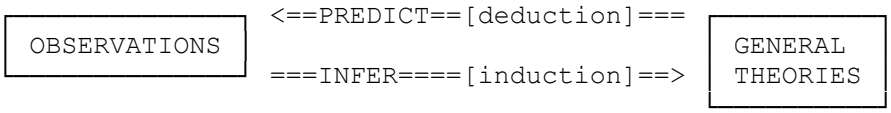
Preventive Medicine - 2nd year

3rd Medical Faculty

B i o s t a t i s t i c s

Statistics and statistical inference

Statistics may be defined as a discipline concerned with the treatment of information collected from groups of individuals (so called data). These individuals may be people, animals, other organisms, or administrative units. Statistics can be used for summarizing the data (descriptive statistics) but the major part of statistics involves drawing inferences from samples to a population in regard to some characteristics of interest. The reliability of such inferences may be objectively evaluated in terms of probability statements. Thus, statistics makes generalizations (inductive inferences) from the particular (sample) to the general (population).



Biostatistics

The tools of statistics are employed in many fields - business, education, psychology, agriculture, and economics, etc. When the data being analyzed are derived from the biological sciences and medicine, we use the term biostatistics.

Why do doctors need to know about statistics?

- Many doctors will need to do their own research at some time (during the work for higher degrees, or preparing manuscripts for publication) and/or consult a statistician
- One of the most important skills a physician should have is the ability to critically analyze original contributions to the medical literature; to check out the interpretation
- Even many GP's will participate in epidemiological studies as data collectors
- Modern computer programs make available to every physician a range of statistical techniques which, however, may be way beyond his/her statistical knowledge (Altman & Bland, 1991)

Individual and aggregated data

Clinical approach (= individual approach)

At any time, the physician makes use of the data of a single patient in his treatment. But every patient differs in important respects from every other patient.

Aggregated approach

Since the individuals greatly vary, no general information follows from a single observation. Statistical information (e.g. on mortality and morbidity experience) of groups of people (units) is needed and the data distribution has to be studied. Moreover, the preventive medicine makes an effort to improve the health status of the whole group of people (vaccination, work-place improvment, etc.).

Application of statistics to the data

STAGE	OPERATIONS
Study design	Planing the study; Definition of study subject; Computing required sample size; Formulation of hypotheses
Data processing	
Collecting	Recording of the observations
Transferring	Coding; transfer of data to the computer
Check-up	Checking; correcting; cleaning
Statistical analysis	
Organizing and summarizing	Categorization; derivation of new variables; tabulating; sorting
Statistical testing	Application of statistical methods
Interpretation	Explanation of computed results in medical terms
Presentation	Selecting suitable tables and graphs

Event - Variable

As a rule, some outcomes of interest (called events) are in the focus of the study. Most frequently, the (random) variable is a numerical expression of the quality or quantity of the event. (Whenever we determine the height, weight, or age, etc. of an individual, the result is referred to as a value of the respective variable). The exact value of the random variable cannot be exactly predicted in advance since it arises as a result of chance factors.

Types of variables

Qualitative (categorical):

binary - observations fall into one of two categories, i.e. 0-1 or yes-no.

nominal - several values of variable which cannot be ordered

ordinal - several values which can be ordered

Quantitative (numerical):

discrete - possible values are quite distinct and separated (usually counts)

continuous - can assume a continuous, uninterrupted range of values limited by the measurement precision only

Frequency

represents the number of individuals (observations) having a particular value (or being within a range) of the variable.

Absolute frequency - number of individuals

Relative frequency - number of individuals expressed as proportions or percentages of appropriate total (total number of individuals)

Cumulative frequency - makes sense only for variables whose values can be ordered. The cumulative frequencies are obtained by the process of successive cumulation of absolute or relative frequencies. The cumulative frequency for the variable value y_0 is a sum of frequencies for all variable values $y \leq y_0$.

Frequency distribution

is determined by numbers of observations at different values (or ranges, categories) of the variable. As a rule, it is presented as a table of relative frequencies. The distribution of outcomes is of primary interest of every study.

Histogram

is a 'bar chart' for diagrammatical presentation of the frequency distribution. The numbers are represented by the area of blocks (bars). Sometimes, the variable (especially the continuous one) has to be grouped in order to reduce the number of bars (class frequency distribution). The classes are then defined by their limits or midpoints. Generally, the areas of histogram bars (rather than the bar heights) are to be made proportional to the corresponding frequencies (when the bars are of the same width, even the heights of the bars are proportional to the frequencies).

Cumulative frequency

makes sense only for variables whose values can be ordered. The cumulative frequencies are obtained by the process of successive cumulation of absolute or relative frequencies. The cumulative frequency for the variable value y_0 is a sum of frequencies for all variable values $y \leq y_0$.

Types of graphs

scatter plots
line charts
pie charts
histograms
frequency polygons
bar charts
boxplots (box-and-whisker plot)
stem-and-leaf plot

Measures of central tendency (location)

(Arithmetic) mean

It is defined for quantitative variables only. It represents the centre of the frequency distribution histogram. The mean is simply the sum of the values divided by the number of values. This is calculated and denoted by 'x bar':

$$\bar{x} = (x_1 + x_2 + x_3 + \ldots + x_n)/n ,$$

where $x_1, x_2, ..$ denote the individual values of the variable and n is the number of the observations. For symmetrical distributions, the mean is close to median.

Percentile (quantile)

It is defined for quantitative and ordinal variables only. It is such a value for which the cumulative frequency equals the given proportion (percentage).

Median

It is the 50% percentile. If the observations are arranged in increasing order, the median lies in the middle of them.

Lower and upper quartile

These are 25% and 75% percentiles.

Mode

It is the value which occurs most often.

