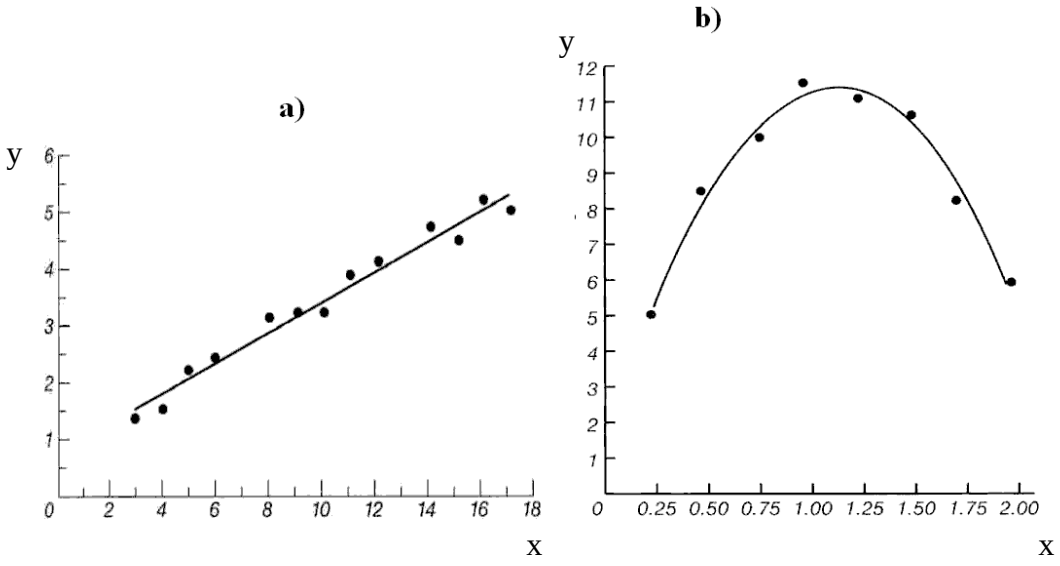
Fig. 8.5 No correlation



If we wish to know the strength of any relationship we have observed, and how reliable this observation is, we need to employ statistical techniques called *correlation and regression analysis*. Correlation is a measure of the relationship between two (or more) variables that helps us determine whether the variables really are related and the degree to which they vary together. Regression is a statistical tool for determining the mathematical relationship between one or several independent or *predictor* variables and a single dependent or *criterion* variable, allowing us to calculate the value of one variable given known values of the other variables. If we need to evaluate and describe the statistical relation in a graphical presentation - we have to estimate the *best-fit function* that can be used for description of this relationship, and to determine its equation (to calculate coefficients for this equation – either linear or nonlinear).

According to the allocation of points in the scatter diagram we can distinguish between two types of correlative relation: Linear or Non-linear correlation (Fig. 8.6). These two types of correlative relationships differ in the way of their statistical evaluation and analysis.

Fig. 8.6 Linear (a) and Non-linear (b) correlation



8.2 Linear Correlative Relationship

The linear function is the most frequently used equation that we can use for estimation and description of some correlative relationship between two variables (biological characters) monitored in biology and related sciences. We need to employ the simplest case of regression analysis, the simple linear regression, in this situation. Data amenable to simple regression analysis will consist of a dependent variable that is a random-effect factor and an independent variable that is either a fixed-effect or a random-effect factor. The data can be visualised in a scatter-plot and analysed by fitting the ***best straight line*** to the points. The simplest and most commonly used fitting technique of this sort is named *least squares*. The name comes from minimizing the sum of squared vertical distances from the data points to the proposed line.

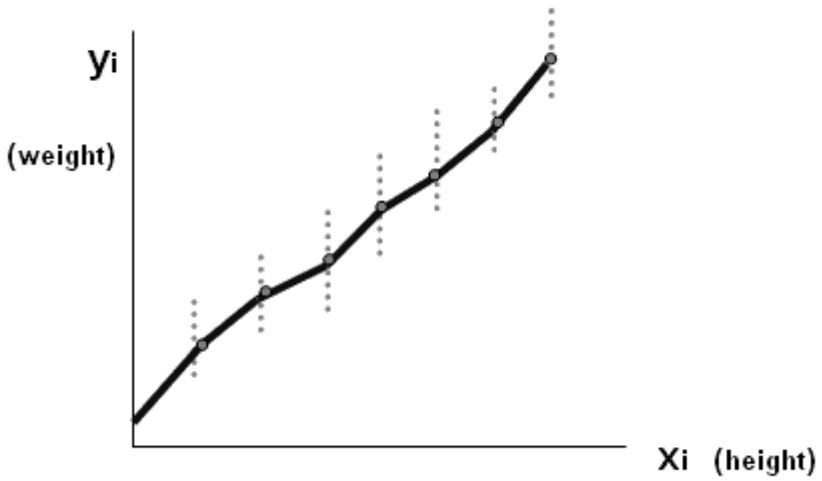
The analysis and description of the statistical relation is usually performed in the following steps:

1) The construction of an ***empirical curve*** that describes the relation in a ***sample*** (estimates the supposed theoretical line for the whole population)***:***

We measure several values y_i for the same value x_i (e.g. in several men that have the same height (x_i) we measure their weights; we obtain several random values y_i). We calculate an average from these values y_i in the appropriate x_i, and then we join these averages in order to construct the empirical curve that describes the relation in the particular sample monitored in our study. This empirical curve can serve as the estimation of the best-fit linear function.

An example for the empirical curve in the case of relationship between height and weight in men is presented in the Figure 8.7.

Fig. 8.7 Empirical curve for the correlative relation



2) The construction of a ***regression line*** (i.e. calculation of equation for this best-fit line) that can be used for description of the relation in the whole ***population.***

We need to calculate coefficients of the best-fit regression equation using regression analysis: $\mathbf{y = a + bx}$

Coefficients ***a*** and ***b*** in the regression equation determine ***properties of the line***:

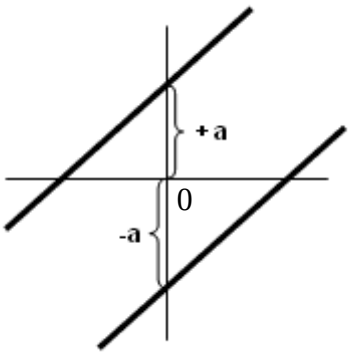
a (called intercept) – represents the intercept point on the axis y for x=0,

b (slope, regression) = tg α (α - an angle that is formed by the line and the axis x).

We need always keep in mind that coefficients *a* and *b* are only the best estimates of the true coefficients denoted α and β of the theoretical regression line that would uniquely describe the functional relationship existing in the whole population.

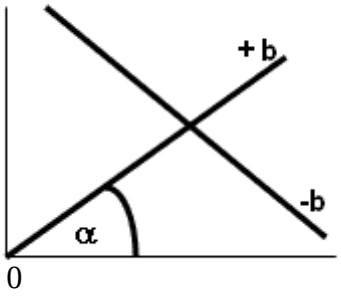
Figures 8.8 and 8.9 demonstrate properties of the regression line that are determined by coefficients *a* and *b* in linear equation.

Fig. 8.8 Coefficient *a* represents the intercept point on the axis y



If the coefficient *a* is a positive value, then the line intersects the axis *y* above the value 0, if the coefficient *a* is a negative value, the line intersects the axis y below the value 0.

Fig. 8.9 Coefficient *b* represents the slope of the line



If the coefficient *b* is a positive value, then the line is ascending (it indicates a direct relation between *x* and *y* variables), if the coefficient *b* is a negative value, then the line is descending (it indicates an inverse relation between *x* and *y*).

