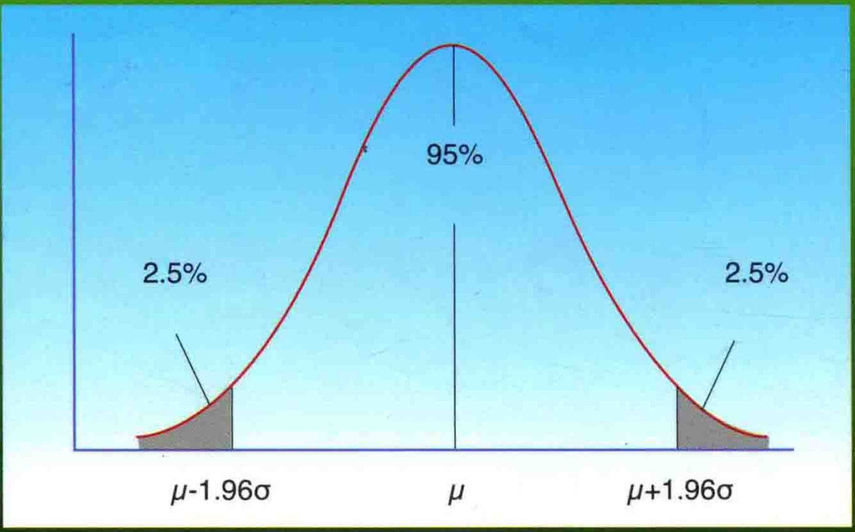
医学教育改革系列教材

Medical Statistics

Chief Editor Xiuhua Guo



高等教育出版社
HIGHER EDUCATION PRESS

医学教育改革系列教材

Medical Statistics

Chief Editor

Xiuhua Guo

Associate Chief Editors

Shicheng Yu Xinghua Yang

Contributors

Dan Feng Xiang Guo Xiuhua Guo Hong He

Xia Li Fen Liu Yanxia Luo Manshu Song

Youxin Wang Lijuan Wu Yuxiang Yan Xinghua Yang

Shicheng Yu Ling Zhang Huiping Zhu

高等教育出版社·北京
HIGHER EDUCATION PRESS BEIJING

图书在版编目（C I P）数据

医学统计学 = Medical Statistics : 英文 / 郭秀花主编. -- 北京 : 高等教育出版社, 2013.5
ISBN 978-7-04-037064-5

Ⅰ. ①医… Ⅱ. ①郭… Ⅲ. ①医学统计－统计学－高等学校－教材－英文 Ⅳ. ① R195.1

中国版本图书馆 CIP 数据核字（2013）第 064197 号

总 策 划 林金安 吴雪梅 席 雁

策划编辑 瞿德竑 责任编辑 瞿德竑 封面设计 张 楠 责任印制 毛斯璐

出版发行	高等教育出版社	咨询电话	400－810－0598
社 址	北京市西城区德外大街 4 号	网 址	http://www.hep.edu.cn
邮政编码	100120		http://www.hep.com.cn
印 刷	北京中科印刷有限公司	网上订购	http://www.landraco.com
开 本	889mm×1194mm 1/16		http://www.landraco.com.cn
印 张	11.25	版 次	2013 年 5 月第 1 版
字 数	380 千字	印 次	2013 年 5 月第 1 次印刷
购书热线	010－58581118	定 价	23.00 元

物 料 号 37064－00

医学教育改革系列教材编委会

《医学统计学》编委会

主　编　郭秀花

副主编　于石成　杨兴华

编　委（以姓氏拼音为序）

冯　丹　郭　翔　郭秀花　和　红　李　霞
刘　芬　罗艳侠　宋曼殳　王友信　吴立娟
闫宇翔　杨兴华　于石成　张　玲　祝慧萍

Foreword

Global developments in medicine and health shape trends in medical education. And in China education reform has become an important focus as the country strives to meet the basic requirements for developing a medical education system that meets international standards. Significant medical developments abroad are now being incorporated into the education of both domestic and international medical students in China, which includes students from the districts of China's Hong Kong, Macao and Taiwan that are taught through mandarin Chinese as well as students from a variety of other regions that are taught through the English language. This latter group creates higher demands for both schools and teachers.

Unfortunately there is no consensus as to how to improve the level and quality of education for these students or even as to which English language materials should be used. Some teachers prefer to directly use original English language materials, while others make use of Chinese medical textbooks with the help of English language medical notes. The lack of consensus has emerged from the lack of English language medical textbooks based on the characteristics of modern medical education in China.