Geometric mean $\bar{x}_g$

It is defined for positive quantitative variables as the n-th root of the product of all the observations where n is the number of observations. Usually, it is obtained by calculating an arithmetic mean of data logarithms and taking the antilogarithm of the result. The geometric mean is always less than the arithmetic mean in value. Unlike the arithmetic mean, it is not overly influenced by the very large values in a positively skewed (nonsymmetric) distribution (deviations of large values from the centre are larger than those of the small ones), and so gives a much better representation of location in this situation.

Measures of variability (spread)

Variance

It is a measure of variation of the observations about the mean. The sample variance is computed as

$$s^2 = (\sum (x_i - \bar{x})^2)/(n-1).$$

Standard deviation (SD, σ, s)

The standard deviation is the square root of the variance. It is expressed in the same units as the mean rather then in the square of the units (as the variance is).

Range

Measure of variation. It is the difference between the maximum and minimum values.

Interquartile range

The distance between the upper and lower quartile.

Coefficient of variation

It is a dimensionless quantity (independent of the the units of observations) expressing the standard deviation as a proportion (or sometimes, as a percentage) of the mean:

$$V_x = s/\bar{x} \ .$$

It is a relative measure of data variation. If the data are evaluated in a logarithmic transformation using the natural logarithms (ln) then $V_x = s_{\ln(x)}$ approximately holds.

Other measures

Skewness

Measure of the asymmetry of the data distribution. If the points above the median tend to be farther from the median in absolute value than points below the median, a distribution is positively skewed (i.e. to the right). In that case, upper tail of a distribution is longer that the lower, and arithmetic mean will tend to be larger than the median. Similarly, the negatively skewed distribution has the tendency to stretch out to the left.

Kurtosis

Describes the proportions of observations found in the center and in the tails of the distribution. The extent to which the peak of frequency distribution departs from the shape of normal distribution, either being more pointed, or flatter.

Logarithmic transformation

When analyzing data whose distribution is positively skewed rather than symmetrical, it is recommended to transform the data to a logarithmic scale of measurement. The logarithmic transformation can be only applied to positive values. For variables expressed as concentrations (titres), weights, etc., the asymmetrical distribution is typical. In other situations, other transformations like reciprocal (1/x), or square-root are appropriate.

Probability

Probability of the occurrence of a particular event equals the proportion of times that the event would (or does) occur in a large number of similar repeated trials. It has a value between 0 and 1, equalling 0 if the event can never occur and 1 if it is certain to occur.

Probability distributions

The probability distribution of a random variable is a graph, table, or formula that specifies the probability associated with each possible value, or range, the random variable can assume.

Discrete probability distributions

Binomial distribution - A random event that has only two possible outcomes (say X and Y) that occur with fixed probabilities is referred to as a Bernoulli trial. Tossing a coin is an example. The probability distribution of the total number of X's occurring in a given number (n) of independent Bernoulli trials is called a binomial distribution. The examples of binary variables, for which the number (and proportion) of one outcome in a series of n realizations can be modelled by the binomial distribution, include survival status (alive - died), specimen positivity (positive - negative), sex of newborns (male - female).

Poisson distribution - This distribution is called the law of rare events and is appropriate for describing numbers of occurrences of an event during a period of time, provided that these events occur independently of each other and at random. The examples are numbers of particles found in a unit of space, numbers of radioactive emissions detected in a scintillation counter in 5 minutes, and some morbidity statistics, such as numbers of injuries or numbers of suicides.

Continuous probability distributions, density

The probability density function of the continuous random variable X is a curve such that the area under the curve between any two points a and b is equal to the probability that the value of X falls between a and b. Thus, the total area under the curve over the possible range of values for the variable is 1. The probability density function takes on high values in regions of high probabiliy and low values in regions of low probability.

Normal (Gauss) distribution - symmetrical distribution with a bell-shaped frequency distribution which is fully determined by the mean μ and the standard deviation σ. Linear measures, such as height or magnitude of measurement error, blood pressure and vital capacity and many other biological variables follow this distribution law.

Log-normal distribution - Many biological variables are distributed asymmetrically. In such situations, it is often useful to use the log-normal distribution, when the logarithms of the original values are normally distributed (see logarithmic transformation). Weights, survival time after radiation dose, triceps skinfold, minimum lethal dose in a homogeneous population of experimental animals and observations expressed as titres may be modelled by this distribution. It is fully determined by the geometric mean and the coefficient of variation.

Student's t-distribution - when individual observations come from an underlying normal

distribution with the mean μ and the standard deviation σ, the _ distribution of the quantity

$$(\bar{x}-\mu)/(s/\sqrt{n}),$$

where s is the estimated standard deviation and n is the sample size, is referred to as Student's distribution with n-1 degrees of freedom. The shape of the distribution highly depends on the number of degrees of freedom (here n-1), thus it is in fact the family of distributions rather than a unique distribution. As n gets larger the shape of the t-distribution approaches that of the normal distribution with zero mean and unique standard deviation. It is used in the procedure of the t-test.

Chi-square (χ^2) distribution - family of distributions (indexed by the degrees of freedom parameter) which describes the distribution of the sum sqaures of n values coming from normal distribution zero mean and unique standard deviation. In is particularly useful if the evaluation of categorical data in contingency tables.

Degrees of freedom

The quantity indexing the distributions, like t or chi-square. The term essentially means the number of independent units of information in a sample relevant to the estimation of a parameter or calculation of a measure (statistic). It is the number of independent comparisons that can be made between the members of a sample. In many cases, it is one or more less than sample size n.

Non-normal data - general approach

1. Transform (logarithm, square-root, reciprocal,etc.) original variable giving a new variable that behaves like a normal random variable.
2. Analysis is performed in terms of transformed variable.
3. Results of the analysis can be back-transformed into the natural measurement units for presentation
   The conclusions remain valid.
   OTHER POSSIBILITY: Non-parametrical methods

Sampling

It is usually the case in medicine, that we make a relatively small number of observations of some characteristic and want to make inferences about what we would have obtained if we had made very many more observations. The essential purpose of sampling is to gain information about the whole (the 'population') by examining only a part (the 'sample').
A population census, as carried out by the Central Statistical Office, is one of the rare occasions in which a complete population is measured.