8.2.1 Regression Analysis

Regression analysis is a statistical technique that calculates the coefficients (parameters) of the linear function: $y = a + bx$. The calculation results from sample data - correlation pairs (x_i, y_i), measured for each of *n* number of individuals in the sample under study.

Calculation formula for regression coefficient *b (slope)*:

$$b = \frac{n \cdot \sum x_i y_i - \sum x_i \cdot \sum y_i}{n \cdot \sum x_i^2 - \left(\sum x_i\right)^2}$$

The coefficient *a* (intercept) is derived from calculated *b* through the formula:

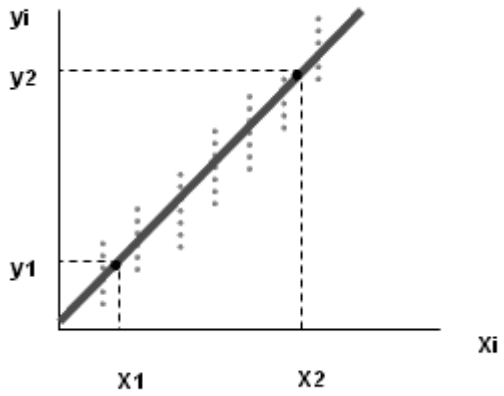
$$a = \frac{\sum y_i - b \cdot \sum x_i}{n}$$

After calculation of the equation for regression line, we need to determine two points for construction of the ***theoretical regression line.*** We can choose any value x_1 and calculate its appropriate value y_1 (according to the calculated regression line equation), and then choose another x_2 and calculate appropriate y_2: $y_1 = a + bx_1$

$y_2 = a + bx_2$

Figure 8.10 demonstrates construction of the best-fit regression line for the scattered diagram in the case of direct correlative relationship.

Fig. 8.10 Construction of the theoretical regression line



Knowing the parameter estimates *a* and *b* for the linear regression equation, we can predict the value of the dependent variable expected at a stated value x_i. A word of caution is in order concerning predicting y_i values from regression equation. Generally, it is an unsafe procedure to extrapolate from regression equations – that is, to predict y_i values for x_i values outside the observed range of x_i. What the linear regression actually describes is *Y* as a function of *X within the range of observed values of X*. For values of *X* above or below this range, the function may not be the same (i.e., α and/or β may be different); indeed, the relationship may not even be linear in such ranges, even though it is linear within the observed range. If there is good reason to believe that the described function holds for *X* values outside the range of those observed, then we may cautiously extrapolate. Otherwise, beware.

8.2.2 Correlation Analysis

Correlation analysis is the statistical technique used for determination of association level between variables in the analysis of the correlative relation monitored. In simple linear correlation, we consider the linear relationship between two variables *X* and *Y*, whereas neither is assumed to be functionally dependent upon the other. An example of a correlation situation is the relationship between the wing length and tail length of a particular species of bird.

We calculate a ***correlation coefficient r*** that determines tightness (closeness) of the relation between variables *X* and *Y* (and also determines the measure of dispersion of points around the theoretical regression line in the scatter diagram). The calculation results from sample data - correlation pairs (x_i, y_i), measured for each of individuals in the sample under study.

Calculation formula for correlation coefficient r:

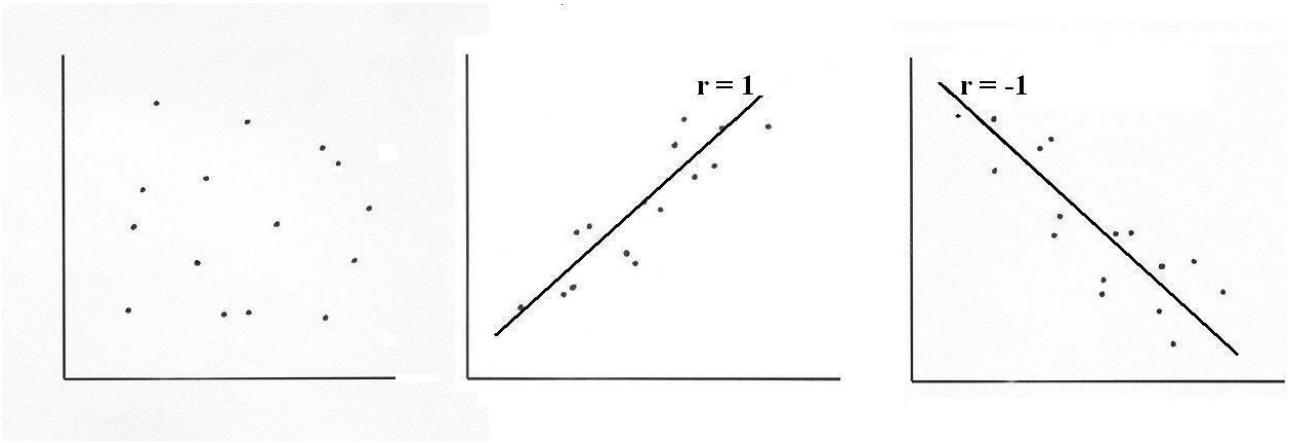
$$r = \frac{\sum (x_i - \bar{x}).(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

Correlation coefficient *r* is called “Parametric” (or Pearson’s) correlation coefficient, as we need parameters (means of variables *X* and *Y*) for its calculation. Therefore it should be used in data that follow Gaussian normal distribution only.

Values of correlation coefficient *r* are located within the interval ⟨-1 ; +1⟩. The larger is the absolute value of ***r***, the closer is the correlation between *X* and *Y* variable. A positive correlation coefficient implies that for an increase in the value of one of the variables, the other variable also increases in value; a negative correlation coefficient indicates that an increase in value of one of the variables is accompanied by a decrease in value of the other variable. If the correlation coefficient *r* = 0, and one has a zero correlation, denoting that there is no linear association between the magnitudes of two variables; that is, a change in magnitude of one does not imply a change in magnitude of the other. Correlation coefficient *r* = +1 represents total (functional) direct relation

(ascending line), correlation coefficient r = -1 represents total (functional) inverse relation (descendent line). Figure 8.11 presents these considerations graphically.

Fig. 8.11 Dot plots of correlations with different correlation coefficients



$r = 0$	$r > 0$	$r < 0$
No correlation	Direct (positive) correlation (X, Y increase together)	Inverse (negative) correlation (When X increases, Y decreases)
	$r = +1$ (total relation: ascending line)	$r = -1$ (total relation: descending line)

8.2.3 Significance of the Correlation Coefficient

The correlation coefficient r that we calculate from a sample is only an estimate of an actual correlation coefficient in the population (denoted ρ). If we need to know whether the correlation in the population really exists, we have to test a hypothesis of the independence (**H_0: ρ=0**) using ***t*-test**:

Test statistic: $$t = \frac{r}{s_r}$$

Where s_r is the ***standard error of the correlation coefficient r*** and is calculated using the following formula:

$$s_r = \sqrt{\frac{1-r^2}{n-2}}$$

Degree of freedom needed for tabulated critical value: ν = n-2

We compare the calculated ***t*** with the critical value (Appendix 2: Critical values for Student's *t*-distribution) according to the chosen α and given v = n-2:

If $t > t_{\alpha(v)}$ => H_0 is not true, the correlation between *X,Y* really exists in the population sampled (*r* is significant),

If $t \leq t_{\alpha(v)}$ => H_0 is true, the correlation between *X,Y* really does not exist in the population sampled (*r* is insignificant).