In fact, most Chinese teachers involved in medical education have already attained an adequate level of English language usage. However, English language medical textbooks that reflect the culture of the teachers would in fact make it easier for these teachers to complete the task at hand and would improve the level and quality of medical education for international students. In addition, these texts could be used to improve the English language level of the medical students taught in Chinese. This is the purpose behind the compilation and publishing of this set of English language medical education textbooks.

The editors in chief are mainly experts in medicine from Capital Medical University (CCMU). The editorial board members are mainly teachers of a variety of subjects from CCMU. In addition, teachers with rich teaching experience in other medical schools are also called upon to help create this set of textbooks. And finally some excellent scholars are invited to participate as final arbiters for some of the materials.

The total package of English medical education textbooks includes 63 books. Each textbook conforms to five standards according to their grounding in science; adherence to a system; basic theory, concepts and skills elucidated; simplicity and practicality. This has enabled the creation of a series of English language textbooks that adheres to the characteristics and customs of Chinese medical education. The complete set of textbooks conforms to an overall design and uniform style in regards to covers, colors, and graphics. Each chapter contains learning objectives, core concepts, an introduction, a body, a summary, questions and references that together serve as a scaffold for both teachers and students.

The complete set of English language medical education textbooks is designed for teaching overseas undergraduate clinical medicine students (six years), and can also serve as reference textbooks for bilingual teaching and learning for 5-year, 7-year and 8-year programs in clinical medicine.

We would like to thank the chief arbiters, chief editors and general editors for their arduous labor in the writing of each chapter. We would also like to acknowledge all the contributors. Finally, we would like to acknowledge Higher Education Press. They have all provided valuable support during the many weekends and evening hours of work that were necessary for completing this endeavor.

President of Capital Medical University
Director of English Textbook Compiling Commission
Zhaofeng Lu
August 1st, 2011

Preface

Medical statistics is an applied science which combines theory of statistics with medical science. It is an indispensable theoretical science to conduct scientific research and a mandatory course for all medical majors in universities. Currently, although there are many publications about medical statistics and manuals of statistical software, most of the textbooks of medical statistics emphasize on the theoretical aspects and lack of straight and thorough introduction of statistical software. Also, most of the manuals of statistical software mainly focus on the introduction of the software while lack concepts and theory of statistics. This textbook combines statistical methods with the common manipulation of SPSS software, which makes up the shortcomings mentioned above.

There are thirteen chapters in this textbook: Introduction to Medical Statistics; Tables and Graphs; Descriptive Statistics of Continuous Variables; Descriptive Statistics of Categorical Variables; Inferential Statistics—Confidence Intervals; Inferential Statistics—*t* Tests; Analysis of Variance; Chi-square Test; Non-parametric Statistics; Correlations and Simple Linear Regressions; Multiple Linear Regression; Logistic Regression; Study Design, Sample Size Estimation and Selection of Statistical Methods. In addition, this book has several appendices, including tables of probability distributions and comprehensive self-test. Authors referred to a lot of books and information related to theories and exercises of medical statistics, as well as the manual of SPSS software during writing this book. This textbook has the following characteristics compared to others. First, it introduces basic concepts, principles and methods of medical statistics systematically and practically, especially in the statistical design of the experiment in terms of the specific problems, adequate use of statistical methods based on actual data and the reasonable explanation for statistical results. Second, flexible, convenient and user-friendly SPSS is applied, saving a lot of statistical computation work and time for learners; thus, students can focus on the deep understanding of statistics. Lastly, we tried to emphasize the application and generalization of statistical methods, and combine these methods with the modern statistical theory, such as sequential contingency table and multivariate statistical modeling.

I appreciate the help and guidance from Fuhua Xian, Vice President of Capital Medical University, and Li Fu, Dean of Department of Academic Affairs of the university, both are in charge of the publication of this book. Also, I am grateful to those who participated in compiling this book. Thanks for the advice given by the Dean of Wei Wang and Secretary Aimin Guo of the School of Public Health of Capital Medical University. Thanks to the associate editors of Professor Shicheng Yu and Associate Professor of Xinghua Yang and academic secretary Xia Li. The postgraduate students Wei Wang, Lixin Tao, Lei Pan, Da Huo, Xiangtong Liu, Chao Wang, et al have contributed to checking and typesetting. At last, I appreciate the support from my family.

Although we are aiming at the innovation of the content and application of statistics in this textbook, mistakes are unavoidable since our knowledge is limited. We will appreciate any comments and suggestions towards the book so as to improve it for the second edition. Please contact me at guoxiuh@ccmu.edu.cn. Thanks!