Population

The term population is used to describe all possible observations of a particular variable or all units on which the observation could have been made. As

a rule, it is the group of individuals or inanimate objects to which the empirically observed frequency distribution is generalized. Population may be:

real - for instance in estimating some health indicators; population of hospitals; population of lung cancer deaths

hypothetical - when the population represents all possible ill persons (e.g. with coronary heart disease) with similar features as those in the experimental group. Then the generalization holds only for patients defined the same way.

finite - population consists of a fixed number of values of interest

infinite - population consists of an endless succession of values.

Sample

A subgroup of individuals in the population which is used to study the properties of the population by observing the values of variables. The reasons for studying samples are: Firstly, it is usually too expensive and time-consuming to study an entire population, secondly, sometimes it is not possible to define the population precisely (hypothetical population), thirdly, sufficiently sized representative sample can give an information concerning the population to whatever degree of accuracy is required.

Sample surveys

Simple random sample - every unit of the population has the same chance of being included. Such samples are said to be representative (meaning that no particular block of population is more likely to be represented than any other).

Stratified random sample - the population is divided into groups, or strata, on the basis of certain characteristics (age, sex, etc.). A simple random sample is selected form each stratum and the results for each stratum are combined.

Multistage random sample - sampling is organized stage by stage using random sample based on different sample units in each stage. Stages are logically successive, e.g. region - municipality - house, or primary school - child.

Cluster sample - a simple random sample of groups (e.g. families) is chosen, and everyone in the chosen group is studied.

Systematic sample - the method is based on a list of population. One starts at random somewhere among the first N members of the list. Then, every N-th individual is chosen.

Quota sample - obviously only co-operative individuals are chosen to fill certain quotas of individuals in well-defined groups, such as males aged 25 to 34 years. One is free to choose anyone who will fit the requirements but there is no guarantee of representativeness and randomness.

Estimation

A 'true' value of characteristic of interest in the population (so called 'parameter') is usually regarded as a fixed number, but its value is generally unknown. A 'statistic' is a numerical characteristic of a sample which is used to estimate the population parameter. However, the value of a statistic will usually vary from sample to sample.

Standard error of the mean (S.E.)

The sample mean is the so called point estimate of the population mean μ and it is unlikely to be exactly equal to μ. Imagine collecting large number of independent samples of the same size and calculating the sample mean of each of them. The mean of this distribution would be the population mean, and it can be shown that its standard deviation, which is called the standard error of the sample mean, would be

$$s_{\bar{x}} = s/\sqrt{n} \ .$$

It measures how precisely the population mean is estimated by the sample mean.

Confidence interval (CI)

In generalizing the sample distribution to the population, the observed distribution is taken as the best estimate of the population distribution. This is true not only for distributions, but also for their characteristics made for quantitative data. These are called distribution parameters in the population, and generally are not identical to the corresponding sample characteristics. For any parameter, it is possible to construct the so-called confidence interval, which covers the unknown value of the parameter with a given probability (=confidence). Usually, it is 0.95 or 0.99, generally it is $1-\alpha$, where α is small.

The limits of the $(1-\alpha).100\%$ confidence interval for the mean are constructed as follows:

$$\bar{x} \pm t.s_{\bar{x}} \ ,$$

where t is the tabulated value for the Student distribution corresponding to probability α and n-1 degree of freedom. The confidence interval for the geometric mean is constructed similarly, except the data are replaced by their logarithms. The antilogarithms of the obtained limits are the confidence limits for the geometric mean.

The confidence interval for the relative index (expressed as a percentage) is constructed as follows:

$$p \pm u.(p(1-p)/n)^{½} \ ,$$

where u is the argument of the Gauss distribution function corresponding to the probability $1-\alpha$ (for $1-\alpha = 0.95$, u equals 1.96).

Statistical model

The statistical model usually deals with the structure and distribution in the population and its corresponding data attributes in the sample.

The model involves firstly assumptions about the distribution of the variables in the population, particularly in case of the hypothetical population. Secondly, the assumptions about the regression relationship, i.e. to what extent and how the mean (of the dependent variable) is conditioned by covariates (independent variables) or classification variables. Any statistical analysis is determined by the statistical model and its validity depends on the adequate fit of the abstract statistical model with the structure and distribution of the observed data.

Statistical test - hypothesis testing

The statistical test verifies a hypothesis about the distribution, parameters of distributions for quantitative variables, and generally about statistical models. The statistical (significance) test is a rule for deciding whether any particular sample is in the 'likely' or 'unlikely' class of possible models of interest. One of the assumption that specifies values of the parameters of the population is called the null hypothesis (H_0). Other possible values of the parameters form the so-called alternative hypothesis (H_A). The test to be used in any situation will depend on what alternative to the null hypothesis are considered. The one-sided alternative is used when the effect under study is supposed in only one direction. The two-sided alternative is appropriate when the direction of effect cannot be determined beforehand. The statistical test evaluates the probability corresponding the observed data under the assumption that the null hypothesis is true. Sample values always differ somewhat, and the question is whether the differences among samples signify genuine population differences or whether they represent merely chance variations such as to be expected among several random samples from the same population.

If the observed data are unlikely under the null hypothesis, the hypothesis is rejected. The limit for the above low probability is given in advance and is called the significance level α. The most frequent are $\alpha=0.05$ or $\alpha=0.01$, i.e. 5% or 1%. The hypothesis is then rejected on 5% or 1% level of significance.

p-value

Sometimes in statistical results (especially those obtained by computer software), the so-called p-value is given. This value means the probability that a result, such as the one obtained, could have occurred if the null hypothesis were true. The smaller the p-value, the less likely is the observed

statistic (or greater) when the null hypothesis is true, and the stronger the evidence for rejection of the hypothesis. When the p-value is less than the decision criteria (the significance level α), the null hypothesis is rejected.

