„TÁMOP-4.1.2/A/1-11/1-2011-0015 Egészségügyi Ügyvitelszervező Szakirány: Tartalomfejlesztés és Elektronikus Tananyagfejlesztés a BSc képzés keretében”

ÚJ SZÉCHENYI TERV

Biostatistic

e-Book

Dr. Dinya Elek

Semmelweis Egyetem
Cím: 1085. Budapest, Üllői út 26.
Telefon: +36 (1) 459-1500
E-mail: hirek@semmelweis-univ.hu
Honlap: http://semmelweis-egyetem.hu

MAGYARORSZÁG MEGÚJUL
A projektek az Európai Unió támogatásával valósulnak meg.

Table of contents

Introduction

What are we going to learn about?

Figures do not lie: statistics is a scientific method and practice dealing with systematic collection and analysis of data of facts (reality) expressed in figures. Means of gaining information is mathematical statistics. In medical science the term "Biostatistics" is used.

Carl Friedrich Gauss (1777-1855)

What skills are needed in the field of Biostatistics?

For example, one of the most important data distribution (density function) the normal distribution. Generally the normal distribution is called Gauss-distribution or sometimes, due to its shape, bell-curve. The shape of the function:

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

The fact, that *x* probability variable follows normal distribution is identified as follows:

$$X \sim \mathcal{N}(m, \sigma^2).$$

Specially, if x *N*(0, 1) form, then *x* is termed standard distribution. The root of word *statistics* is Latin, *status* (state); the German Gottfried Achenwall was the first to use it for the analysis of data pertaining to state activities.

By now the meaning of the word *statistics* have significantly been expanded and it basically means the inferential statistics (based on mathematical knowledge, mainly probability theory).

Gottfried Achenwall (1719-1772)

How shell I put them into the practice?

Biostatistics is an applied science used to process and analyze our data. In the word of computers we hardly do any calculations by hand as excellent statistical analyses softwares are at our computers. One of the many great options is the Excel software program because either its functions or its statistic module is able to perform routine-like tasks.

Augustus De Morgan (1806–1871)

Analysis steps:

- The use of appropriate variable (according to the protocol!).
- Use of universal data format (e.g.: Excel).
- Descriptive statistical reports.
- Scatter plot diagram.
- Use of simple, effective and practical comparative methods. Use of complicated methods is not always appropriate.
- Use of validated statistical software packages.

Conclusions and inferences:

- Type I-II. errors checking.
- Performing multiple comparisons.

- One hypothesis: only one response.
- Is it the issue of overestimation vs. underestimation?
- Is it the issue of statistical vs. clinical significance?
- Is the result relevant professionally (clinically)?

Adrien-Marie Legendre (1752-1833)

I may have heard about it

Please read the below study and give your answers!

In a clinical trial of 12 COPD patients the effect of a new bronchodilator was studied for the Forced Expiratory Volume (FEV_1 %). In each patient the value was measured before the treatment and 10 minutes after the treatment, too.

Can the affect of the preparation be decided by mere inspection of the data?

a) No.

b) Yes.

c) I am uncertain about the answer.

Is biostatistics necessary for the objective investigation of the problem?

a) No.

b) Yes.

c) I am uncertain about the answer.

When does biostatistics help?

a) If I am in possession of appropriate knowledge of biostatistics.

b) The computer solves tasks even without me, all I have to do is to put the data in.

c) It is enough to test the new preparation even on one patient, no need of involving several individuals into the trial and biostatics is not needed either.

1. Probability theory

The module is to check the correctness of the most important concepts of probability calculations.

Abraham de Moivre (1667-1754)

Probability calculation is one of the very important disciplines of mathematics. it deals with the investigation of possible outcome of repeated experiments (mass phenomenon), it determines the chances of their occurance. It gives help to making correct decision and it is an indispensable means in Biostatistics, too.

We tell, on the basis of p value (p = the symbol of probability), if a treatment is effective during the clinical trial or ineffective.

Jakob Bernoulli (1654-1705)

Its major branches are: the classic probability calculation; mathematical statistics (Biostatistics is one of its disciplines); the theory of stochastic processes and the information theory.

The word "probability" was used first by Jacob Bernoulli, a Swiss mathematician , in his work "The Art of Guessing, (1713)".

In the XX. century it was Kolmogorov who axiomatised probability theory and incorporated it into the measurement theory.

2. Data distributions

The module is to check the properties of the most important discrete and continuous distributions.