8.3 Non-linear Correlative Relationship

There are many difficulties in calculations of non-linear regression equations; therefore it is very convenient to use a computer in case of such non-linear relations. A statistical software with options for so called ***polynomial regression*** is useful. In such cases, we get different regression models (curves) computed by means of this polynomial regression. The most common non-linear regression is the quadratic equation: $y = a + b_1 x + b_2 x^2$ ("second-order polynom"). In this case, we need to calculate regression coefficients a, b_1, b_2 for this equation by means of a computer in order to find out the best-fit curve (parabola).

The calculation of a Spearman rank correlation coefficient represents another method for the analysis of the non-linear relation between variables in biology.

8.3.1 Spearman Rank Correlation Coefficient

If we deals with a non-linear relation between two variables or if we have data obtained from a bivariate population that is far from normal, then the correlation procedures discussed in the chapter 8.2 are generally not applicable. Instead, we may operate with the ranks of the measurements for each variable studied in these situations.

Calculation of the Spearman rank correlation coefficient is a non-parametric method, since we don't need parameters (means of variables *X* and *Y*) for calculation. This method may also be used for data sets that don't follow Gaussian normal distribution, and it can be used more generally – in both linear and non-linear correlations. This method may also be used in normal data sets, but non-parametric correlation coefficient is less forceful (less effective) than the parametric one. Therefore, the non-parametric correlation coefficient is mostly used only for preliminary calculations in normal data.

In the course of the calculation of Spearman rank correlation coefficient r_s, we use only the ranks of values instead of the actually measured values x_i, y_i,. The calculation results from the number of individuals (n) in the sample and correlation pairs (x_i, y_i).

Method:

First, we **arrange observed values** of variables *X* and *Y* separately in **two variant sequences** (ascending or descending rows). Then we assign appropriate ranks to the values in these variant sequences:

e.g.: $x_2 < x_4 < x_1 < x_5 < x_3 < x_8 < x_6 < x_7$ .....
Rank: 1 2 3 4 5 6 7 8 n

$y_3 < y_1 < y_5 < y_2 < y_4 < y_8 < y_7 < y_9$ .....
Rank: 1 2 3 4 5 6 7 8 n

If there are some equal values in the row, they get so called "average ranks": e.g., if x_4 and x_1 are equal then both get the rank 2.5 (calculated as = (2+3)/2)

Then we calculate $\mathbf{D_i}$: **differences between ranks** of corresponding x_i and y_i values:

e.g.: D_1=3-2, D_2=1-4, D_3=5-1, D_4=2-5 .....

Calculation formula for Spearman rank correlation coefficient:

$$r_S = 1 - \frac{6\sum D_i^2}{n.(n^2-1)}$$

:

Where:

D_i - differences between ranks of corresponding x_i and y_i values

n - number of members in the sample

After calculation, we compare computed r_s to the critical rank coefficient found in the statistical tables (Appendix 8: Tables of Spearman rank correlation) according to the chosen α and given n:

If $|r_s| > r_{crit.}$ => There is a ***significant correlation*** between *X* and *Y* variables (relation really exists in the population sampled),

If $|r_s| \leq r_{crit.}$ => There is an ***insignificant correlation*** between *X* and *Y* variables (relation does not really exist in the population sampled).

Example:

Calculate the Spearman rank correlation coefficient for the relation between wing and tail lengths among birds of a particular species:

Method:

1) We arrange observed values of variable *X* and *Y* in an ascending variant sequence and assign appropriate ranks to the values in these variant sequences.

2) We found appropriate values x_i and y_i and assign their ranks (see columns 1 - 4 in the following table).

3) We calculate differences D_i between ranks of corresponding x_i and y_i values (see columns 5 in the following table):

Wing length (X) [cm]	**Rank of X**	**Tail length (Y) [cm]**	**Rank of Y**	**Difference D_i**	**D_i^2**
10.2	1.5	7.1	1	0.5	0.25
10.2	1.5	7.2	2.5	-1	1
10.3	3	7.4	5	-2	4
10.4	4	7.4	5	-1	1
10.5	5	7.2	2.5	2.5	6.25
10.6	6	7.8	9.5	-3.5	12.25
10.7	7	7.4	5.	2	4
10.8	8.5	7.6	7	1.5	2.25
10.8	8.5	7.8	9.5	-1	1
11.1	10	7.9	11	-1	1
11.2	11	7.7	8	3	9
11.4	12	8.3	12	0	0

4) We calculate sum of squared differences : $\Sigma D_i^2 = 42.00$ (number of pairs: n =12)

5) We calculate Spearman rank correlation coefficient:

$$r_S = 1 - \frac{6\sum D_i^2}{n(n^2-1)} = 1 - \frac{6(42.00)}{1716} = 0.853$$

6) Critical r_s (for α=0.05, n=12) **= 0.587** $\Rightarrow$ Spearman rank correlation coefficient is statistically significant (at the level α=0.05).

7) ***Conclusion***:

Correlation between wing and tail lengths among birds of a particular species is statistically significant (it really exists in the population).

Chapter 9

Categorical Data

(Qualitative Data: On a Nominal Scale)

Sometimes an observed variable in biology can be described as *nominal data* ("*qualitative" data*) – they can have only 2 levels of their "quality": 0-1, yes-no, true-false, etc. We can not measure any values in these biological characters.

We can determine only a presence (or absence) of this quality in an individual in the samples under study.

E.g. when we evaluate survival rate (alive or dead), presence or absence of a disease (healthy or ill), anatomic anomaly (yes or no), incidence of parasites (positive or negative), gender (male or female), vaccinated or not vaccinated, etc.

Data of this sort are placed into named ***categories*** - so called "*qualitative classes*" (as opposed to being measured as a point on a scale or ranked in order); therefore they usually are referred to as *categorical data*. Categories of such data represent different variants of the observed biological character. In some biological characters, there can be only two categories or, in other biological character monitored, there may be more categories, e.g.:

- 2 categories: healthy-ill, alive-dead, male-female,
- More categories: eye colour – blue, brown, green, grey
  hair colour, hair type - long, short, medium, smooth, curly, etc.
  (each of these categories has 2 levels – "qualities": yes – no).

Categorical groups are formed in a natural way in most qualitative biological characters. However, sometimes, categorical groups may be formed also artificially, by dividing the scale upon which continuous data occur. If we were to categorize age by decade (50-59, 60-69, and 70-79 years), we would have age groupings, which we could name 1, 2, and 3. These groups could be considered as categories and categorical methods used. However, they fall into a natural rank order, as group 1 clearly comes before group 2, etc. Rank methods give better results than categorical methods, since rank methods are more powerful than categorical methods, which are not very powerful. *When ranking is a natural option, rank methods should be used.*

When dealing with categorical data, the basic statistic, ***count = frequency,*** is obtained by counting the number of "events" per category (individuals that possess the appropriate "quality") in a sample of total number of *n* individuals.

The symbol for number of "events" used in calculation formulas is:

$\boldsymbol{f_i}$ – *frequency* (count) in class *i*, and

n – *total number* of "events" (individuals) summed over all categories in a sample.

Another important statistic obtained from categorical data is the ***proportion*** of data in a category, which is the count in the category divided by the total number of ***n*** individuals in the sample under study. Proportion of data is a relative measure (in contrast to absolute frequencies) and gives us ***probability*** of data in the category. Multiplication by 100 yields percent (denoted %). Percent is useful in that most of the public is used to thinking in terms of percent, but statistical methods have been developed for proportions. A symbol for a sample proportion could be *P*:

$$P = \frac{f_i}{n}$$

E.g.: If 5 animals out of 50 have a disease $\Rightarrow$ (Proportion) Probability of the disease in this sample is P=5/50=0.1 (incidence of disease is 10%)

In statistical methods intended and used for categorical data, we can distinguish between empirical and theoretical counts (*in the terms of the sample and population*):

f_i – ***Empirical count*** (frequency) – ***observed*** in a sample (actually found)

$\hat{f}_i$ – ***Theoretical count*** (frequency) – theoretically ***expected*** in a population sampled (this theoretical count may be obtained in various ways in particular statistical methods for categorical data, e.g. according to some literature sources, from a long-term monitoring of the "event" under study in the past or by means of calculation from tables of empirical counts).