In other words, p-value represents the likelihood that the result observed in the data, or one more extreme, is due to chance, given that the null hypothesis is true.

General structure of a significance test

1. Formulate the problem and determine the type of the test (two-sample, one-sample, paired)
2. State the null hypothesis H_0 and the alternative hypothesis H_A (one-sided, or two-sided)
3. Set the significance level α
4. Parametric or nonparametric test?
5. Calculate the test statistic
6. Find the critical values of the distribution of the test statistic.
7. If the test statistic lies in the critical region, reject H_0 and accept H_A, otherwise do not reject H_0.

6. Find the p-value corresponding to the test statistic.
7. If the p-value is less than the significance level, reject H_0 and accept H_A, otherwise do not reject H_0.

One-sample Student t-test about the mean

The null hypothesis is $\mu = C$. The sample mean $\bar{x}$ and standard deviation s are calculated from the data. The equality is tested by the so-called t-statistic:

$$t = \frac{|\bar{x} - C|}{s} \sqrt{n} ,$$

where n is the number of observations. The t-statistic is compared to the tabulated critical value for the Student distribution corresponding to probability α and n-1 degrees of freedom. If the t value is greater, the null hypothesis is rejected at the 100.α % significance level.

Two-sample Student t-test about means (equal variances)

It is used for comparison of two independent groups. The population means in the first and second groups are denoted by μ_x and μ_y, respectively. The variances in both populations are assumed to be the same σ^2. The null hypothesis is $\mu_x - \mu_y = 0$. The test is based on t-statistic:

$$t = \frac{|\bar{x} - \bar{y}|}{s} \left(\frac{n_x n_y}{n_x + n_y}\right) ,$$

where $\bar{x}$ and $\bar{y}$ are sample means in groups, n_x and n_y the corresponding numbers of observations and s is the pooled sample standard deviation calculated by:

$$s^2 = \frac{\sum(x_i-\bar{x})^2 + \sum(y_i-\bar{y})^2}{n_x + n_y - 2} .$$

The t-statistic is compared to the tabulated critical value of the Student distribution corresponding α and n_x+n_y-2 degrees of freedom. If the t value is greater than the critical one, the difference of means is statistically significant at level α (the null hypothesis is rejected).

Two-sided and one-sided Student t-test

The above t-tests verifies the null hypothesis against the two-sided alternative that the means are not equal. If the difference is large enough, the hypothesis is rejected whatever is the sense of the difference. This is the two-sided test. However, in practical situations a different alternative is to be tested. for instance, if one of the studied groups is exposed to some agent, it is reasonable to assume the one-sided alternative, i.e. the mean of the exposed group may be be different only in one sense. This is the one-sided test and the obtained t-statistic is compared to the tabulated critical value corresponding to probability 2α. The significance level of the test is naturally α.

Two-sample Student t-test about means (unequal variances)

Is is used instead of the above test if the variances in both groups differ (by F-test). The calculation proceeds with means $\bar{x}$ and $\bar{y}$ in both groups, variances s_x^2 and s_y^2 and variances of the means $s_{\bar{x}}^2$ and $s_{\bar{y}}^2$ (squares of standard errors of the mean)

The test is based on the quantity

$$t = \frac{|\bar{x} - \bar{y}|}{(s_{\bar{x}}^2 + s_{\bar{y}}^2)} .$$

The resulting value of t is compared to the tabulated critical value of the Student t-distribution corresponding to probability α and to degrees of freedom given by

$$DF = \frac{(s_{\bar{x}}^2 + s_{\bar{y}}^2)^2}{\frac{s_{\bar{x}}^4}{n_x-1} + \frac{s_{\bar{y}}^4}{n_y-1}}$$

truncated to its integer part.

Paired t-test

Suppose the values of x and y are measured either on the same individuals, or on the pairs of individuals, who are very similar or are under the same conditions. In such a situation, the comparison is based on pair differences

$$d_i = x_i - y_i \ .$$

The one-sample t-test comparing the mean with zero is applied to these differences:

$$t = \frac{\bar{d}}{s_{\bar{d}}} \sqrt{n} \ .$$

When t exceeds the tabulated critical value of the Student t-distribution corresponding to probability α and n-1 degrees of freedom, the average difference significantly differs from zero on $100\alpha\%$ level of significance.

F-test for comparison of two variances

In many situations, the equality of two variances is studied. So, $\sigma_x^2 = \sigma_y^2$ is the null hypothesis H_0.

The variances are compared by considering the ratio of the two sample variances

$$F = \frac{s_x^2}{s_y^2} \ .$$

The obtained value of F is compared to the tabulated critical value of the Fisher-Snedecor F-distribution with n_x-1 and n_y-1 degrees of freedom. If the calculated F value is larger than the critical one, the hypothesis is rejected and the variances significantly differ.

The F-test is used in

(1) a simultaneous comparison of more than two means ANOVA;

(2) testing the equality of variances in two-sample t-test.

Two-dimensional frequency distribution

If two variables are measured on each individual,they may either be treated separately, or their simultaneous distribution may be examined (the so called two-dimensional frequency distribution). In the case of qualitative variables, the data can be expressed in the form of contingency table, see below. For quantitative variables, the two-dimensional distribution is characterized by means and standard deviations of both variables and by the correlation coefficient measuring the strength of linear relationship (association) between them.