Andrey Kolmogorov (1903-1987)

The task of probability calculation is to explore the types of data distribution, recognize their characteristics and to describe them. The mathematical statistics uses and relies heavily on this knowledge.

It is important to recognize our data distribution, because it facilitates the task solutions, if we know what distribution our investigational data come from.

The properties of the known distributions (e.g.: Gauss distribution) can be well circuscribed, this knowledge is worth acquisition so that we should be able solve statistical problems with success.

Blaise Pascal (1623-1662)

3. Data types

The module is to check the properties of every single learnt scale type by the given examples.

Thomas Bayes (1701-1761)

Our data fall into four data types (data scales) knowing, which data scale the tested random variable comes from, is essential in terms of the biostatistical analysis as each comparative statistical method is developed fro a certain type of data. The data can be divided into data of categorical (classifying) and non-categorical (quantitative) character. The categorical data can be nominal (named) and ordinal (organiser).

The quantitative data are of continuous (eventually) discrete character, which are often termed metric data.

Data with arbitrary chosen (interval scale) 0-point are distinguished from the ones for which the multiplicative arithmetic operations can be applied, too. (the have ratio scale, with fixed 0-point).

Data scales were published by Stevens Stanley Smith (1946) in his article "On the Theory of Scales of Measurement" Science 103 (2684): 677–680 pgg.

Stanley Smith Stevens (1906–1973)

4. Data reduction

The module is to check your knowledge in terms of data reduction.

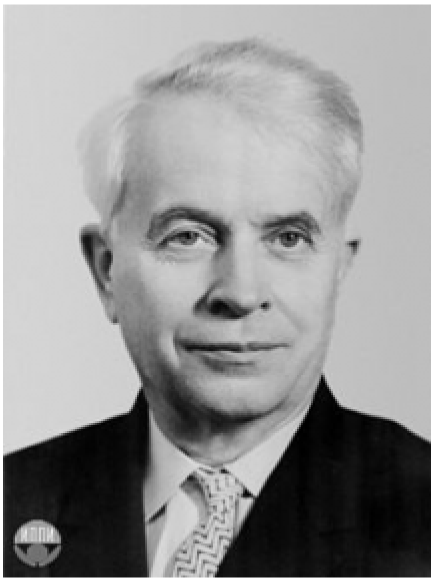
Siméon Poisson (1781-1840)

The data reduction procedure is (relatively simple calculation) when the properties of our data are compressed into a single data (figure).

The knowledge and interpretation of these data (e.g. mean value, standard deviation) is important because it will help to know the behavior of data, to provide useful information for further analysis. In fact, it is a transformation performed for the shake of data size to be processed.

Data reduction in the broader sense may mean:

- aggregation,
- attribute subset selection,
- dimensionality reduction,
- size reduction (alternative representation, modeling).



Andrey Markov (1856-1922)

5. Confidence interval

The module is to check the knowledge in terms of confidence interval.

Pierre-Simon de Laplace (1749-1827)

The confidence interval (CI) or reliability interval or probability interval of the calculated parameters is the interval prediction, which at a given probability level (1-α) provides the bands of the given parameters (e.g. mean), the population value falls in.

The confidence interval - at a given level of significance - is the lower and upper limit of the predicted variables. To be able to calculate CI different assumptions have to be used, e.g. assuming that the distribution of the errors of the prediction is normal. At a given reliability level, a lesser confidence interval indicates a better estimation. This depends on several factors like the size of the sample (N).

Pafnutyij Lvovics Csebisov (1821-1894)

Interval estimates can be contrasted with the point estimates. A point estimate is a single value estimate for a given parameter (e.g. expected value) which is probably close to this value. The generalisation of the confidence interval is a reliability interval of several dimensions.

This is appropriate not only to predict the error of assessment but also to show that a parameter could not be estimated correctly. It the estimate of a parameter is not estimated correctly enough, the estimates for the rest of the parameters are incorrecz, too..

6. Power-analysis

The module is to check the facts learnt about power-analysis (prediction of sample size).

Gerolamo Cardano (1501-1576)

The task of power-analysis is to help determine the necessary and sufficient (minimum) number of samples to achieve the desired accuracy. The "sufficient" is underlined, because the larger number of samples will dramatically increase the costs of testing and duration of the research, which may raise ethical issues in clinical trials.

The aim is to find the minimum size of sample at which the desired effects can safely be proved.

I course of experiment, one of the most important task is to achieve accuracy, reliability and harmony between the available financial sources and the time.

Pierre de Fermat (1601-1665)

When planning clinical trials, one of the basic steps is to determine the minimally necessary number of observations.