In the terms of a sample and population, we can also distinguish between:

Empirical proportion (probability) – in the sample: $P = \frac{f_i}{n}$

Theoretical proportion (exact probability) in the population (for N=∞): $\pi = \frac{\hat{f}_i}{N}$

Theoretical (population) proportion is symbolized π, in keeping with the convention of using Greek letters for population values and Latin letters for sample values. Population proportion $\boldsymbol{\pi}$ is a theoretical value, as we cannot calculate it in practice. We can only estimate its value by sample proportion ***P***. The exactness of the estimate is dependent (direct relation) on the sample size used for the estimate calculations; it holds generally that the larger is ***n***, the better is our estimation, i.e. sample estimate ***P*** is close to the true population $\boldsymbol{\pi}$).

9.1 Analysis of Categorical Data

For nominal (categorical) data we can only use categorical methods that are based on **frequencies** (counts) or **proportions** of "events" in statistical sets (we can't use any measured values and parameters like in methods used for numerical data). When dealing with counts in statistical methods for categorical data, we usually arrange them into tables of counts that allow us a better technique for all calculations used in the course of analysis of nominal data.

Nominal data analyses give us a possibility to assess:

A) **Difference between counts** in statistical sets:

- Sample vs. Population comparison,

- Sample1 vs. Sample 2 comparison.

B) **Relationships** between categorical data – "**Contingency tables"**

It is frequently desired to obtain a sample of nominal data and to infer whether the population from which it came conforms to a specified theoretical distribution. For example, a plant geneticist may rise 100 progeny from a cross that is hypothesized to result in a 3:1 phenotypic ratio of yellow-flowered to green-flowered plants. Perhaps a ratio of 84 yellow : 16 green is observed, although out of this total of 100 plants, the geneticist's hypothesis would predict a ratio of 75 yellow : 25 green. The question to be asked, then, is whether the observed frequencies (84 and 16) deviate significantly from the frequencies expected if the hypothesis were true (75 and 25).

The statistical procedure for attacking the question first involves the concise statement of the hypothesis to be tested. The hypothesis in this case is that the population which was sampled has a 3 : 1 ratio of yellow-flowered to green-flowered plants. This is referred to as a *null hypothesis* (abbreviated H_0), because it is a statement of "no difference"; in this instance, we are hypothesizing that the population flower colour ratio is not different from 3 : 1. If it is concluded that H_0 is false, then an *alternate hypothesis* (abbreviated H_A) will be assumed to be true. In this case, H_A would be that the population sampled has a flower-colour ratio which is *not* 3 yellow : 1 green. Recall that we state a null hypothesis and an alternate hypothesis for every statistical test performed, and all possible outcomes are accounted for by the two hypotheses.

The following calculation of a statistic called ***Chi-Square*** is used as a measure of how far a sample distribution deviates from a theoretical distribution. This *Chi-Square analysis* represents the basis of all calculations and techniques used for nominal data – we calculate test statistic of χ^2-**test** for differences between observed counts (in a sample) and those that would be expected theoretically (in all population).

The null hypothesis used in this Chi-Square test is $\mathbf{H_0}$**:** observed counts = expected counts ("no difference").

Calculation of test statistic Chi-square:

$$\chi^2 = \sum_{i=1}^{m} \frac{\left(f_i - \hat{f}_i\right)^2}{\hat{f}_i} \qquad \text{or} \qquad = \sum_{i=1}^{m} \frac{f_i^2}{\hat{f}_i} - n$$

Where:

f_i - observed frequencies *(in the sample class i),*

$\hat{f}_i$ - expected frequencies *(in the whole population class i),* i.e. frequencies expected in class *i* if the null hypothesis is true.

m – number of classes *(categories)* in the sample or population.

n – total nimber members in the sample

The summation is performed over all *m* categories of data; in the example with the flower-coloured ratio, there are two categories of data (i.e. m = 2): yellow-flowered plants and green-flowered plants. The expected frequency, $\hat{f}_i$, of each class is calculated by multiplying the total number of observations, *n*, by the proportion of the total that the null hypothesis predicts for the class. Therefore, for the two classes in the example, $\hat{f}_1 = 100.\frac{3}{4} = 75$ and $\hat{f}_2 = 100.\frac{1}{4} = 25$.

In the calculation of Chi-square test:

If calculated $\chi^2 = 0$ then the observed and theoretical frequencies are exactly identical.

The bigger is the value of calculated statistic χ^2 the bigger is the difference between observed and theoretical frequencies. Thus, this type of calculation is referred to as a measure of *goodness of fit ("goodness of fit test").*

If we compare calculated χ^2 with the *critical value* χ^2 for the specific ***α and DF: ν = m-1*** from the statistical tables , then:

- If $\chi^2 > \chi^2_{\alpha,\nu} \Rightarrow$ difference between the observed and expected counts is ***significant*** (at the level α). The null hypothesis is not true, i.e., sample distribution (empirical frequencies) *deviates* from a theoretical distribution (theoretical frequencies).
- If $\chi^2 \leq \chi^2_{\alpha,\nu} \Rightarrow$ difference between the observed and expected counts is ***not significant*** (at the level α). The null hypothesis is true, i.e., sample distribution (empirical frequencies) does *not deviate* from a theoretical distribution (theoretical frequencies).

In practice, **Chi-Square test is usually used** in testing for**:**

- *difference between observed frequencies of patients in a sample and the statistical probability of a disease in the population,*
- *difference between observed frequencies of patients in 2 (or more) samples (e.g. groups or herds of animals),*
- *contingency tables (analysis of relations between categorical data).*

9.2 Test for Difference between Empirical and Theoretical Counts

(Sample vs. Population)

In practice, this test is usually used in the situations when the theoretical probability for a studied "event" is known (e.g. predicted ratios in genetics, probabilities for the incidence of a particular disease according to literature sources, from a long-term monitoring of the "event" under study in the past, etc.). In the chi-square analysis we compare the expected frequencies (calculated from theoretical probabilities) with the empirical frequencies observed in the sample under our study in order to asses the statistical significance of differences between these counts.

Example:

From the total number of 146 calves in a sample 13 have enteritis. In the whole population the probability of this disease is 4.5%. Is the enteritis occurrence in the sample different from the whole population?

Method:

We can distinguish between **2 categories** in the sample and population:

- ill animals
- healthy animals

Sample: n = 146

Enteritis: 13

Population: π=0.045 (4.5%)

Empirical counts (observed in the sample):

f_1 13 (ill)

f_2 133 (healthy)

Theoretical counts (calculated):

$\hat{f}_1$: π.n = (0.045).146= 6.57 (ill)

$\hat{f}_2$: 146 - 6.57=139.43 (healthy)

Calculation of Chi-square test statistic:

$$\chi^2 = \frac{(13-6.57)^2}{6.57} + \frac{(133-139.43)^2}{139.43} = 6.5895$$

Degree of freedom: ν= m-1=1

Critical values from the tables of chi-square distribution:

$\chi^2_{crit.0.05}$ = 3.84

$\chi^2_{crit.0.01}$ = 6.63

$\chi^2 > \chi^2_{crit.0.05} \Rightarrow$ difference between empirical and theoretical counts is **significant** (at the $\alpha = 0.05$ level of significance).

