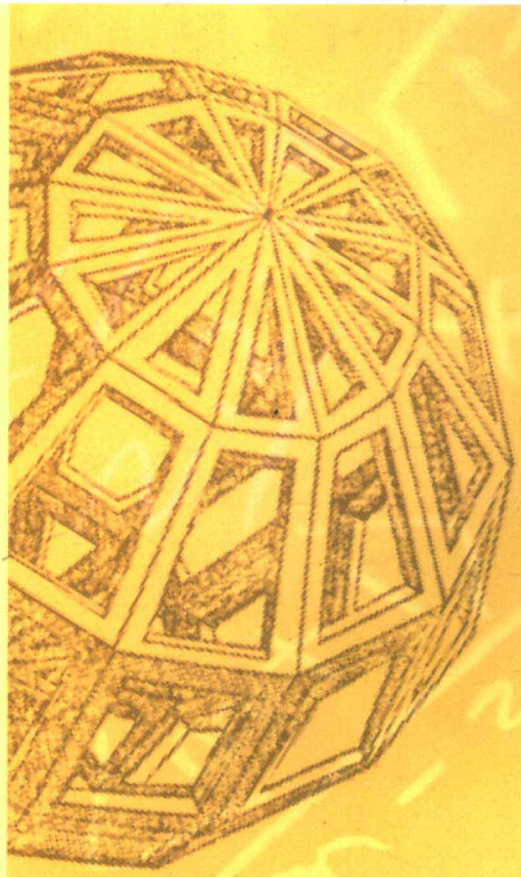
高等学校教材

Probability and Statistics

主编 赖虹建 郝志峰



高等教育出版社

高等学校教材

Probability and Statistics

主编　赖虹建　郝志峰
编者　陶志穗　林健良　朱锋峰
　　洪　毅　薛　菲

高等教育出版社

内容提要

本书是在教育部大力推进双语教学的大背景下推出的，结合当前开展概率论与数理统计课程双语教学的实际情况，以教育部数学与统计学教学指导委员会制定的本课程教学基本要求为依据，同时兼顾概率论与数理统计课程的考研大纲要求，突出统计学的思想，着重讲解概率论与数理统计的基本概念、基本理论及基本方法，注重体现概率统计在各个领域中的广泛应用，将大量的实际应用贯穿于理论讲解的始终，将经典和近代统计思想与算法阐述得更为具体。

本书介绍了随机事件及其概率、随机变量与概率分布、连续型随机变量、多维随机变量和中心极限定理、统计描述、参数估计、假设检验、非参数统计、回归分析以及方差分析。本书汲取了英文原版教材中流行的直观、灵活的教学方法及通过图表和案例进行概率统计教学的长处，突出应用概率统计技能培养的特点，强调学生分析问题和解决问题的实际能力，具有覆盖面全、实用性强、示例与习题丰富、内容新颖、图解清晰、难易适度等特点。

本书邀请了一些有英美留学背景的人士进行编写和审定，在文字表述方面，英文语言简单易懂，写作风格简约。本书已在校内外经过多次试用并取得了较好的教学效果，可供高等学校理工非数学类专业概率论与数理统计课程双语教学使用。

图书在版编目(CIP)数据

概率论与数理统计＝Probability and Statistics:英文/(美)赖虹建，郝志峰主编. —北京：高等教育出版社，2008.5

ISBN 978－7－04－023605－7

Ⅰ.概… Ⅱ.①赖…②郝… Ⅲ.①概率论－高等学校－教材－英文②数理统计－高等学校－教材－英文 Ⅳ.O21

中国版本图书馆 CIP 数据核字(2008)第 048142 号

策划编辑 于丽娜 **责任编辑** 崔梅萍 **封面设计** 张申申 **责任绘图** 宗小梅
版式设计 余 杨 **责任校对** 刘 莉 **责任印制** 韩 刚

出版发行	高等教育出版社	**购书热线**	010－58581118
社 址	北京市西城区德外大街 4 号	**免费咨询**	800－810－0598
邮政编码	100120	**网 址**	http://www.hep.edu.cn
总 机	010－58581000		http://www.hep.com.cn
		网上订购	http://www.landraco.com
经 销	蓝色畅想图书发行有限公司		http://www.landraco.com.cn
印 刷	北京中科印刷有限公司	**畅想教育**	http://www.widedu.com
开 本	787×960 1/16	**版 次**	2008 年 5 月第 1 版
印 张	19.25	**印 次**	2008 年 5 月第 1 次印刷
字 数	350 000	**定 价**	22.30 元

物料号 23605－00

Preface

This book is intended to serve as a text for an elementary introduction in probability and statistics. Probability theory makes explicit references to the nature and effects of the chance phenomena and serves as the foundation upon which statistics methods are base. Probability theories and concepts, and statistical methods have found increasing applications in many other areas such as physics, biological sciences, economics and social sciences. It is doubtless that even more extensive applications of probability and statistics can be seen in the future.