Correlation coefficient

Correlation coefficient is given by the formula

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

The correlation coefficient is always a value between -1 and 1. When r equals 1 or -1 the value of one

variable is exactly determined as a linear function of the second one. If both variables tend to be high or low correspondingly, r is positive and the greater its value the closest the association, i.e. if x is greater than the mean, y is likely to be greater than the mean, and vice versa. If one variable decreases as the other increases, r is negative. If there is no relationship at all between the two variables, they are said to be uncorrelated or linearly independent and r will be 0.

Regression

The linear regression gives the equation of the straight line (called regression line) that best describes how the dependent variable y increases (or decreases) with an increase of the independent (explanatory) variable x and enables the prediction of one variable from the other. In the population, this linear relationship is supposed to be given by the equation

$$\mu(x) = \alpha + \beta x \ .$$

The equation of the regression line in the sample is

$$y = a + bx \ ,$$

where

$a = \bar{y} - b\bar{x}$ is the intercept and

$$b = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$$ is the slope (regression coefficient).

The values of a and b are calculated by the least square fit.

Residual variance (of the points about the regression line)

It is a measure of data variability about the regression line. The sample residual variance is calculated by the formula

$$s^2 = \frac{\sum(y_i - a - b\,x_i)^2}{n-2} \ ,$$

where the denominator is often referred to as the residual degrees of freedom.

Standard error of the regression coefficient b

Like the mean, the regression coefficient is subject to sampling variation and its precision is measured by its standard error

$$s_b = s/\sqrt{(\sum(x_i-\bar{x})^2)} \ ,$$

where s is the square root of the residual variance with n-2 degrees of freedom.

Residual standard deviation (about regression line)

It is the square root of the residual variance.

Analysis of variance (ANOVA)

A method for testing the null hypothesis that several samples come from the same populations or from identical populations with respect to means. The estimates of variability between groups and within groups of the data are used to achieve the comparison of means. Depending on the number of qualitative variables (factors) which define the groups, one-, two-, or three-way ANOVA is used. One-way analysis of variance is an extension of the two-sample t-test giving the same results when there are only two groups. Comparing more than two groups in pairs by t-tests is not quite a correct technique.

Multivariate analysis

Multivariate methods investigate how a dependent variable is related to more than one explanatory variable, or how several dependent variables vary together. Some examples are multiple regression, analysis of covariance, multiple logistic regression, log-linear models, factor analysis, cluster analysis, logrank methods.

Chi-square (χ^2) test for contingency table or proportions

The frequency distributions for a categorical variable in two (or more) groups is compared. The standard χ^2 (chi-square) test for a 2 x m table is a general test to assess whether there are differences among the m proportions, i.e. the null hypothesis is tested, stating that the frequency distribution of individuals in the population is the same in all compared groups. The data are arranged into the so called contingency table (2 compared groups, m categories of the variable):

Group	Category 1	2	...	m	
1	n_{11}	n_{12}	...	n_{1m}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2m}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	...	$n_{.m}$	$n_{..}$

Under the null hypothesis, the relative frequencies would be alike in both rows and would not differ so much from that in the last row of the table. For each cell of the table (with indices ij), the expected value e_{ij} is calculated. It is determined by the proportion in a given column of the last summary row:

$$e_{ij} = \frac{n_{i.} n_{.j}}{n_{..}} .$$

The test of the hypothesis consists in the comparison of both distributions expressed by the actually

observed (n_{ij}) and expected hypothetical values (e_{ij}). We use the criterion

$$\chi^2 = \sum_{ij} \frac{(n_{ij}-e_{ij})^2}{e_{ij}} \quad .$$

The chi-square test can also be applied to larger tables consisting generally of r rows and c columns. The null hypothesis is that the distributions of the qualitative variable in the different populations are the same or, equivalently, that two qualitative variables in a single sample are not associated. In every case, obtained χ^2 value is compared to the tabulated critical value of the chi-square distribution corresponding to probability α and to (r-1)(c-1) degrees of freedom. If the critical value is exceeded by obtained χ^2 value (or, equivalently, if the p-value is small), it is concluded that there is a statistically significant difference at chosen significance level (e.g. 5%) between the groups in the proportions that are compared.

Fourfold table (2-by-2 table, 2 x 2 table)

With only two categories and two groups, the resulting table reduces to four cells and is called fourfold. The test formula is

$$\chi^2 = \frac{(n_{11}n_{22} - n_{12}n_{21})^2 \, n_{..}}{n_{1.}n_{2.}n_{.1}n_{.2}} \, ,$$

corresponding to one degree of freedom.

Comparison of paired proportions - McNemar's test

Chi-square tests for proportion mentioned above assumes that the groups being compared are independent. If a sample is studied twice - separated by a certain interval of time or under different conditions - then we are dealing with dependent samples; we have two trials on the same individuals. For the general layout

		STUDY 2 +	STUDY 2 -
STUDY 1	+	a	b
	-	c	d

the null hypothesis that the frequencies in the population do not differ for the two studies is tested by McNemar χ^2 based on 1 degree of freedom:

$$\chi^2 = (b - c)^2 / (b + c) \, .$$

Nonparametric tests

Many statistical procedures make assumptions about the distributions of variables in the populations studied. For example, an underlying assumption of normally distributed data is required for a valid application of the t-test (though with a large sample size violations of this assumption could be tolerated). The significance tests which make distributional assumptions about the variable being analyzed are called the ***parametric tests***. In some situations, the assumptions are not satisfied or, it is not possible to check whether such assumptions are true, particularly when the sample size tends to be small, and there is often doubt about whether or not a particular test is valid for a set of quantitative data. In such situations, the assumption-free tests should be used. Such tests are called ***nonparametric, distribution-free*** or ***rank tests***.

The nonparametric equivalent of the t-test is Wilcoxon (Mann-Whitney) test; the equivalent of classical correlation coefficient is Spearman rank correlation coefficient.