Professor Xiuhua Guo

January, 2013

Contributors

Dan Feng　冯丹

Institute of Hospital Management PLA General Hospital

Beijing,China

Chapter 7

Xiang Guo　郭翔

Peking University Clinical Reasearch Institute

Beijing,China

Chapter 12

Xiuhua Guo　郭秀花

School of Public Health

Capital Medical University, Beijing, China

Chapter 11

Hong He　和红

School of Sociology & Population Studies

Renmin University of China, Beijing, China

Chapter 9

Xia Li　李霞

School of Public Health

Capital Medical University, Beijing, China

Chapter 10

Fen Liu　刘芬

School of Public Health

Capital Medical University, Beijing, China

Chapter 2

Yanxia Luo　罗艳侠

School of Public Health

Capital Medical University, Beijing, China

Chapter 8

Manshu Song　宋曼殳

School of Public Health

Capital Medical University, Beijing, China

Chapter 6

Youxin Wang　王友信

School of Public Health

Capital Medical University, Beijing, China

Chapter 4

Lijuan Wu　吴立娟

School of Public Health

Capital Medical University, Beijing, China

Chapter 3

Yuxiang Yan　闫宇翔

School of Public Health

Capital Medical University, Beijing, China

Chapter 5

Xinghua Yang　杨兴华

School of Public Health

Capital Medical University, Beijing, China

Chapter 13

Shicheng Yu　于石成

Division of Health Statistics National Center for Public Health Surveillance and Information Services

Chinese Center for Disease Control and Prevention (China CDC), Beijing, China

Chapter 1

Ling Zhang　张玲

School of Public Health

Capital Medical University, Beijing, China

Chapter 8

Huiping Zhu　祝慧萍

School of Public Health

Capital Medical University, Beijing, China

Chapter 3

CONTENTS

Chapter 1

Introduction to Medical Statistics

▪ Objectives

Statistics is the science of collecting, analyzing, and interpreting data. The objective of this chapter is to introduce basic statistical concepts, such as variables, types of data, probabilities, populations and samples. Students should also understand the steps of statistical work after finishing this chapter.

▪ Key Concepts

Quantitative data and qualitative data; Continuous variables and discrete variables; Ordinal variables; Categorical variables; Probabilities; Populations and samples; Parameters and statistics.

1.1 Definition of Medical Statistics

Medical statistics plays a key role in medical research, evidence-based medicine, and evidence-based public health policy making. In medical research, data must be collected and analyzed in order to answer a specific research question. During this process medical statistics provides knowledges of the study design, data collection, data management, data analysis, and interpretation of the results. Before a new drug, treatment, or device can be marketed, an experimental study or quasi-experimental study must be conducted to evaluate the effectiveness and safety of the new health technique and service. In addition, many other questions need to be addressed. What major health problems are currently present in the region? From where do increases in healthcare spending originate? Where should a government invest its resources if it wishes to reduce the rate of birth defects? What are the effects of workplace health and safety on the nurses' employment rates? What types of services are used by long-term home care users, and how has this service changed over time? What factors are associated with the increased risk of ischemic heart disease (IHD)? To answer these questions and many others, methods and skills of medical statistics are essential. Most importantly, researchers can inform policy makers and influence public health policy making with evidence from the research, analysis, data series, and evaluation, in which medical statistics plays an important role.

Statistics is the science of collecting, organizing, analyzing, and interpreting numerical facts that we call data. It also involves planning a study in terms of principles of the study design, such as surveys and experiments. Biostatistics is the application of statistics to a wide range of topics in biology. The science of biostatistics involves ① the design of biological experiments, especially in medicine and agriculture; ② the collection, summarization, and analysis of data from those experiments; ③ the interpretation of, and inference from, the results. As opposed to biostatistics, medical statistics involves applications of statistics to medicine and health sciences, including epidemiology, public health, forensic medicine, and clinical research. Simply, medical statistics is the practice of statistics in medical sciences.

1.2 Variables and Types of Data

A set of numbers is composed of data that have been manipulated to obtain an average or a graph to gain insight into the data. To do so, classifying data and identifying types of data are essential for thoroughly understanding the context of data.

A variable is a characteristic of interest about each individual element of a population or sample. Variables could be weight, age, blood pressure, or occupation, and many other things. Variables can be classified as dependent variables and independent variables. Dependent variables are dependent on the independent variables. For example, several researchers wish to determine how high density lipoprotein (HDL) cholesterol may influence the risk of an individual for developing IHD. The level of HDL cholesterol is the independent variable. The observed result of the development of IHD is the dependent variable, since HDL cholesterol is causally associated with the development of IHD. In this context, an independent variable is also referred to as an "explanatory variable" or a "predictor variable"; a dependent variable is also known as a "response variable" or an "outcome variable".