There have been many excellent textbooks on probability. Many of them are written for mathematics students with a sufficient level of mathematical maturity. From our teaching of this course in the past, we found that for students with a working knowledge of differential and integral calculus, they might benefit by a treatment that emphases breadth rather than details, though of course both approaches are equally important. Non mathematics major students who are expecting to apply the theories and methods in the studies of their specific areas may benefit even more with such a treatment. This becomes an aim when we write this book. As a result, this is not the same detailed material that most efficiently leads young mathematics students to the frontiers of professional mathematics. Instead, it requires attitudes not necessarily compatible with the attitudes of such mathematics courses. In this book, the subject matter that is central to the theory is also central to the applications; and the implications of the theorems become more important than their proofs. We are hopeful that such a basic text for probability and statistics will become a useful text for students who might not have a high level of mathematical maturity, and who expect to use the theory and methods in the studies in their own fields. We are also hopeful that this book would prove an alternative way of teaching such a course.

We have run the course of Probability and Statistics in English since 1995 at South China University of Technology. The book has been developed from course notes prepared for such a course. Students taking this course have been in mathematics, statistics, engineering and other. A successful calculus sequence serves as the prerequisite for the course.

Chapters 1—5 provide most of the basics of probability theory. It has been noted

that an understanding of the basic definitions, theorems and methods in the finite case will make it much easier for students with necessary preparations to master the corresponding ideas, concepts and theories in the infinite case. Chapter 6 presents a catalog of the most frequently used distributions. The other chapters provide introductions and discussions of several topics in several statistics and statistical methods, and their applications.

Instructors using this text for a one semester course will need to be picky in choosing the materials. Chapters 1—5 are basic to any course.

Comments from readers are always welcome.

Contents

1 Introduction

Historically, the origin of probability comes from gambling. Girolamo Cardano, who lived in 1501—1576, wrote a "Gambler's manual", in which he discussed probability in terms of gambling games. In 1654, the Chevalier de Méré, a gambler, was considering the following problem: A game is played between two persons, and any one who firstly scores three points wins the game. In the game, each of the participants' places at stake 32 pistoles and the winner will take entire stake of the 64 pistoles. The Chevalier was concerned that if the players left off playing when the game was only partially finished, how should the stakes be divided? Unable to find an answer to this problem, he consulted Blaise Pascal. Pascal soon solved the problem and communicated his solution to Fermat. Later, Fermat and Pascal, two of the greatest mathematicians of their times, laid a foundation for the theory of probability in their correspondences following Pascal's solution.

Probability can be viewed as a study of the likelihood of a possible outcome to occur in a experiment. Like other branches of mathematics, there are certain undefined concepts and assumptions (called axioms) in the theory of probability. The undefined concepts of probability theory are experiment, event, and probability. We are here to give a brief informal description these undefined concepts and assumption. Formal definition will be given in the following chapters.

An experiment usually means an act such that there is uncertainty about the outcomes after it is performed. A typical example of an experiment is the act of observing the number of dots on the top face of a die upon rolling it. The mathematical counterpart of an experiment is usually called a sample space. The potential outcomes of a probabilistic experiment are called *events*. There are many experiments other than gambling games can be seen in our daily life. For example, will tomorrow be sunny, or clouded, or raining? Will the new teaching technique improve the students' learning? Will the students in your class become successful engineers? Will the next patient entering the doctor's clinic have a higher temperature? Must I wait for more than 10 minutes for the next bus? The answers to all these questions are uncertain. These are

good examples of experiments.

Probability is not only a tool for us to understand experiments with uncertain outcomes, but also a useful tool in solving problems in other areas closely related to our life. When a life insurance company sells a life insurance policy to a person, the insurance company must determine the fair amount of premium this new customer must pay for next year. How much should the fair amount of premium be? Graunt and Halley first applied probability to this problem. When the insurance company determines the premium of a customer, the insurance company must know how likely, or in mathematical terms, what is the probability of, a male in his 40s will die within one year. In other words, the insurance company must know the distribution of the probability of death, known as a mortality table in life insurance. The foundation for mortality determinations was laid by John Graunt and Edmund Halley in the late seventeen century. Graunt first made a careful analysis of London bills of mortality, and he published a book titled *Natural and Political Observations Made upon the Bills of Mortality*, in which he studied a number of interesting problems, including the ratio of new born infants, the age distribution of the population the migration into and out of city London. Later, Halley prepared the first mortality table using data from another British city Breslau. Their work became the foundation of the mortality studies, an application of probability theory to life insurance.