NONPARAMETRIC METHOD	USE	PARAMETRIC EQUIVALENT
Wilcoxon signed rank test	Test of difference between paired observations	Paired t-test
Wilcoxon rank sum test (Mann-Whitney test)	Comparison of two groups	Two-sample t-test
Kruskall-Wallis one-way analysis of variance	Comparison of several groups	One-way analysis of variance
Friedman two-way analysis of variance	Comparison of groups, defined by their values on two variables	Two-way analysis of variance
Spearman's rank correlation coeff.	Measure of association between two variables	Correlation coefficient

Statistical software

The statistical package integrates together a broad set of programs covering many different statistical methods, graphics and data management.

Full-range statistical analysis systems

BMDP, SAS, SOLO, SPSS, Stata, Statgraphics, Statistica, SYSTAT, S-Plus, GLIM

Epidemiological and special purpose statistical programs

EpiInfo - epidemiological and public health system
EGRET - epidemiological package focused on regression
EPICURE - special types of regression for various types of studies; calculations based on person-years

S t a t i s t i c s i n E p i d e m i o l o g y

Epidemiology

may be defined as the study and quantification of the occurrence of illness in groups of people. It deals with evaluation of hypotheses about the causation of illness and it tries to relate disease occurrence to characteristics of people and their environment.

Rates

Vital statistics measures are commonly expressed in the form of rates because absolute numbers (of some event in the population) are not very informative because of different size of compared populations. A rate is a proportion involving a numerator and a denominator, which is connected to an element of time. The rate of occurrence of an event in a population is the number of events which occur during a specified time interval, divided by the total number of observation time accumulated during that interval. For convenience, it is multiplied by 1000 (or 100, etc.) to avoid decimal numbers. In contrast, the numerator of measures of risk refers to a different group from the denominator.

crude rate - presented for the entire population; summary measure calculated by dividing the total number of cases of the outcome in the population by the total number of individuals in that population in a specified time period.

category-specific rates - presented for categories of the population defined on the basis of particular characteristics such as age, sex, or race

adjusted (standardized) rates - single statistically constructed summary rates that account for the difference between the populations with respect to some other variables. When comparing rates adjusted for a particular factor, any remaining observed differences between groups cannot be attributed to confounding by that variable.

Measures of disease frequency - prevalence and incidence

The prevalence quantifies the proportion of individuals in a population who have the disease of interest at a specific time, and it provides an estimate of the probability (risk) that an individual will be ill at a particular point in time. The prevalence may be measured either at a single point in time (point prevalence) or over a period of time (period prevalence). In contrast to the prevalence, the incidence quantifies the number of new cases of a disease that develop in the population during a specified time interval. Both measures are usually expressed as a percentage or as per 1000 population.

$$\text{(Point) prevalence rate} = \frac{\text{No of persons with disease at a particular point in time}}{\text{total population}}$$

$$\text{Period prevalence rate} = \frac{\text{total No of persons with disease at some time during specified period}}{\text{total population at mid-point of interval}}$$

$$\text{(Cumulative) incidence} = \frac{\text{No of new cases of disease in a specified period of time}}{\text{total population at risk}}$$

$$\text{Incidence density} = \frac{\text{No of new cases of disease in a specified period of time}}{\text{total person-time of observation}}$$

To summarize, we can say that the incidence is an interval measure measuring the number of ***new cases*** of a disease during a specified period of time related to the number of ***persons at risk*** of contracting the disease; while the prevalence is a point measure based on the total number of ***existing cases*** among the ***whole population***. Prevalence and incidence are the principal measures of morbidity (relative frequency of illness).

Standardization

Since both the risk of dying and the risk of contracting most diseases are related to age, and since they often differ for the two sexes, the crude mortality rate and overall incidence and prevalence rates depend critically on the age-sex composition of the population. For example, a relatively older population would have a higher crude mortality rate than a younger population, even if, age for age, the mortalities were the same. It is therefore misleading to use these overall rates when comparing two different populations and the overall standardized rates must be used. An age-standardized rate is the theoretical rate which would have occurred if the observed age-specific rates applied in a reference population: this population is commonly referred to as the "standard population". Standardized rates are obtained using a standard population with two basic methods:

direct method - the age-sex specific rates from each of the populations under study are applied to a standard population to give age-sex adjusted mortality (morbidity) rates. Age-sex specific rates for the study populations and age-sex composition for the standard population are required. The standard could be the distribution of one of the populations to be compared, two populations combined, or an outside standard of interest (such as the European or World standards).

indirect method - the age-sex specific rates from a standard population are applied to each of the populations of interest to give standardized mortality (or morbidity) ratios (SMR), which in turn may give the adjusted rates. It is convenient to think of this method in terms of a comparison between observed and expected numbers of cases. Age-sex composition + total deaths for the study populations and age-specific rates + overall rate for standard population are required. The standard population should be as similar as possible to the studied on with respect to other risk factors.
Standardized mortality ratio (SMR) measures how much more (or less) likely a person is to die in the study population compared to someone of the same age and sex in the standard population. A value of 1 means that they are equally to die.

$$SMR = \frac{\text{observed number of deaths}}{\text{expected number of deaths if the age-sex specific rates were the same as those of the standard population}}$$

Confidence intervals for age-standardized rates can be computed which are rather easier to interpret.

	S T A N D A R D I Z A T I O N	
	DIRECT	INDIRECT
DATA REQUIRED Study population(s)	Age-sex specific rates	Age-sex composition + total deaths (or cases)
Standard population	Age-sex composition	Age-sex specific rates (+ overall rate)
METHOD	Study rates applied to standard population	Standard rates applied to study population
RESULT	Age-adjusted rate	Standardized mortality (morbidity) ratio (+ age-sex adjusted ratio)

Person-time units

In research studies, not only may mortality be measured over the instantaneous periods such as 1 year, but also people may be followed for differing lengths of time due to various reasons. In this subject-time approach, the number of deaths observed is therefore related to the total number of person-years (person-months, etc.) of observation, rather than to a mid-population. Note that one person-year of observation may result from one person being observed for a whole year or, for example, from 12 persons being observed for just one month each.