When the sample size increases, the confidence interval will be decrease, so the statistical tests have been able to detect significantly lower, existing real differences.

The same holds for the standard deviation, too, which result from the homogeneity of test conditions and from improving the research protocols. Methods for determining the required sample size are generally based on the normal(Student's) distribution.

Methods for determining the required sample size are generally based on the normal (Student's t) distribution.

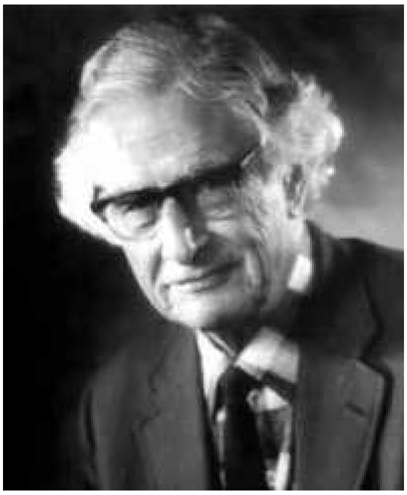
7. t-test

The module is to check the knowledge in terms of relatively simple parameter statistics (t-tests, z-test).

William Sealy Gosset (1876-1937)

It is a common problem that we would like to know whether the mean of two normal distributions is the same.

For example: whether the data measured before treatment are the same as the data obtained after the treatment, i.e. whether the treatment significantly affected the value of the variable under consideration. t-test or z-test is used to compare two data sets, of course under different conditions.



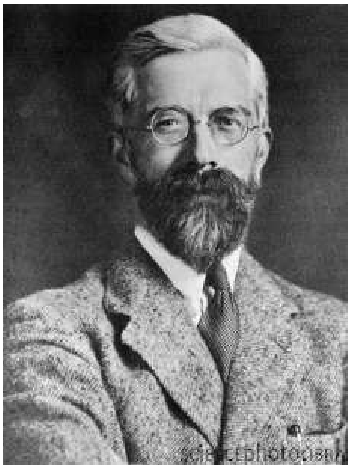
Frank Yates (1902-1994)

It is also true not only for the comparative statistical tests used in this chapter but also in general: every statistical procedure (or test) has a condition system of usage, which the investigational data have to meet, otherwise the conclusions will not be reliable!

It is just like this e.g. in terms of the two sample test: the normality of the variables, homogeneity of variances and the independence of the groups.

8. ANOVA

The module is to check the knowledge in terms of use of ANOVA (ANalysis Of VAriance).



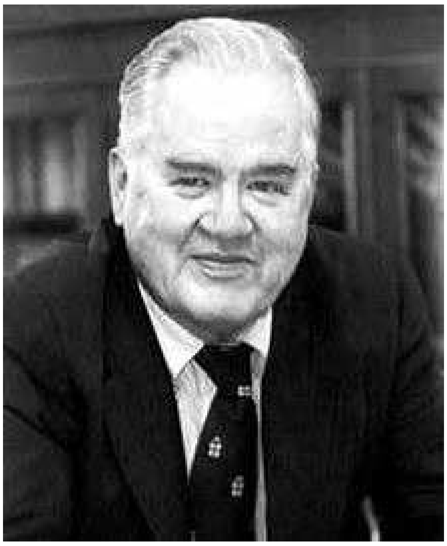
Sir Ronald Aylmer Fisher (1890-1962)

Variance analysis (ANOVA) is a statistical method for comparing the means of more than two groups with homogenous standard deviation and the variables come from normal distributions.

It is one of the most important methods: it investigates whether differences between mean values in the single groups are significant or not. The method can also be regarded as the generalization of t-tests with two samples.

It compares the different mean values of population to each other by variances: starting from the sum of variances of the total data quantity; trying to find the answer if the differences in standard deviations between the groups are accidental or are caused by the effect of an other explanatory factor (e.g.: drug).

It divides the sum of variances into two groups: varinace between the groups and variance within the groups, and compares them by means of F-test. The variance within the groups is in denominator.