A natural question is: can the mortality table made by Halley be used to determine the death rate of an American city like New York? One will certainly doubt about it. Even restricted to London, one would also ask the question how closely Halley mortality table reflected the reality in Breslau in his time. These are questions to be studied in Statistics.

When using experimental and observational methods to study a problem, one must collect data by means of observations and/or experiments. These data will inevitably have some kind of uncertainty: they may be affected by the time when the data are collected, the place where the data are collected, and the mechanism with which the data are collected. The randomness of the data is also from the fact that we some times can only study a portion of the whole population, and which portion are selected to be studied is totally random. After the data are collected, one needs to analyze the data to come up with conclusions. How do we have conclusions with a reasonable level of assurance from such data with certain randomness? How big a portion we should single out to study so that the analysis will closely reflect the total population? We will inevitably encounter many problems. In order to solve these problems, statisticians

have developed many techniques and theories. These techniques and theories constitute the content of statistics. Informally speaking, statistics is a branch of mathematics that studies how to effectively collect and use the data with randomness.

In order to be the object of statistics, the data must be random. If you want to know the percentage of all male students in a university, you can just do the computation. The answer will be precise, and no statistical methods will be needed. This is not an object to be studies in statistics. On the other hand, if a city has 1 000 000 adult males, in which there are m of them are smoker, then the adult male smoking rate in this city is $m/1\ 000\ 000$. In order to determine this rate, one can survey all 1 000 000 adult males in the city to get the precise smoking rate of the city. But this is very difficult to do and in fact it is totally unrealistic. Another way to do so is to randomly select a portion of the total population, such as a group of 1 000 adult males among the total population, and then use the information collected from this group to estimate the adult male smoking rate. Then this becomes an object to be studied in statistics as the collection of data has randomness.

What do we mean by effectively collecting data? The effectiveness is in two aspects: on one hand, we want to have a simple mathematical model to treat the collected data, and so the smaller the data is, the simpler the model will be. On the other hand, we also want the collected data to contain as much related and interested information as possible, and so the data cannot be too small. Return to the example of selecting 1 000 adult males to estimate the adult male smoking rate problem. Is the number 1 000 too small or too big or just right? If too small, the data may not have sufficient information for us to make a close to reality estimation. If it is too big, then the cost for conducting the survey may be unnecessarily wasted. To determine a just right number, we need the help of statistics.

What do we mean by effectively using the collected data? The purpose of collecting data is to obtain useful information from the data. Usually, the useful information cannot be seen transparently from the collected data. Thus it requires us to use certain methods to analyze the data to come up with conclusions related to the problems being studied. To effectively use the data, one must use effective methods to analyze the data to pull out conclusions as surely as possible, as close to reality as possible.

In order to effectively collect and use data, many mathematical methods and models will be involved. Some of the most commonly used methods and models will be discussed in the chapters that follow.

While data collecting were found in history long ago, statistics became rapidly

developed starting from the early 20th century. The earlier pioneers were R. A. Fisher and K. Pearson. Fisher's two books "Experimental Design" and "Statistical Methods for Research Workers" were once viewed as most important references. Later in 1946, H. Cramer published his "Mathematical Methods of Statistics", in which he applied mathematical methods to summarize the most important achievements and discoveries in statistics in that time. Cramer's book marks the moment that statistics has become a matured branch of mathematics. Statistics has been rapidly developed due to its wide applications to many areas. It has been applied in almost all areas in today's society: statistical methods have been used in forecasting the weather, predicting earth quakes, describing the effect of medicines, and predicting social and economical activities in a society, to list just a few. More and more statistical methods are now found for their application purposes. Statistics has becoming one of the most useful tools in our society.

2 Probability

2.1 Sample Space

If we toss a properly balanced coin, it will fall with either a head or a tail showing. For a particular locality, it will probably be sunny or rain in tomorrow. An event with two or more possible outcomes is called a random event. The main purpose of probability theory is to study random events.