A variable takes a number or quantity as its value. If a person weights 76 kilograms (kg), the value of the variable weight is 76 kg.

A survey or experiment yields a set of data ranging from a few measurements to thousands of observations. Table 1-1 shows aggregated data of hand-foot-mouth disease (HFMD) cases in Beijing from September 1, 2010 to September 7, 2010. During that one-week period of time, there were 563 HFMD cases reported to the China CDC through the National Disease Reporting Information System (NDRS).

Table 1-1 Reported HFMD cases during a one-week period in Beijing

Areas	Cases	Deaths	With severity
District	542	2	12
County	21	3	12
Total	563	5	24

Tables, graphs, and numerical summary measures can all be used to present data; the table above is a summary of descriptive statistics. However, the type of data must first be determined prior to the presentation of the data. The types of data are usually defined as either quantitative or qualitative data.

Quantitative variables take numerical values that arithmetic operations can be applied to; they are further divided into two types: continuous and discrete variables. The following table (Table 1-2) presents the individual records for the HFMD cases from the data in Table 1-1. In Table 1-2, age is a continuous variable, since it takes numerical values and arithmetic operations can be applied to it. For example, the average age and difference of ages between the two groups could be calculated from the data. A discrete variable only takes integers, such as the number of patients in a hospital in January, the number of newborns in a county, and the number of injuries that a person sustained in the last three months, etc.

Qualitative data also consist of two types of variables: categorical and ordinal variables. Categorical variables describe the data in terms of some qualities or categorizations, including well-defined aspects of the variable (e. g. gender, nationality, ethnicity, yes or no, life or death, pass or fail). Severity, another type of qualitative variable, is denoted as mild, moderate, or severe in Table 1-2. It is known as an ordinal variable, since serious order of mild, moderate, and severe cases exists.

Table 1-2 Individual HFMD cases reported to the China CDC during a one-week period in Beijing

ID	Age (years)	Gender	Severity	Lab-confirmed
1	5	Female	Mild	No
2	3	Female	Mild	No
3	1	Male	Moderate	No
⋮	⋮	⋮	⋮	⋮
562	7	Male	Severe	Yes
563	9	Female	Mild	No

1.3 Probabilities

The result of an observation or experiment is an event or basic element to which probability can be applied. An outcome of the lung cancer, whether a student passes an examination, and whether a 40-year-old man lives in the next 5 years are all events. Uppercase, or capital, letters usually represent such events. There are several operations that can be performed on the event.

The intersection of two events, A and B, is denoted by A∩B, and defined as the event "both A and B" (as shown in Figure 1-1a). Let A represents the event that a man has a stomach cancer; let B represents the event that the same man's wife suffers from a stomach cancer as well. The intersection of events A and B would be the event that both the man and his wife sustain stomach cancers.

The union of two events, A and B, is denoted by A∪B, and defined as the event "both A and B" or "either A or B" (as shown in Figure 1-1b). In the example mentioned above, the union of events A and B would be the event that either the man or his wife has a stomach cancer, or that they both have stomach cancers.

The complement of an event, A, is denoted $\overline{A}$ (as shown in Figure 1-1c). The complement is the event that is "not A". Event A indicates that the man has the disease, so the complement of event A, denoted as $\overline{A}$, indicates that the man does not have the disease.

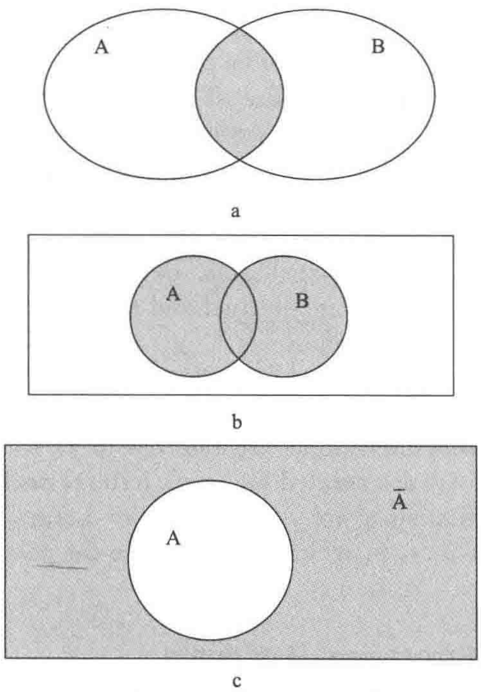


Figure 1-1 Venn diagrams showing the operations of events

After numerous repeated trials under virtually identical conditions the probability of an event A is the relative frequency of the occurrence according to the frequentist definition. If an experiment is repeated n times under essentially identical conditions, and if an event A occurs m times, then as n grows larger, the ratio of m to n approaches a fixed limit referred to as the probability of event A: $P(\mathrm{A})=\frac{m}{n}$.