Conclusion: There is a significantly higher proportion of ill animals in the sample than in the population ($P<0.05$).

(Empirical frequency of ill animals is 13, but theoretically it should be only 6.57 (to have the same probability like in the whole population).

9. 3 Test for Difference between 2(or more) Empirical Counts

(Sample1 vs. Sample2)

In most of situations in practice, we don't know theoretical probabilities or expected frequencies for the whole population – more often we have to compare 2 or more groups of empirical frequencies known from samples monitored, and have to decide whether these samples differ in their empirical frequencies.

In the Chi-Square analysis we work with 2 or more **groups** having several qualitative **classes** (unlike in the previous case, where there was only one group with 2 classes).

We mark: ***number of groups* as *r*** and frequencies in groups as f_i

number of classes* as *c and empirical counts in classes as f_j

In the course of the Chi-Square analysis, the empirical frequencies are usually arranged into a table. By means of the double subscript, f_{ij} refers to the frequency observed in **row *i*** *(group)* and **column *j*** *(class);* see the following example.

***Example*:**

The number of live- and dead born piglets was observed in 3 farms (A, B, C) in a region. We have to decide whether the frequencies of dead born piglets differ in the farms monitored. Frequencies obtained from 10 litters in each farm are summarized in the following table:

3 groups - rows (A, B, C) – in general ***r*,** (***i***)

2 classes - columns (live, dead) – in general ***c*,** (***j***)

r \ c	Live	Dead
A	**96**	**25**
B	**121**	**22**
C	**89**	**16**

Empirical frequencies in the table - f_{ij}

For the Chi-Square analysis we also need theoretical (expected) counts – we are able to calculate them from the *sums in rows and columns* in the table; thus the next step in the analysis is to ***sum empirical frequencies in rows and columns***:

r \ c	Live	Dead	Row Σ (R_i)
A	***96*** **(100.34)**	***25*** **(20.66)**	**121**
B	***121*** **(118.59)**	***22*** **(24.41)**	**143**
C	***89*** **(87.07)**	***16*** **(17.93)**	**105**
Col. Σ (C_j)	**306**	**63**	**369 (n)**

Then, we can calculate theoretical frequencies $\hat{f}_{ij}$ for each cell in the table.

Calculation formula for the theoretical frequencies in each table cell (row *i*, column *j*):

$$\hat{f}_{ij} = \frac{Ri \,.\, Cj}{n}$$

Where :

Ri = sum of empirical counts in row *i*

C_j = sum of empirical counts in column *j*

E.g. calculation of the theoretical frequency for the cell in the first row and the first column:

$$\hat{f}_{11} = \frac{121 . 306}{369} = 100,34$$

Calculation of the **Chi-square test statistic:**

$$\chi^2 = \sum_{i,j=1}^{6} \frac{f_{ij}^{\ 2}}{\hat{f}_{ij}} - n = \frac{96^2}{100,34} + \frac{25^2}{20,66} + \frac{121^2}{118,59} + \frac{22^2}{24,41} + \frac{89^2}{87,07} + \frac{16^2}{17,93} - 369 = 1,637$$

Degrees of freedom for the test (needed for the critical value χ^2 from the statistical tables) are calculated according the following formula:

Degree of freedom: $v = (r-1).(c-1) = 2$

We compare the calculated Chi-square statistic with the critical value:

$\chi^2_{\text{crit. 0,05}} = 5.99$

Conclusion:

$\chi^2 < \chi^2_{\text{crit. 0.05}} \Rightarrow$ Difference between the empirical and theoretical counts is **insignificant** (P>0.05).

It means that the farms monitored don't differ in the mortality of born piglets; i.e. frequencies of live - and dead born piglets don't differ among farms A, B, and C .

9. 4 Contingency Tables

(Analysis of relations between categorical data)

In many situations, nominal data for two variables may be tested for a hypothesis H_0: frequencies in categories of one variable are independent on the frequencies in the second variable. *E.g. whether the incidence (frequency) of a parasitic infection in dogs is the same in vaccinated as in the non-vaccinated individuals (= Does the incidence of parasites depend on the vaccination? Is the vaccination effective?)*

Observed data are arranged in a ***contingency table***:

- number of rows – ***r*** (categories of variable1: incidence of parasites)
- number of columns – ***c*** (categories of variable2: vaccination)

The null hypothesis H_0 *for this contingency table: frequencies in columns are independent on the frequencies in rows.*

According to the number of rows and columns we can distinguish between:

- Contingency **table *r* x *c*** *("r* by *c")*
- Contingency **table *2* x *2*** (special case of the table *r* x *c* for 2 categories in each variable)

9.4.1 Contingency table *r* x *c*

Method for the analysis of the contingency table *r* x *c* is the same as in the previous testing (*see the test for differences between empirical counts*) by means of Chi-square statistic calculation. We test the null hypothesis (independence of variables) through the testing for difference between empirical and theoretical counts in this contingency table *r* x *c*.

E.g. it can be a situation, when we are concerned with the question whether some diseases are associated with special breeds of cattle:
Variable 1 – Breeds of cattle (A, B, C)

Variable 2 – Diseases (1, 2)

Method:

1)We create the contingency table 3 x 2:

r \ *c*	**Disease 1**	**Disease 2**
Breed A	f_{11}	f_{12}
Breed B	f_{21}	f_{22}
Breed C	f_{31}	f_{32}

f_{ij} - Empirical frequencies

2) We calculate theoretical frequencies $\hat{f}_{ij}$ for all table cells (according to the known formula):

$$\hat{f}_{ij} = \frac{R_i \,.\, C_j}{n}$$

R_i – sums in rows

C_j – sums in columns

3) We calculate test statistic χ^2 (according to the known formula):

$$\chi^2 = \sum_{i=1}^{m} \frac{\left(f_i - \hat{f}_i\right)^2}{\hat{f}_i} \quad \text{or} \quad \chi^2 = \sum_{i,j} \frac{f_{ij}^{\,2}}{\hat{f}_{ij}} - n$$

4) Degrees of freedom: $\nu = (r-1)\,.\,(c-1)$

5) Conclusion:

If the calculated Chi-square statistic is small, there is a little dependence between the variables.

A large statistic indicates the positive dependence. If the critical value is exceeded, then the H_0 (independence) is rejected and dependence of variables monitored is statistically proved (In this case it would mean that there is some dependence between breeds (A, B, C) and monitored diseases).

9.4.2 Contingency Table 2 x 2

Contingency table 2 x 2 is a special case of the table *r* x *c* for 2 categories in each variable only. We can solve such table either in the same way like the previous table *r* x *c* or by means of the special (shortened) method.

Following situation can be solved by contingency table 2 x 2:

E.g.: Does the vaccination affect the incidence of parasites?

(Does the incidence of parasites depend on vaccination?)

Variable A – vaccine application

Variable B – incidence of parasites

We test the null hypothesis H_0: the incidence of parasites is not dependent on the vaccination.