Statisticians use the word experiment to describe any process that generates a set of data, which comes from random events. An example of a statistical experiment is the tossing of a coin. In this experiment there are only two possible outcomes, heads or tails. Another experiment might be measuring the diameters of ball bearing produced by a certain company; here the possible outcomes contain all real numbers in a certain interval. We are particularly interested in the observations obtained by repeating the experiment several times. In most cases the outcomes will depend on chance and cannot be predict with certainty. Such experiments are called random experiments. When a coin is tossed repeatedly, we cannot be certain that a given toss will result in a head.

Definition 2.1.1 The set of all possible outcomes of a statistical experiment is called the **sample space**.

Each outcome in a sample space is called a sample point of the sample space.

Example 2.1.1 Consider the experiment of tossing a die. If we are interested in the number that shows on the top face, the sample space would be

$$S_1 = \{1,2,3,4,5,6\}.$$

If we are interested only in whether the numbers is even or odd, the sample space is simply

$$S_2 = \{\text{even},\text{odd}\}.$$

Example 2.1.2 illustrates the fact that sometimes more than one sample space can be used to describe the outcomes of an experiment.

Example 2.1.2 Measuring the diameters of ball bearing produced by a certain company, this experiment can be described by the sample space

$$S_1 = (0, \infty)$$

or, if it is known that the maximum and minimum diameter of the ball bearing are 22.5 mm and 22 mm, respectively, then the sample space is

$$S_2 = [22, 22.5].$$

Example 2.1.3 An experiment consists of flipping a coin and then flipping it a second time if a head occurs. If a tail occurs on the first flip then a die is tossing once. To list the elements of the sample space, we construct a diagram of Figure 2.1.1, which is called a tree diagram. Now the various paths along the branches of the tree give the distinct sample points. Starting with the top left branch and moving to the right along the first path, we get the sample point HH, indicating the possibility that heads occurs on two successive flips of the coin. The possibility that coin will show a tail followed by a 4 on the toss of the die is indicated by T4. Thus the sample space is

$$S = \{HH, HT, T1, T2, T3, T4, T5, T6\}. \qquad \square$$

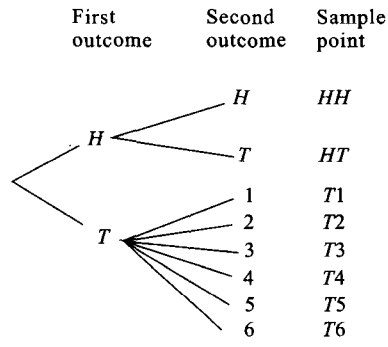


Figure 2.1.1 Tree diagram for Example 2.1.3

2.2 Events

In engineering or economic problems, we may be interested in some subset of the sample space rather than in a specific sample point in the sample space. For example, a game player may be interested in the event A that the outcome when a die is tossed is greater than 3. This will occur if the outcome is in the subset $A = \{4, 5, 6\}$ of the

sample space. In Example 2.1.2, we are concerned about the event A that the diameters of a ball bearing is in a specific interval, say $[22.2, 22.3]$. Thus

$$A = \{d \mid 22.2 \leqslant d \leqslant 22.3\}.$$

Definition 2.2.1 An **event** is a subset of a sample space.

Example 2.2.1 Given the sample space $S = \{t \mid t \geqslant 0\}$, where t is the life in hours of a certain type of light bulbs, we are interest in the event B that a bulb will be burnt out in less than 200 hrs, i.e. the subset $B = \{t \mid 0 \leqslant t < 200\}$ of S. □

Example 2.2.2 Assuming that the unemployment rate r of a region is between 0 and 15%, we have the sample space $S = \{r \mid 0 \leqslant r \leqslant 0.15\}$. If the event C "unemployment rate is low" means that $r \leqslant 0.04$, then we have the subset $C = \{r \mid 0 \leqslant r \leqslant 0.04\}$ of S. □

As events are subsets, the operations of set theory can be used in the discussion of events. So we can say about the complement of an event, the union, difference and intersection of events.

The sample space S itself, is certainly an event, which is called a certain event, meaning that it always occurs in the experiment. The empty set, denoted by $\varnothing$, is also an event, called an impossible event, meaning that it never occurs in the experiment.