Epidemiologic studies

cross-sectional - It is carried out at just one point in time or over a short period of time. That is why it is suitable for measuring prevalence but not incidence. Cross-sectional studies are useful in showing associations (especially between stable indicators such as blood groups), in providing early clues of aetiology.

case-control study - It is used to investigate the association between a certain factor and a particular disease. The sampling is carried out according to the disease status rather than the exposure one. A group of individuals identified as having disease (the cases) is compared with a group not having the disease (the controls). Case-control studies are useful for rare diseases or conditions, or when the disease takes a very long time to become manifest.

cohort study - In this type of study, a group or groups of individuals are defined on the basis of presence o absence of exposure to a suspected risk factor for a disease. When exposure status is defined, all potential subjects must be free from the disease under investigation, and eligible participants are then often followed up over a period of time to assess the occurrence of that outcome. Cohort studies are useful for providing stronger evidence of causality, and less subject to biases due to errors of recall or measurement.

clinical trial - It is an prospective experiment carried out to assess the effectiveness of a new treatment regime. Clinical trials are prospective experimental studies used for providing the most rigorous evidence of causality.

Case-control study and cohort study with count denominators

In epidemiologic research, the calculations are often based on two-dimensional distribution with first variable measuring the presence of disease or other harm and the second measuring the exposition of individuals. Results are presented in the form of the fourfold (2 x 2) table.

	Disease (Cases)	No disease (Controls)	Total
Exposed (Factor present)	a	b	a+b
Non-exposed (Factor absent)	c	d	c+d
Total	a+c	b+d	N = a+b+c+d

Each of the cell represents the number of individuals having the particular combination of exposure and disease status:

a = the number of individuals who are exposed and have the disease
b = the number of who are exposed and do not have the disease
c = the number of who are not exposed and have the disease
d = the number of who are both nonexposed and nondiseased

The margins of the table represent:
a+b = the total number of individuals exposed
c+d = the total number of nonexposed
a+c = the total number with the disease
b+d = the total number without the disease

Odds:
a/b = odds in exposed
c/d = odds in nonexposed

Cohort study with person-time denominators

	Cases	Person-time units
Exposed	a	PY_1
Non-exposed	c	PY_0
Total	a+c	$PY_1 + PY_0$

Rates:
a/PY_1 = rate in exposed
c/PY_0 = rate in nonexposed

Epidemiologic measures of risk from the exposure

In this section, several measures of association used to assess the strength of the relationship between a risk factor and the subsequent occurrence of disease will be presented.

The ***risk*** of a disease (or of death) is the number of events occurring in a specific period of time divided by the total number of persons alive at the start of the period. It can be calculated in both exposed and non-exposed group as a/(a+b) and c/(c+d), respectively. The ***relative risk*** is defined as the ratio of these two quantities. It estimates the magnitude of an association between exposure and disease in a cohort study and indicates the likelihood of developing the disease in the exposed group relative to those who are not exposed.

A relative risk of 1.0 indicates that the incidence rates of disease in the exposed and nonexposed groups are identical and thus that there is no association observed between the exposure and the disease in the data. A value greater than 1.0 indicates a positive association., or an increased risk among those exposed to a factor.

The formula cannot be applied to data from a case-control study, since the participants are there selected on the basis of disease status and the value of RR strongly depends on the proportions of cases and controls. The relative risk can be estimated, however, by means of the odds ratio.

The ***odds*** ("chance") of disease to non-disease equals the total number of cases divided by those still at risk at the end of the study. The odds among the exposed is a/b and among non-exposed c/d. Their ratio is called ***odds ratio***.

Relative risk (cohort studies with counts)

$$RR = \frac{a / (a+b)}{c / (c+d)}$$

Rate ratio/relative risk (cohort studies with person-time units)

$$RR = \frac{a / PY_1}{c / PY_0}$$

Odds ratio

$$OR = \frac{a / b}{c / d} = \frac{a \cdot d}{b \cdot c}$$

In the cohort study, the ***attributable risk*** provides an information about the absolute effect of the exposure or the excess risk of disease in those exposed compared to those nonexposed. It is calculated as the difference between the two risks.

Attributable risk

$$AR = \frac{a}{a+b} - \frac{c}{c+d}$$

Attributable fraction is used to estimate the proportion of the disease among the exposed that is attributable to the exposure, or the proportion of the disease in that group that could be prevented by eliminating the exposure.

Attributable fraction (Aetiologic fraction)

$$AF = \frac{RR - 1}{RR} \quad \text{(cohort)}$$

$$AF = \frac{OR - 1}{OR} \quad \text{(case-control)}$$

Confidence intervals for RR and OR

There are several methods for calculating the asymmetrical confidence limits for OR and RR. On the basis of logarithmic transformation, we can calculate the error factor EF

for case-control or cohort study with count denominators

$$EF = \exp \{ 1.96 \sqrt{(1/a + 1/b + 1/c + 1/d)} \}$$

for cohort study with person-time denominators

$$EF = \exp \{ 1.96 \sqrt{(1/a + 1/c)} \}$$

95% confidence interval is formed by

lower limit	OR / EF
upper limit	OR * EF .

Alternatively, 95% confidence limits for both OR and RR can be calculated according to Miettinen (test-based approach)as

$$RR^{(1\pm1.96/\chi)} \qquad OR^{(1\pm1.96/\chi)},$$

where χ is a square-root of χ^2 obtained in the fourfold table. When the confidence interval does not include 1.0 (i.e. the expected value if there is no association), the associated p-value is small and this is the indication of statistically significant association between the disease and risk factor.