Variable 1 – vaccine application (A- yes, A′-no)

Variable 2 – incidence of parasites (B-yes, B′-no)

Method:

1) We create the contingency table 2 x 2:

	B	B′	Row Σ
A	***a***	***b***	a + b
A′	***c***	***d***	c + d
Column Σ	a + c	b + d	n= a + b + c + d

a – Frequency of animals that have A and B (vaccinated animals that have parasites)

b – Frequency of animals that have A and B′ (vaccinated animals that have no parasites)

c – Frequency of animals that have A′ and B (not vaccinated animals that have parasites)

d – Frequency of animals that have A′ and B′ (not vaccinated animals that have no parasites)

(a, b, c, and d represent the empirical frequencies)

n – Total number of animals in the experiment

2) We calculate the test statistic χ^2 according to the formula:

$$\chi^2 = \frac{n.(a.d - b.c)^2}{(a+b).(c+d).(a+c).(b+d)}$$

3) Degree of freedom: $\nu = (r-1) \,.\, (c-1) = 1$

4) We compare the calculated Chi-square statistic with the critical value:

If $\chi^2 > \chi^2_{crit.} \Rightarrow$ H_0 (independence of A and B) is rejected

If $\chi^2 \leq \chi^2_{crit.} \Rightarrow$ H_0 (independence of A and B) is true

***Example*:**

In a sample of 50 dogs: 25 dogs got an experimental anti-parasitic substance

25 dogs did not get the substance

Does the substance affect the incidence of parasites in dogs?

	Without sub.	***With sub.***	**Total**
With parasites	**15**	**9**	**24**
Without parasites	**10**	**16**	**26**
Total	**25**	**25**	**50**

Test statistic: $\chi^2 = \frac{n.(15.16 - 9.10)^2}{(15+9).(10+16).(15+10).(9+16)} = 2.885$

$\nu = 1$

$\chi^2_{\text{crit. } 0.05} = 3.84$

Conclusion:

$\chi^2 < \chi^2_{\text{crit}}. \Rightarrow H_0$ is not rejected i.e. frequencies in columns are independent on the frequencies in rows ($P>0.05$).

(It means that the tested substance does not affect the incidence of parasites in dogs.)

Appendix

Statistical Tables

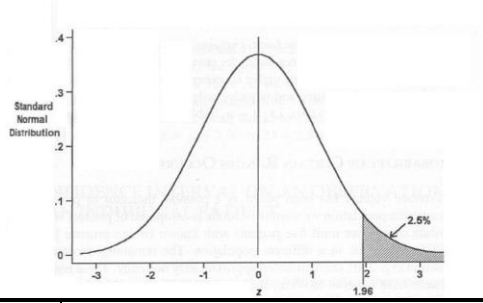
Source:

Riffenburgh, R. H.: Statistics in medicine. ACADEMIC PRESS, San Diego, USA 1999

Zar, J. H.: Biostatistical Analysis. Prentice Hall, Upper Saddle River, N.Y. 1999

List of Tables:

Appendix 1 Normal Distribution [1)]



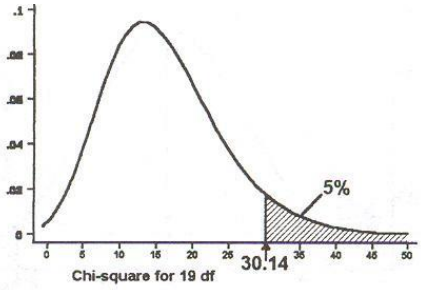
	One-tailed applications		Two-tailed applications	
z	One-tailed α (area in right tail)	1- α (area except right tail)	Two-tailed α (area in both tails)	1- α (area except both tails)
0	.500	.500	1.000	.000
.10	.460	.540	.920	.820
.20	.421	.579	.842	.158
.30	.382	.618	.764	.236
.40	.345	.655	.690	.310
.50	.308	.692	.619	.381
.60	274	.726	.548	.452
.70	.242	.758	.484	.516
.80	.212	.788	.424	.576
.90	.184	.816	.368	.632
1.00	.159	.841	.318	.682
1.10	.136	.864	.272	.728
1.20	.115	.885	.230	.770
1.281	.100	.900	.200	.800
1.30	.097	.903	.194	.806
1.40	.081	.919	.162	.838
1.50	.067	.933	.134	.866
1.60	.055	.945	.110	.890
1.645	.050	.950	.100	.900
1.70	.045	.955	.090	.910
1.80	.036	.964	.072	.928
1.90	.029	.971	.054	.946
1.960	.025	.975	.050	.950
2.00	.023	.977	.046	.934
2.10	.018	.982	.036	.964
2.20	.014	.986	.028	.972
2.30	.011	.989	.022	.978
2.326	.010	.990	.020	.980
2.40	.008	.992	.016	.084
2.50	.006	.994	.012	.088
2.576	.005	.995	.010	.990
2.60	.0047	.9953	.0094	.9906
2.70	.0035	.9965	.0070	.9930
2.80	.0026	.9974	.0052	.9948
2.90	.0019	.9981	.0038	.9962
3.00	.0013	.9987	.0026	.9974

1) For selected distances (*z*) to the right of the mean are given (a) one-tailed α, the area under the curve in the positive tail; (b) one-tailed 1- α, the area under all except the tail; (c) two-tailed α, the areas combined for both positive and negative tails; and (d) two-tailed 1- α, the area under all except the two tails. Entries for the most commonly used areas are italicized.

Appendix 2 Critical values for Student´s *t*-distribution.

DF = ν	α (1): 0.05	0.025	0.01	0.005	0.0025	0.001
	α (2): 0.10	0.05	0.02	0.01	0.005	0.002
1	6.314	12.706	31.821	63.657	127.321	318.309
2	2.920	4.303	6.965	9.925	14.089	22.327
3	2.353	3.182	4.541	5.841	7.453	10.215
4	2.132	2.776	3.747	4.604	5.598	7.173
5	2.015	2.571	3.365	4.032	4.773	5.893
6	1.943	2.447	3.143	3.707	4.317	5.208
7	1.895	2.365	2.998	3.499	4.029	4.785
8	1.860	2.306	2.896	3.355	3.833	4.501
9	1.833	2.262	2.821	3.250	3.690	4.297
10	1.812	2.228	2.764	3.169	3.581	4.144
11	1.796	2.201	2.718	3.106	3.497	4.025
12	1.782	2.179	2.681	3.055	3.428	3.930
13	1.771	2.160	2.650	3.012	3.372	3.852
14	1.761	2.145	2.624	2.977	3.326	3.787
15	1.753	2.131	2.602	2.947	3.286	3.733
16	1.746	2.120	2.583	2.921	3.252	3.686
17	1.740	2.110	2.567	2.898	3.222	3.646
18	1.734	2.101	2.552	2.878	3.197	3.610
19	1.729	2.093	2.539	2.861	3.174	3.579
20	1.725	2.086	2.528	2.845	3.153	3.552
21	1.721	2.080	2.518	2.831	3.135	3.527
22	1.717	2.074	2.508	2.819	3.119	3.505
23	1.714	2.069	2.500	2.807	3.104	3.485
24	1.711	2.064	2.492	2.797	3.091	3.467
25	1.708	2.060	2.485	2.787	3.078	3.450
26	1.706	2.056	2.479	2.779	3.067	3.435
27	1.703	2.052	2.473	2.771	3.057	3.421
28	1.701	2.048	2.467	2.763	3.047	3.408
29	1.699	2.045	2.462	2.756	3.038	3.396
30	1.697	2.042	2.457	2.750	3.030	3.385
31	1.696	2.040	2.453	2.744	3.022	3.375
32	1.694	2.037	2.449	2.738	3.015	3.365
33	1.692	2.035	2.445	2.733	3.008	3.356
34	1.691	2.032	2.441	2.728	3.002	3.348
35	1.690	2.030	2.438	2.724	2.996	3.340
36	1.688	2.028	2.434	2.719	2.990	3.333
37	1.687	2.026	2.431	2.715	2.985	3.326
38	1.686	2.024	2.429	2.712	2.980	3.319
39	1.685	2.023	2.426	2.708	2.976	3.313
40	1.684	2.021	2.423	2.704	2.971	3.307
41	1.683	2.020	2.421	2.701	2.967	3.301
42	1.682	2.018	2.418	2.698	2.963	3.296
43	1.681	2.017	2.416	2.695	2.959	3.291
44	1.680	2.015	2.414	2.692	2.956	3.286
45	1.679	2.014	2.412	2.690	2.952	3.281
46	1.679	2.013	2.410	2.687	2.949	3.277
47	1.678	2.012	2.408	2.685	2.946	3.273
48	1.677	2.011	2.407	2.682	2.943	3.269
49	1.677	2.010	2.405	2.680	2.940	3.265
50	1.676	2.009	2.403	2.678	2.937	3.261
52	1.675	2.007	2.400	2.674	2.932	3.255
54	1.674	2.005	2.397	2.670	2.927	3.248
56	1.673	2.003	2.395	2.667	2.923	3.242
58	1.672	2.002	2.392	2.663	2.918	3.237
60	1.671	2.000	2.390	2.660	2.915	3.232