Example 2.2.3 Consider the experiment of tossing a die, then

$$S = \{1, 2, 3, 4, 5, 6\}.$$

Let x be the number that shows on the top face, then the event $A = \{x \mid x \in S, x \leqslant 10\}$, is the certain event, i.e. $A = S$. Then even $B = \{x \mid x \in S, x$ is an irrational number$\}$, is the impossible event, i.e. $B = \varnothing$. □

Example 2.2.4 Consider the sample space S consists of all positive integers less than 10, i.e.

$$S = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}.$$

Let A be the event consisting of all even numbers and B be the event consisting of numbers divisible by 3. Find $\bar{A}, A \cup B, A \cap B, \bar{A} \cap B$.

Solution We have $A = \{2, 4, 6, 8\}$, $B = \{3, 6, 9\}$. Thus

$$\bar{A} = \{1, 3, 5, 7, 9\}, \quad A \cup B = \{2, 3, 4, 6, 8, 9\}, \quad A \cap B = \{6\}, \quad \bar{A} \cap B = \{3, 9\}.$$ □

The relationship between events and the corresponding sample space can be illustrated graphical by means of Venn diagrams. In a Venn diagram, we represent the sample space by a rectangle and represent events by circles drawn inside the rectangle.

Example 2.2.5 In Figure 2.2.1,

$A \cap B$ = regions 1 and 2, $A \cup D$ = regions 1, 2, 3, 4, 5 and 7,

$(A \cup B) \cap \bar{D}$ = regions 2, 6 and 7. □

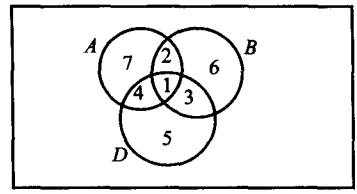


Figure 2. 2. 1 Venn diagram of Example 2. 2. 5

The following list summarizes the rules of the operations of events.

(1) $A \cap \varnothing = \varnothing$

(2) $A \cup \varnothing = A$

(3) $A \cup A = A$

(4) $A \cap \bar{A} = \varnothing$

(5) $A \cup \bar{A} = S$

(6) $\bar{S} = \varnothing$

(7) $\bar{\varnothing} = S$

(8) $\bar{\bar{A}} = A$

(9) $\overline{A \cup B} = \bar{A} \cap \bar{B}$

(10) $\overline{A \cap B} = \bar{A} \cup \bar{B}$

(11) $A \cup B = B \cup A$

(12) $(A \cup B) \cup C = A \cup (B \cup C)$

(13) $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$

(14) $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$

2. 3 Probability of Events

Experience resulting from repeated experiment is always used to predict the outcome of future events. Tossing a properly balanced coin, we can predict with certainty that the chance of its showing a head about one-half of the time. Thus we say that for a single toss, the probability of showing head is 0. 5. Knowledge of probability makes it possible to understanding statistics, to interpret statistical results.

First we consider an example. In studying the effect of seed treatments on emergence of cotton seedlings, it is necessary to know the percent of emergence of untreated seed. To do this we plant 100 untreated cotton seeds. If 49 seeds germinate, that is, if there are 49 success(by success in statistics we mean the occurrence

of the event under discussion) in 100 trials, we say that the relative frequency of success is 0.49. If we plant more and more seeds, a whole sequence of values for the respective relative frequencies is obtained. In general, these relative frequencies approach a limit value; we call this limit the probability of success in a single trial. From the data of Table 2.3.1 it appears that the relative frequencies are approaching the value 0.51, which we call the probability of a cotton seedling emerging from an untreated seed.

Table 2.3.1 Relative Frequency of Emergence of Cotton Seedlings from Untreated Seeds

Number of trial(n)	Number of successes(s)	Relative frequency(s/n)
100	49	0.49
500	261	0.522
1 000	508	0.508
5 000	2 549	0.509 8

Definition 2.3.1 If the number of successes in n trails is denoted by s, and if the sequence of relative frequencies s/n obtained for larger and larger value of n approaches a limit, then this limit is defined as the **probability** of success in a single trial.

By the definition, the probability of the certain event is 1, since its relative frequency is always 1. Similarly, the probability of the impossible event is 0, and the probability of any event is always between 0 and 1.

Note In this definition, the word "limit" has a meaning which is different from that in calculus. We will discuss this problem later.

Example 2.3.1 200 bulbs produced by company X are selected at random. Among them, 150 have life longer than 300 hrs. Find the probability that the bulbs produced by company X have life longer than 300 hrs.

Solution $p = \dfrac{150}{200} = 0.75.$ □

In many cases, the probability may be stated without an experiment. If we toss a properly balanced coin, we believe that the probability of getting a head is 0.5. We make this statement since in tossing a properly balanced coin, only two outcomes are possible and both outcome are equally likely to occur.