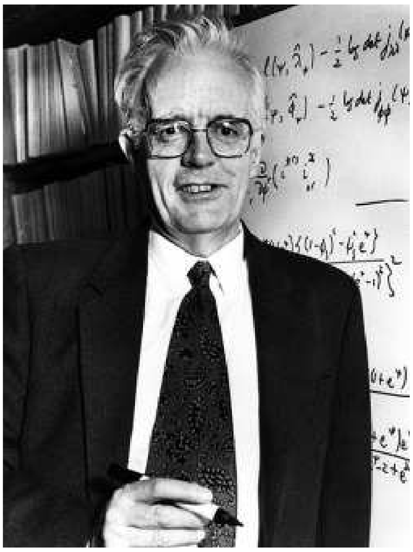


John Wilder Tukey (1915-2000)

If the analysis reveals that not all the means are equal between the treatment groups, then the group-pairs differing significantly are screened by an appropriate post-hoc test (postanalysis). Several methods are known for screening, such as Tukey's method, which usually gives good results. Classification of ANOVA methods: according to the nunber of test criteria (one-way and more than two-way) and according to the independence of the samples. The repeated measures design uses the same subjects with every condition of the trial, including the control group, too. A popular repeated measures design is the crossover study.

9. Rank-statistics

The module is to check the knowledge in terms of nonparametric statistics.



Sir David R. Cox (1924-)

The common feature of these methods is that it does not assume a specific distribution of the data (as opposed to parametric methods e.g. normality). That is way these methods are -in broader sense- called distribution free methods, as well. Such a process is called rank-statistics. The power (reliability) of nonparametric test is less than that of the parametric test.

What does the rank transformation means? All data are arranged in ascending order (regardless of which group they belong to), the data are indexed by natural numbers (serial numbers or rank order numbers are given), and instead of the data, the rank numbers are used in the following calculations.

If you find two or more of the same data, they will be replaced by the average of the rank numbers. The ranking numbers thus obtained are separated according to the original groups. This transformation, naturally expresses the original observations in ordinal scale.

George Edward Pelham Box (1919-2013)

If there is no difference (i.e. H0 is met) between the position mean values (median), then both groups will have observations of low and high rank number, and the mean rank values are nearly identical, too.If H0 is rejected, the average rank number will be very likely higher than in the other. If there are many values of the same rank (tied or liked ranks), which is not favourable, and this time, the test slightly underestimates the significance level.

10. Correlation

The module is to check the knowledge in terms of correlation calculation.

Karl Pearson (1857-1936)

The power, direction, the degree of dependence of relationship between the variables of population, is determined by the correlation coefficient defined as a parameter.

The linear correlation coefficient has two important features: 1) In case of independent variables, this coefficient is 0 and in this case the connection is uncorrelated 2) In case of variables in linear function (non-stochastic) variables, the maximal values of correlation coefficients are [-1,1] interval.

In general statistical use, the correlation indicates that the two values are not independent of each other as they are adjusted to the type of data.

Do you remember what is the main features of normal probability variables?

It is characteristic of the normally distributed random variables if they are uncorrelated, they are independent, too. Thus, the correlation may be used for measuring the power of relationship between measurable quantities of normal distribution.

The rank correlation coefficients measure if the two sets change together. If one sequence increases, the other decreases, the rank correlations will be negative.

Charles Spearman (1863-1945)

Several rank correlations are known: the Spearman's rank and Kendall's rank correlations are most commonly used. Similarly to the linear correlation, their values fall into the [-1,1] interval.

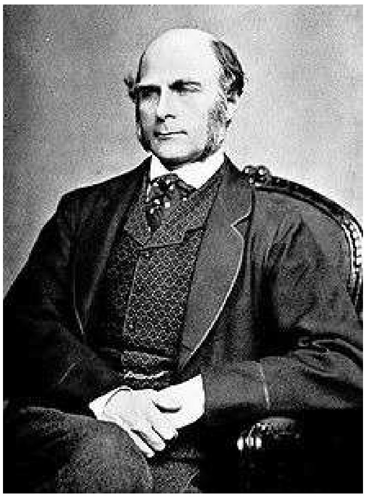
Their values:

- +1, if the two rankings are the same,
- 0, if they are independent of each other,
- -1, if their rankings are another's reversal.

The rank correlations are less distribution sensitive alternatives of the linear correlation coefficient.

11. Linear and nonlinear regression

The module is to check the knowledge in terms of regression calculations.



Sir Francis Galton (1822-1911)

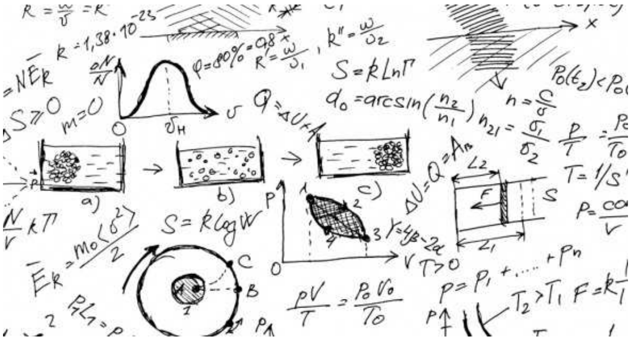
If there is a connection between two (or more) variables, we often want to predict (calculate) one value from the other in form of a function. The simplest regression relationship between two variables is the linear function relationship.