Appendix 3 Critical values for χ^2 distribution, Right tail



DF =ν	α (area in right tail):				
	0.05	0.025	0.01	0.005	0.001
1	3.84	5.02	6.63	7.88	10.81
2	5.99	7.38	9.21	10.60	13.80
3	7.81	9.35	11.34	12.84	16.26
4	9.49	11.14	13.28	14.86	18.46
5	11.07	12.83	15.08	16.75	20.52
6	12.59	14.45	16.81	18.54	22.46
7	14.07	16.01	18.47	20.28	24.35
8	15.51	17.53	20.09	21.95	26.10
9	16.92	19.02	21.67	23.59	27.86
10	19.31	20.48	23.21	25.19	29.58
11	19.68	21.92	24.72	26.75	31.29
12	21.03	23.34	26.22	28.30	32.92
13	22.36	24.74	27.69	29.82	34.54
14	23.69	26.12	29.14	31.32	36.12
15	25.00	27.49	30.57	32.81	37.71
16	26.30	28.84	32.00	34.27	39.24
17	27.59	30.19	33.41	35.72	40.78
18	28.87	31.53	34.80	37.16	42.32
19	30.14	32.85	36.19	38.58	43.81
20	31.41	34.17	37.57	39.99	45.31
21	32.67	35.48	38.94	41.40	46.80
22	33.92	36.78	40.29	42.80	48.25
23	35.17	38.08	41.64	44.19	49.75
24	36.41	39.36	42.97	45.56	51.15
25	37.65	40.65	44.31	46.93	52.65
26	38.88	41.92	45.64	48.30	54.05
27	40.11	43.20	46.97	49.65	55.46
28	41.34	44.46	48.28	51.00	56.87
29	42.56	45.72	49.59	52.34	58.27
30	43.77	46.98	50.89	53.68	59.68
35	49.80	53.20	57.34	60.27	66.62
40	55.76	59.34	63.69	66.76	73.39
50	67.51	71.42	76.16	79.50	86.66
60	79.08	83.30	88.38	91.96	99.58
70	90.53	95.02	100.43	104.22	112.32
80	101.88	106.63	112.32	116.32	124.80
100	124.34	129.56	135.81	140.16	149.41

Appendix 4 Critical values for χ^2 distribution, Left tail

DF =ν	α (area in left tail):				
	0.001	0.005	0.01	0.025	0.05
1	.000016	.000039	.00016	.00098	.0039
2	.0020	.010	.020	.051	.10
3	.024	.072	.12	.22	.35
4	.091	.21	.30	.48	.71
5	.21	.41	.55	.83	1.15
6	.38	.68	.87	1.24	1.64
7	.60	.99	1.24	1.69	2.17
8	.86	1.34	1.65	2.18	2.73
9	1.15	1.73	2.09	2.70	3.33
10	1.48	2.16	2.56	3.25	3.94
11	1.83	2.60	3.05	3.82	4.57
12	2.21	3.07	3.57	4.40	5.23
13	2.61	3.57	4.11	5.01	5.89
14	3.04	4.07	4.66	5.63	6.57
15	3.48	4.60	5.23	6.26	7.26
16	3.94	5.14	5.81	6.91	7.96
17	4.42	5.70	6.41	7.56	8.67
18	4.90	6.26	7.01	8.23	9.39
19	5.41	6.84	7.63	8.91	10.12
20	5.92	7.43	8.26	9.59	10.85
21	6.45	8.03	8.90	10.28	11.59
22	6.99	8.64	9.54	10.98	12.34
23	7.54	9.26	10.20	11.69	13.09
24	809	9.89	10.86	12.40	13.85
25	8.66	10.52	11.52	13.12	14.61
26	9.23	11.16	12.20	13.84	15.38
27	9.80	11.81	12.88	14.57	16.15
28	10.39	12.46	13.57	15.31	16.93
29	10.99	13.13	14.25	16.05	17.71
30	11.58	13.79	14.95	16.79	18.49
35	14.68	17.19	18.51	20.57	22.46
40	17.93	20.71	22.16	24.43	26.51
50	24.68	27.99	29.71	32.36	34.76
60	31.73	35.53	37.49	40.48	43.19
70	39.02	43.28	45.44	48.76	51.74
80	46.49	51.17	53.54	57.15	60.39
100	61.92	67.32	70.07	74.22	77.93

Appendix 5 Critical values for Snedecor's F – test (two-tailed, $\alpha = 0.05$)

Denominator DF	Numerator DF								
	1	2	3	4	5	6	7	8	9
1	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28
2	38.506	39.000	39.165	39.248	39.298	39.331	39.355	39.373	39.387
3	17.443	16.044	15.439	15.101	14.885	14.735	14.624	14.540	14.473
4	12.218	10.649	9.979	9.605	9.365	9.197	9.074	8.980	8.905
5	10.007	8.434	7.764	7.388	7.146	6.978	6.853	6.757	6.681
6	8.813	7.260	6.599	6.227	5.988	5.820	5.696	5.600	5.523
7	8.073	6.542	5.890	5.523	5.285	5.119	4.995	4.899	4.823
8	7.571	6.060	5.416	5.053	4.817	4.652	4.529	4.433	4.357
9	7.209	5.715	5.078	4.718	4.484	4.320	4.197	4.102	4.026
10	6.937	5.456	4.826	4.468	4.236	4.072	3.950	3.855	3.779
11	6.724	5.256	4.630	4.275	4.044	3.881	3.759	3.664	3.588
12	6.554	5.096	4.474	4.121	3.891	3.728	3.607	3.512	3.436
13	6.414	4.965	4.347	3.996	3.767	3.604	3.483	3.388	3.312
14	6.298	4.857	4.242	3.892	3.663	3.501	3.380	3.285	3.209
15	6.200	4.765	4.153	3.804	3.576	3.415	3.293	3.199	3.123
16	6.115	4.687	4.077	3.729	3.502	3.341	3.219	3.125	3.049
17	6.042	4.619	4.011	3.665	3.438	3.277	3.156	3.061	2.985
18	5.978	4.560	3.954	3.608	3.382	3.221	3.100	3.005	2.929
19	5.922	4.508	3.903	3.559	3.333	3.172	3.051	2.956	2.880
20	5.872	4.461	3.859	3.515	3.289	3.128	3.007	2.913	2.837
21	5.827	4.420	3.819	3.475	3.250	3.090	2.969	2.874	2.798
22	5.786	4.383	3.783	3.440	3.215	3.055	2.934	2.839	2.763
23	5.750	4.349	3.751	3.408	3.184	3.023	2.902	2.808	2.731
24	5.717	4.319	3.721	3.379	3.155	2.995	2.874	2.779	2.703
25	5.686	4.291	3.694	3.353	3.129	2.969	2.848	2.753	2.677
26	5.659	4.266	3.670	3.329	3.105	2.945	2.824	2.729	2.653
27	5.633	4.242	3.647	3.307	3.083	2.923	2.802	2.707	2.631
28	5.610	4.221	3.626	3.286	3.063	2.903	2.782	2.687	2.611
29	5.588	4.201	3.607	3.267	3.044	2.884	2.763	2.669	2.592
30	5.568	4.182	3.589	3.250	3.027	2.867	2.746	2.651	2.575
40	5.424	4.051	3.463	3.126	2.904	2.744	2.624	2.529	2.452
60	5.286	3.925	3.343	3.008	2.786	2.627	2.507	2.412	2.334
120	5.152	3.805	3.227	2.894	2.674	2.515	2.395	2.299	2.222
∞	5.024	3.689	3.116	2.786	2.567	2.408	2.288	2.192	2.114