Bias

It may be defined as any factor or process which tends to produce erroneous results or conclusions that differ systematically from the truth. Many of the biases can arise in the designing, executing, analyzing, and interpreting of the study. Some of them are avoidable, but a badly run study cannot be rescued by statistical manipulation.

On the other hand, precision of a measurement relates to the amount of random variation about a fixed point (be it the true value [unbiased], or not [biased]).

Main bias types

- selection bias
- confounding bias
- information bias
- recall bias
- diagnostic bias

Confounders (from the Latin *confundere*, to mix together)

Confounding variable is one that distorts an existing real relationship by the fact of being related to both disease and exposure (or to both response variable and treatment assignment). A confounder can make it appear that there is a relationship present where there is none, or, at the other extreme, hide a real relationship.. Confounding can be controlled by using an appropriate study design, or using special statistical techniques (e.g.

Mantel-Haenszel, matching, multivariate analysis, logistic regression). Unless it is possible to adjust for confounders, their effects cannot be distinguished from those factors being studied. A confounder, when properly controlled for, can explain away an apparent association between the treatment and the response.

Confounding factor must
- be a risk factor for the disease among nonexposed
- be associated with exposure
- not be an intermediate step in the causal path between exposure and disease

Mantel-Haenszel technique, stratification

When confounding is present, it is important to analyze relevant subsets (e.g. age groups) of the data separately. Stratification is a technique that involves the evaluation of the association within homogeneous categories, or strata, of the confounding variable (e.g. age). It is often useful, however, to apply a summary test which pools the evidence from the individual subsets but which takes into account the confounding factor(s). When the data are stratified into several 2 x 2 tables, the Mantel-Haenszel χ^2 is used to test the null hypothesis of no overall relationship in a series of such a tables.

The Mantel-Haenszel technique is used to calculate an overall summary measure of association (like Mantel-Haenszel summary relative risk or odds ratio):

crude OR = ad / bc

Mantel-Haenszel OR = (ad/N) / (bc/N)

Logistic regression

A form of regression analysis used when the reponse variable is a binary variable (denoting e.g. whether or not a person develops a given disease). The probalility p that a person develops the health outcome is modelled as a function of explanatory variables of interest in the form $\ln \frac{p}{1-p} = \alpha + \beta_1 x_1 + ... + \beta_q x_q$, where $x_1,...,x_q$ is a set of q explanatory variables (which can include exposure variables and those used for adjustment, e.g. age). Parameters β_i, i=1,..,q are called regression coefficients. For binary variables, the $\exp(\beta_i)$ is the odds ratio for the presence of the health outcome. For continuous variables, the $\exp(\beta_i)$ is OR corresponding to unit change in x_i.

Diagnostic tests - sensitivity and specificity

A good diagnostic or screening test is one which iproves your guess about the patient's disease staus over the guess you would make based on just the general prevalence of disease.

		TRUE CONDITION: Disease +	No disease -	Total
DIAGNOSTIC TEST	+	a	b	a+b [all testing positive]
	-	c	d	c+d [all testing negative]
Total		a+c [all diseased]	b+d [all non-diseased]	N = a+b+c+d [total population]

Tests: T+ positive T- negative
Diseased: D+ yes D- no

Sensitivity - the proportion of diseased persons the test classifies as positive

= a/(a+c) = P[T+/D+] (probability of positive test,given disease)

Specificity - the proportion of non-diseased persons the test classifies as negative

= d/(b+d) = P[T-/D-] (probability of negative test, given no disease)

False-positive rate - the proportion of non-diseased persons the test classifies (incorrectly) as positive

= b/(b+d) = P[T+/D-]

False-negative rate - the proportion of diseased persons the test classifies (incorrectly) as negative

= c/(a+c) = P[T-/D+]

Predictive value of a positive test - the proportion of positive positive tests that correctly identify diseased persons (who really have the disease)

= a/(a+b) = P[D+/T+]

Predictive value of a negative test - the proportion of negative tests that correctly identify non-diseased persons

= d/(c+d) = P[D-/T-]

Accuracy of the test - the proportion of all tests which are correct classifications

= (a+d)/(a+b+c+d)

Sensitivity and specificity are characteristics of the test itself. Of primary interest to a clinician,however, are the predictive values of a positive test(PV+) and of a negative test(PV-).Predictive values are very much influenced by how common the disease is. Thus, the practical value of a giagnostic test is dependent on combination of sensitivity, specificity and disease prevalence, all of which determine the predictive values. Nevertheless, a good test should have a high predictive value, even though prevalence of the disease is low.

Hypotheses and types of errors

		DECISION ON BASIS OF SAMPLE (CONCLUSION OF SIGNIFICANCE TEST)	
		No Effect DO NOT REJECT H_0	Effect REJECT H_0 (ACCEPT H_A)
TRUE STATE OF NATURE (REALITY)	No Effect H_0 true	No Error Probability = 1-α *(contingency coeff.)*	Type I Error Probability = α *(significance level)*
	Effect H_0 true H_A false	Type II Error Probability = β	No Error Probability = 1-β *(power)*

Sample size determination

It is the mathematical process of deciding, before a study begins, how many subjects should be studied in order to reach the desired precision. That would depend on how large a difference we think is important. The factors to be taken into account include:

for cohort study:

- confidence level (the probability that if the two ***samples*** differ this reflects a true difference in the two ***populations***);
- desired power (the probability that if the two ***populations*** differ, the two ***samples*** will show a "significant" difference - it is the ability of a study to demonstrate an association if one exists);
- ratio of number of unexposed to number of exposed;
- expected frequency of disease in unexposed group (attack rate);
- odds ratio or relative risk worth detecting (the value of OR or RR closest to 1 which is to be detected as significantly different from 1)

for case-control study:

- confidence level;
- desired power;
- ratio of controls per case;
- expected frequency of exposure among controls;
- OR worth detecting

Computed sample sizes must be increased to allow for nonresponse, tabulation of more than one question, etc.

Index