What does it mean in the practice?

This mean that, in course of estimating the linear regression, we try to fit a straight line to the scatter plot of sampling data.

When performing estimation of linear regression we get two coefficients: a (intercept or constant) and b (slope and covariance between two variables) values whiches are estimated from the sample in a way that e.g. the average squared error is minimalized. The simplest and most common method of estimation is the least squares fitting.

The term "constant" can be omitted from the model, which usually results in poorer fitting.



The investigation of functional relation of problems of several linear or e.g. nonlinear problems.

The simple linear regression model may be generalized so that, instead of one, n explanatory variables are involved in the model, and their effect on y is estimated. Thus, we also have the opportunity to incorporate the nonlinearity of the explanatory variables into the model.

Abraham Wald (1902-1950)

The multiple linear regression model allows for explanatory variables that they should be correlated with each other (partial effect).

12. Contingency table statistics

The module is to check the knowledge in terms of contingency table statistics.

Harald Cramér (1983-1985)

If one of the two properties of the basic multitude can be characterized by r-value, the other by k-value discrete random variable then, their combined behaviour can be described by a frequency table consisting of r number of rows and k number of columns (r x k field contingency table). In addition to general use (independence, homogeneity, fit testing) it plays an important role in the epidemiology as well. E.g. if there is no relationship between the variables, if the two variables are independent, then the differences will be close to 0, so the chi-square value will be close 0, too. If chi-square value is far from being 0, there is a high probability that the variables are not independent.

In these types of chi-square tests, the null hypothesis means the independence of variables. In general, the calculation of the degrees of freedom is: df=(rows-1) x (columns-1).

The chi-square test can be performed:

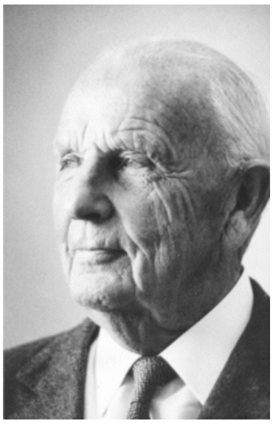
- 1) If N (the sample size) is large enough, will be greater than 30. 2) The expected values in each cell will not be 0. 3) The cell number of the expected values between 1 and 5 can only make up 20% of total cell numbers.

At for example four-field table cell contents should be greater than 5 in each cell.

Alan Agresti Distinguished Professor Emeritus

13. Hipothesis testing

The module is to check the knowledge in terms of hypothesis testing.



Ernst Hjalmar Waloddi Weibull (1887-1979)

Hypothesis: it is a statement on a tested population distribution or any of its parameters. It is a proposed explanation of a phenomenon which still has to be rigorously tested. It may be: a simple e.g. equality or a composite (sum of multiple hypotheses). Purpose: to verify (based on a given sample) the hypothesis for the investigated population.

The investigation of a hypothesis is a statistical decision: i.e. rejection of a conclusion or the H0 hypothesis in favor of an alternative (or counter) hypothesis H1 or acceptance of H0. It is true for H0 and H1: in general the hypothesis is a simple statement; they mutually exclude each other.

The following concepts are of major importance:

1. Statistical test: the process of deciding on the acceptance or rejection of H0 on the basis of the sample.

2. Test function: is a function of sample elements with a known distribution if we suppose that the H0 hypothesis is true. Its role is to help make a decision which hypothesis should be accepted.

3. Significance level: is the probability of falling of a trial function into the so called critical range (e.g. in case of normal distribution, in some of the ranges of 2.5%). According to the situation of the critical range left, right and double side significance levels are distinguishe

Summary

1. The right decision can only be made based on data of good quality!
2. Data intended for analysis should always be checked even before analysis!
3. Correctness of our decision always involves a certain degree of probability!
4. In making decision possibly all circumstances should be considered!
5. Statistical significance is not the same as clinical significance!
6. Ask the question at the end of the analysis: is the result significant clinically? Always decide on the basis of this as to the utility of the result!

MAGYARORSZÁG MEGÚJUL

A projektek az Európai Unió támogatásával valósulnak meg.