Nina Trukhacheva

Modern Methods of Statistical Analysis of Medical Data

λογος

Nina Trukhacheva

Modern Methods of Statistical Analysis of Medical Data

Logos Verlag Berlin

λογος

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de .

ISBN 978-3-8325-3818-7

Logos Verlag Berlin GmbH
Comeniushof, Gubener Str. 47,
10243 Berlin
Tel.: +49 (0)30 42 85 10 90
Fax: +49 (0)30 42 85 10 92
INTERNET: http://www.logos-verlag.de

CONTENT

Chapter 6 THE APPLICATION OF THE STATISTICA SOFTWARE PACKAGE FOR DEPENDENCES ANALYSIS

Chapter 7 THE APPLICATION OF STATISTICA SOFTWARE PACKAGE FOR MULTIVARIANCE ANALYSIS

Nina Trukhacheva

Modern Methods of Statistical Analysis of Medical Data

Logos Verlag Berlin

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de .

ISBN 978-3-8325-3818-7

Logos Verlag Berlin GmbH
Comeniushof, Gubener Str. 47,
10243 Berlin
Tel.: +49 (0)30 42 85 10 90
Fax: +49 (0)30 42 85 10 92
INTERNET: http://www.logos-verlag.de

ANNOTATION

The monograph "Modern Methods of Statistical Analysis of Medical Data" covers both parametric and non-parametric methods of analysis which are considered to be fundamental: descriptive statistics, analysis of variance and relations, multiple regression analysis as well as modern methods such as factor analysis, discriminant analysis, cluster analysis and loglinear analysis. Most books on biostatistics can be divided into two categories: first group of books applies academic mathematical approaches which do not give a clear understanding to medics; the second group is based on simplified mathematical apparatus without explanations of basic principles. This book is among the few where the author strictly but perfectly clear describes the principles of each method providing scheme of its application and indicating possible limitations and errors.

The book consists of two parts. The first part outlines the theoretical basis of biostatistics. It determines fundamental concepts which are used for planning the statistical analysis. The research material for statistical processing should be prepared correctly to get the reliable results. For competent statistical analysis the task should be formulated in such a way that appropriate statistical methods can be applied for it solution, in other words, a mathematical and statistical model should be developed.

It is the most difficult point of the working plan for beginners since they need to outline the opportunities given by applied statistics, and understand which of them can be useful for a particular purpose. Finally, it is important to consider what the data requirements accompany these methods, and verify whether these requirements are satisfied. The book provides the schemes which show a wide range of methods used to solve the most common tasks. Traditional textbooks usually focus on statistical conclusions based on the assumption of a normal distribution of variables, including multivariate statistical methods such as regression, factor and dispersion analyses. However, parametric methods cannot be used in every case, even for continuous variables, especially for a small sample. As regards practical tasks, they often require to identify the differences and relations of ordinal or discrete variables or a set of different variables. Taking into consideration these factors, the book covers non-parametric methods of data analysis.

However, the theory becomes clear and understandable if it can be used to solve practical tasks. Therefore, besides the theoretical fundamentals the second part of the text describes application of the software package Statistica for solving medical tasks which simplifies them and reduces task completion time. Such a choice is also connected with the convenience of export and import of data and results in this system. Solving tasks are considered step by step. The text is well illustrated with pictures and screenshots which are extremely helpful in acquiring skills to use theoretical knowledge in practice. The author had to sacrifice strict mathematical descriptions in order to make the text clear for a wide range of readers. Cumbersome proving is replaced by simple and not very strict explanation. All examples are taken from clinical practice or developed hypothetically to illustrate the method briefly. For some readers too many details may seem to be the disadvantage. However, due to minute description the text can solve problems with minimal loss of time. It is up to the reader to decide whether it is good or bad.

The book is intended primarily for medical students and health care professionals who are more interested in applying statistical methods than their strict mathematical justification. It will be useful for practicing physicians and researchers to critically evaluate the medical literature and planning, conducting and analyzing the results of research. However, it may be useful as an introductory course for those who want to study biostatistics more profoundly. This book will give an opportunity to get a clear understanding of biostatistics, its methods and the place occupied in medical education and work of the practicing physicians.

To my mother Anna Trukhacheva

INTRODUCTION

The monograph presents the revised and corrected lectures given by the author to students and researchers of Altai State Medical University. It focuses on the fundamentals of biostatistics for medical data analyzing. Each theory becomes more understandable and accessible, if it is possible to use it for solution of practical problems. For this reason the monograph describes how to solve problems using the software package Statistica in addition to theoretical basics. It allows every interested reader to acquire skills of using theoretical knowledge in practice, and makes the tasks solution significantly faster. The author had to sacrifice strict mathematical descriptions in order to make the text clear for a wide range of readers. Cumbersome proving is replaced by simple and not very strict explanation. All examples are taken from clinical practice or developed hypothetically to illustrate the method briefly. For some readers too many details may seem to be the disadvantage. However, due to minute description the text can solve problems with minimal loss of time. It is up to the reader to decide whether it is good or bad.