Appendix 6 Critical Values for Mann-Whitney *U*-test (2-tailed, $\alpha = 0.05$)

n_2 \ n_1	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
2					16	18	20	22	23	25	27	29	31	32	34	36	38
3		15	17	20	22	25	27	30	32	35	37	40	37	45	47	50	52
4	16	19	22	25	28	32	35	38	41	44	47	50	53	57	60	63	67
5		23	27	30	34	37	42	46	49	53	57	61	65	68	72	76	80
6			31	36	39	44	49	53	58	62	67	71	75	80	84	89	93
7				41	46	51	56	61	66	71	76	81	86	91	96	101	106
8					51	57	63	69	74	80	86	91	97	102	108	114	119
9						64	70	76	82	89	95	101	107	114	120	126	132
10							77	84	91	97	104	111	118	125	132	138	145
11								91	99	106	114	121	129	136	143	151	158
12									107	115	123	131	139	147	155	163	171
13										124	132	141	149	158	167	175	184
14											141	151	160	169	178	188	197
15												161	170	180	190	200	210
16													181	191	202	212	222
17														202	213	224	235
18															225	236	248
19																248	261
20																	273

Appendix 7 Critical values for Wilcoxon signed rank test

n	0.05	0.01	0.001
6	1	-	-
7	2	-	-
8	4	0	-
9	6	2	-
10	8	3	-
11	11	5	0
12	14	7	1
13	17	10	2
14	21	13	4
15	25	16	6
16	30	19	8
17	35	23	11
18	40	28	14
19	46	32	18
20	52	37	21
21	59	43	25
22	66	49	30
23	73	55	35
24	81	61	40
25	90	68	45
30	137	109	78
35	195	160	120
40	264	221	172
45	344	292	232
50	434	373	304

Appendix 8 Critical Values for the Spearman's Rank Correlation Coefficient r_S

n	α: 0.20	0.10	0.05	0.02	0.01	0.005	0.002
5	0.800	0.900	1.000	1.000			
6	0.657	0.829	0.886	0.943	1.000	1.000	
7	0.571	0.714	0.786	0.893	0.929	0.964	1.000
8	0.524	0.643	0.738	0.833	0.881	0.905	0.952
9	0.483	0.600	0.700	0.783	0.833	0.867	0.917
10	0.455	0.564	0.648	0.745	0.794	0.830	0.879
11	0.427	0.536	0.618	0.709	0.755	0.800	0.845
12	0.406	0.503	0.587	0.678	0.727	0.769	0.818
13	0.385	0.484	0.560	0.648	0.703	0.747	0.791
14	0.367	0.464	0.538	0.626	0.679	0.723	0.771
15	0.354	0.446	0.521	0.604	0.654	0.700	0.750
16	0.341	0.429	0.503	0.582	0.635	0.679	0.729
17	0.328	0.414	0.485	0.566	0.615	0.662	0.713
18	0.317	0.401	0.472	0.550	0.600	0.643	0.695
19	0.309	0.391	0.460	0.535	0.584	0.628	0.677
20	0.299	0.380	0.447	0.520	0.570	0.612	0.662
21	0.292	0.370	0.435	0.508	0.556	0.599	0.648
22	0.284	0.361	0.425	0.496	0.544	0.586	0.634
23	0.278	0.353	0.415	0.486	0.532	0.573	0.622
24	0.271	0.344	0.406	0.476	0.521	0.562	0.610
25	0.265	0.337	0.398	0.466	0.511	0.551	0.598
26	0.259	0.331	0.390	0.457	0.501	0.541	0.587
27	0.255	0.324	0.382	0.448	0.491	0.531	0.577
28	0.250	0.317	0.375	0.440	0.483	0.522	0.567
29	0.245	0.312	0.368	0.433	0.475	0.513	0.558
30	0.240	0.306	0.362	0.425	0.467	0.504	0.549
31	0.236	0.301	0.356	0.418	0.459	0.496	0.541
32	0.232	0.296	0.350	0.412	0.452	0.489	0.533
33	0.229	0.291	0.345	0.405	0.446	0.482	0.525
34	0.225	0.287	0.340	0.399	0.439	0.475	0.517
35	0.222	0.283	0.335	0.394	0.433	0.468	0.510
36	0.219	0.279	0.330	0.388	0.427	0.462	0.504
37	0.216	0.275	0.325	0.383	0.421	0.456	0.497
38	0.212	0.271	0.321	0.378	0.415	0.450	0.491
39	0.210	0.267	0.317	0.373	0.410	0.444	0.485
40	0.207	0.264	0.313	0.368	0.405	0.439	0.479
41	0.204	0.261	0.309	0.364	0.400	0.433	0.473
42	0.202	0.257	0.305	0.359	0.395	0.428	0.468
43	0.199	0.254	0.301	0.355	0.391	0.423	0.463
44	0.197	0.251	0.298	0.351	0.386	0.419	0.458
45	0.194	0.248	0.294	0.347	0.382	0.414	0.453
46	0.192	0.246	0.291	0.343	0.378	0.410	0.448
47	0.190	0.243	0.288	0.340	0.374	0.405	0.443
48	0.188	0.240	0.285	0.336	0.370	0.401	0.439
49	0.186	0.238	0.282	0.333	0.366	0.397	0.434
50	0.184	0.235	0.279	0.329	0.363	0.393	0.430
51	0.182	0.233	0.276	0.326	0.359	0.390	0.426
52	0.180	0.231	0.274	0.323	0.356	0.386	0.422
53	0.179	0.228	0.271	0.320	0.352	0.382	0.418
54	0.177	0.226	0.268	0.317	0.349	0.379	0.414
55	0.175	0.224	0.266	0.314	0.346	0.375	0.411

References

Armitage, P., Berry, G., Matthews, J.N.S.: Statistical methods in medical research. Blackwell Publishing, Oxford UK 2002, 814 p.

Ashcroft, S., Pereira, Ch.: Practical Statistics for the Biological Sciences. PALGRAVE MACMILLAN, New York, N.Y. 2003, 167 p.

Carvounis, Ch.: Handbook of Biostatistics. PARTHENON PUBLISHING, New York, USA 2000, 103 p.

Everitt, B. S.: Medical Statistics from A to Z. CAMBRIDGE University Press, Cambridge, UK 2003, 230 p.

Glantz, S. A.: Primer of Biostatistics. McGRAW-HILL, N.Y. USA 2005, 520 p.

Riffenburgh, R. H.: Statistics in medicine. ACADEMIC PRESS, San Diego, California USA 1999, 581 p.

Zar, J. H.: Biostatistical Analysis. Prentice Hall, Upper Saddle River, N.Y. 1999, 663 p.