The probability of an event takes a numerical value that lies between 0 and 1. A value of 1 indicates that a particular event occurs in each of the n trials and that the probability is $n/n=1$. If an event can never happen, it has a probability of $0/n=0$.

1.4 Populations and Samples

The main use of medical statistics is to infer information about the population from samples taken from that population. The key components are sampling methods representative of the population and the statistical technique. The population is a collection of similar people, observations, or measurements, in which certain subjects can be sampled to infer a property or attribute of the population. For instance, Chinese adults living in China, 12-year-old students in Beijing, or patients with diabetes mellitus in Shanghai all are populations.

Sometimes, the number of subjects or individuals in a population might be too large, so a sample of that population is selected. The purpose of sampling is to select and study a part of the population to infer information for the whole population. A sampling survey of 2‰ of the population across China was conducted in 2006 in order to get an idea of the prevalence and distribution of visual, hearing, speech, physical, intellectual and mental disabilities among Chinese people.

The values or numbers calculated from the population are often called parameters, and those derived from the sample are referred to as statistics. Parameters are notated by Greek letters, such as μ, σ, ρ. Statistics are represented by Latin letters, such as $\overline{x}$, s, r, which correspond to the parameters above.

1.5 Errors and Residuals

In statistics, the term error, or the statistical error, is the deviation of the observed value from the true value or the expected value that cannot be measured. It is the amount that an observation differs from its expected value that is based on the entire population. For example, if the mean weight in a population of male adults is 73.5 kg (μ), and a man's weight from the randomly chosen sample is 73.0 kg (x_i), then the error is -0.5 kg ($x_i-\mu$). In fact, the population mean of weight ($\mu=73.5$ kg) for male adults cannot be observed through measurement; therefore the error cannot be determined either. Often, the expected value or true value is considered as an accepted value or a given true value, such as the value from published literatures, or a value from a census or a national survey.

Another term of residual or fitting error is a measurable estimate of the unobservable statistical error. From the previous example of male adults' weight, a random sample with n subjects is chosen, and the sample mean ($\overline{x}$) could be viewed as a good estimator of the population mean of weight; then the difference ($x_i-\overline{x}$) between the weight of a male adult (x_i) from the sample and the sample mean of weight ($\overline{x}$) is referred to a residual. Through the difference ($x_i-\overline{x}$) or residual we could have an insight into the quality of data and errors of the measurement; moreover, residual is employed to calculate the variance and standard deviation for a randomly chosen sample.

There are two types of errors: stochastic errors and non-stochastic errors. In a measurement, the stochastic error or random error is the one that is randomly occurred from one measurement to the next measurement. Some stochastic errors could be identified and controlled for the measurement, such as sampling errors. Non-stochastic errors are composed of systematic errors and non-systematic errors. The values from the systematic error are constant or changes according to a certain rule. Bias is one of the systematic errors; it is stemmed from some non-experimental factors in clinical trials, and it can distort true effect of the treatment in the experiment. Non-systematic errors are caused by occasional errors from researchers during the experiment.

1.6 Steps of Statistical Work

The entire process of statistical work contains four steps: statistical design, data collection, data management, and data analysis. All of the above steps are important and none of them can be ignored.

Statistical design is the first step of statistical work, and is also a key step in medical research. It guides the data collection and data analysis on the right track. In this stage, the study must be carefully devised and arranged in terms of the principles of the study design, especially for sampling, sample size, data collection, quality control, statistical method, as well as organization and implementation of the plan.

Data collection is the basis of statistical work. Its purpose is to collect reliable original data based on the statistical design. Data can be characterized as routine data or one-time collected data. The criteria of data are required for accuracy, completeness, timeliness, comparability, and usability.

Data management involves checking, correcting, and manipulating data to eventually make the data systematic, logical, and usable for analyses.

Data analysis, the last step of statistical work, involves statistical descriptions and inferences. A descriptive statistics describes the properties and distributions of the data using summary measurements, such as mean, standard deviation, frequency, and percentiles. A statistical inference involves inferring some attributes of the population using the information from samples under a certain confidence level or probability.