The book is intended primarily for medical students and health care professionals who are more interested in applying statistical methods than their strict mathematical justification. It will be useful for practicing physicians and researchers to critically evaluate the medical literature and planning, conducting and analyzing the results of research. However, it may be useful as an introductory course for those who want to study biostatistics more profoundly.

I sincerely appreciate the assistance of people who helped me with the monograph. I express my profound gratitude for support and discussion of this work to Professor Yu. A. Vysotskij, Dr. N.P. Pupyrev and Dr. S.V.Hlybova. I thank tremendously for the design E.I.Vorsin and M. V. Nechaev.

N. Trukhacheva

CHAPTER 1

HISTORY OF BIOSTATISTICS

The history of statistical science is said to start in the middle of the XVIII century; although the practical operations to collect data on population, its structure, property status, and other information were known long before. The term "statistics" derived from the Latin *status* designating “state of affairs". Originally statistics described the "sights" of the state, and only in the XIX century statistical information was started to be quantified.

Statistical science is associated with names of English economist William Petty (1623-1687) and John Graunt (1629-1674) whose statistical and demographic ideas were developed by their followers - German pastor Johann Peter Sussmilch II (1707-1767) and the prominent Belgian scientist of the XIX century Lambert-Adolph-Jacques Quetelet (1796-1874). Adolph Quetelet's works showed the importance of statistics in learning laws of social life, detecting that these laws are clearly evident only in the mass of phenomena; that is, studying data on a large number of cases. In addition A. Quetelet founded biometrics. His doctrine of statistical regularity was developed by German statistician and economist William Lexis (1837-1914).

Further development of statistical science is associated with works of Francis Galton (1822-1911), Karl Pearson (1857-1936), Ronald Fisher (1890-1962), William Sealy Gosset (1876-1937) and other Western scholars. A number of indicators and criteria were named after them. F. Galton introduced the term "regression" in 1886. He found out that on average children of tall fathers are not so tall, and sons of fathers of small stature are taller than their fathers. This was interpreted by him as "regression towards mediocrity." K. Pearson improved the methods of correlation and regression proposed by F. Galton. K. Pearson introduced to statistics such concepts as standard deviation and variation. He developed chi-squared test and introduced generally accepted term "normal distribution". The idea of controlled clinical trials appeared in the XII century when Frederick II, Holy Roman Emperor, studied how physical exercises influenced digestion. Two knights had been given the same food; then one of them went to bed, and the other went a-hunting. A few hours later he killed both of them and made a careful study of their digestive tracts. Digestion of the sleeping man was more intensive. In the XVII century Jan Baptist van Helmont decided to call into question the practice of bloodletting and offered the first randomized clinical trial with a large number of participants and statistical analysis. About 500 people were expected to be randomly divided into two groups. Bloodletting was not used in one of the group. As regards the other group, doctors could apply this method as many times as it was necessary. The effectiveness of bloodletting was evaluated due to the number of funerals in each group. For unknown reasons the experiment was not carried out; however, later bloodletting was proved to be ineffective by P. Louis.

However, extensive use of statistical data for medical research was started only in the XIX century. The science of application of mathematics in biology directly intertwined with the development of genetics. It was genetics, and especially Mendel's Genetic Laws, to become the main area of application of statistical methods in biology. Recently statistical methods having penetrated into different branches of medicine have become the principal methods of analysis and processing of the experimental data. However, power of developed statistical research methods tend to come into conflict with the lack of adequate knowledge of doctors who have to use these methods. Any doctor should be familiar with the basic principles and methods of statistics in order to competently discuss new diagnostic techniques and choose the best method of treatment. German scholar H.G. Wells pointed out that not only doctors but any modern person should have such a "statistical thinking":”Statistisches Denken wird für den mündigen Bürger eines Tages dieselbe Bedeutung haben wie die Fähigkeit, lesen und schreiben zu können". Due to the wide spreading of the "evidence based medicine" ideology in the world, recently the fundamentals of biostatistics have become a necessary element of education at the medical schools.

CHAPTER 2

DESCRIPTIVE STATISTICS

2.1. RANDOM EVENTS AND RANDOM VARIABLES

When studying one's health, it is essential to consider many factors - both improving and worsening one's medical condition. All these factors must be expressed in certain quantitative estimations. To obtain necessary numerical data, a number of observations is required, as most of the random and unforeseen events are subjected to some general non-random laws.

The science that studies the pattern of mass random events is called the theory of probability. Application of the probability theory to the processing of a large set of numbers is called mathematical statistics.

The examples of random events are everywhere. For instance, in questions: "Will it snow tomorrow? Which side will a tossed up coin fall down?" In other words, whenever there is no complete information, an accident occurs.

Statistical definition of probability

Ratio limit of the number of trials *m*, in which the event *A* happened, to the total number of trials *n*, providing that the total number of trials *n* goes to infinity, is called probability of the event *A*.

$$P(A) = \lim_{n\to\infty} \frac{m}{n}$$

The number of trials must be large enough. For instance, two trials are insufficient to determine the probability of appearance of heads or tails, as in each of the cases both heads and tails can appear. Therefore the probability of their appearance will be 100%.

The definition of probability given above is named statistical. It allows calculating the probability of such events, the structure of which is unknown and the frequency of which cannot be predicted in advance. For example, only the statistical data collected over the years made it possible to find the boys and girls birth probability. It turned out that these probabilities are different. The boys' birth probability is about 0.52; therefore, the girls birth probability is 0.48.

THE NOTION OF RANDOM VARIABLE

Random variable is a variable whose value is subject to variations due to chance and cannot be predicted on the basis of trial conditions.

Even the most accurate method of analysis gives certain deviation in the results when repeated (repeatability error). It means that every numerical result is a random event. Sugar and hormones content in blood, height and weight etc. of a patient under examination are all random. In medicine and biology the patient is regarded as an *object* of observation. During the observations severity of illness, height, weight, quantitative data of laboratory assessment etc. are defined. Certain parameters such as gender are qualitative; others such as height are quantitative.

Random variables can be divided into two basic classes: *discrete and continuous.*

- **Discrete** random variables take on strictly defined values and there can be no other values between them.
- **Continuous** random variables take on any value within a given interval.

THE TYPES OF SCALES IN STATISTICS (OR TYPES OF VARIABLES)

Variables differ by how "precise" they can be measured or, in other words, how much measureable information is provided by their measurement scale. Type of the scale, in which the measurement is performed, is another factor that determines the amount of information which variable contains. The following types of scales are distinguished: *nominal scale, ordinal scale, interval scale, and ratio scale.*

- **Nominal scale** is used only for proper objects classification in order to distinguish one object from another: number of an animal in the group or the unique code assigned to him, etc. Such variables can be measured only in reference to different classes; however, these classes cannot be arranged. Typical examples of nominal variables are gender, nationality, color etc. Nominal variables are also known as *categorical*. Categorical variables are often presented as the monitoring frequency referred to specific categories and classes. If there are only two classes, the variable is called *dichotomous*. For example: 1 – male gender, 2 – female gender. It is seen that coding of the variable *gender* with the help of numbers 1 and 2 is absolutely arbitrary; they could be interchanged or represented with another numbers. The same situation is with the variable *marital status*. In this case again the correspondence between the numbers and categories of marital status has no empirical value. But in contrast to gender, this variable is not dichotomous – it has four code numbers instead of two: 1 – single, 2 – married, 3 – widower/widow, 4 – divorced. Processing capabilities of nominal scale variables are very limited. For instance, the calculation of mean value for the variable *gender* is completely pointless.

- **Ordinal scale.** This scale only arranges objects assigning to them various grades. In addition it indicates which of them to a greater or lesser extent possess the quality evaluated by the variable. However, values of the variables do not provide the possibility to say “how much bigger” or “how much smaller” one value is than another. Numbers of buildings on the streets are measured in ordinal scale. A typical example of an ordinal variable – clothing size: S, M, L, XL, XXL, XXXL, XXXXL. The Mohs scale of mineral hardness is also ordinal. School grading scale (five points, twelve points, etc.) can be attributed to the ordinal scale. Variable *Smoking* is possible to rank on an ordinal scale from the bottom upwards: 1- do not smoke, 2 – smoke rarely, 3 – smoke often, 4 – smoke very often. Light smoker smokes more than non-smoker, while heavy smoker smokes more than light smoker, etc. The scale of hypertensive disease stages, heart failure degrees scale, scale of coronary insufficiency stages are ordinal scales in medicine. In this case comparing mean values in two samplings makes no sense. The empirical importance of these variables does not depend on the difference between the neighboring numerical values. Thus, despite the fact that the difference between the values of code numbers for non-smoker, light smoker and heavy smoker in both cases equals one, it is impossible to say that the actual difference between non-smoker, light smoker and heavy smoker is the same. These concepts are too vague to draw such conclusions.

- **Interval scale** not only allows arranging the measurement objects, but also makes it possible to express them numerically and compare the difference between them. For example, the temperature, measured in the degrees Fahrenheit or degrees Celsius, generates the interval scale. According to the Celsius scale, as it is known, 0°C was defined as the freezing point of water and 100°C was defined as the boiling point of water. Consequently, the temperatures interval between the freezing point and the boiling point is divided into 100 equal parts. In this case it will be wrong to declare that a body with a temperature of 40°C is two times hotter than a body with a temperature of 20°C. The interval scale keeps the length ratio of the intervals. It is not only possible to say that the temperature of 40°C is higher than the temperature of 30°C, but also that the increase of temperature from 20°C up to 40°C is twice as high as the increase from 30°C up to 40°C. Now consider the intelligence quotient (IQ). Its absolute values show ordinal relation between the respondents, and the difference between the two values is also empirically important. For example, if Fedor’s IQ is 80, Peter’s is 120 and Ivan’s is 160, it is possible to say that Peter is as “intelligent” in comparison to Fedor as Ivan is “intelligent” in comparison to Peter (i.e. – by 40 units). However, it is impossible to conclude that Ivan is twice smarter than Fedor based only on the fact that Fedor’s IQ is two times smaller. Such variables can be processed with any statistical methods without restrictions. It means, for instance, that a mean value is a valid statistical indicator to characterize such variables.