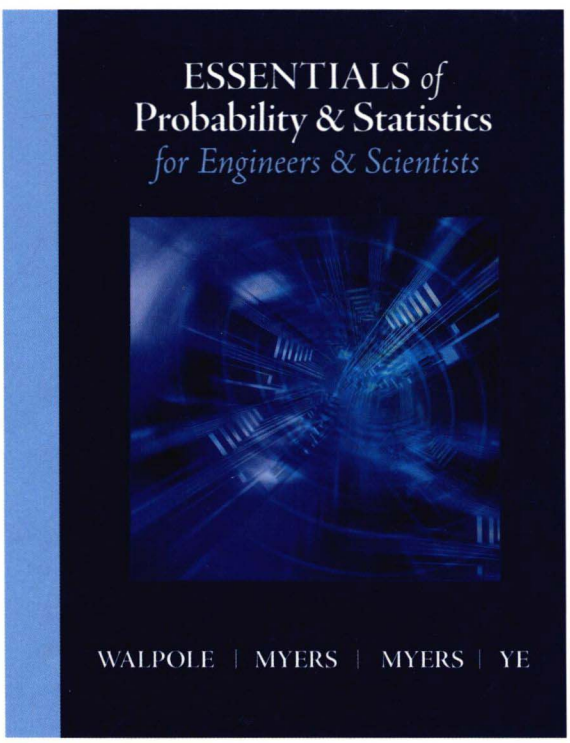
PEARSON

华章统计学原版精品系列

概率与统计

Essentials of Probability & Statistics for Engineers & Scientists

（英文版）



（美）Ronald E. Walpole　Raymond H. Myers　Sharon L. Myers　Keying Ye 著

机械工业出版社
China Machine Press

华章统计学原版精品系列

概率与统计

Essentials of Probability & Statistics for Engineers & Scientists

（英文版）

（美）Ronald E. Walpole　Raymond H. Myers　Sharon L. Myers　Keying Ye 著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

概率与统计（英文版）/（美）沃波尔（Walpole，R. E.）等著. —北京：机械工业出版社，2013.5
（华章统计学原版精品系列）
书名原文：Essentials of Probability & Statistics for Engineers & Scientists

ISBN 978-7-111-42506-9

I. 概… II. 沃… III. ① 概率论—英文 ② 数理统计—英文 IV. O21

中国版本图书馆CIP数据核字（2013）第100857号

机械工业出版社（北京市西城区百万庄大街22号邮政编码 100037）
责任编辑：迟振春
北京瑞德印刷有限公司印刷
2013年6月第1版第1次印刷
186mm×240mm · 30印张
标准书号：ISBN 978-7-111-42506-9
定价：69.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259 读者信箱：hzjsj@hzbook.com

Preface

General Approach and Mathematical Level

This text was designed for a one-semester course that covers the essential topics needed for a fundamental understanding of basic statistics and its applications in the fields of engineering and the sciences. A balance between theory and application is maintained throughout the text. Coverage of analytical tools in statistics is enhanced with the use of calculus when discussion centers on rules and concepts in probability. Students using this text should have the equivalent of the completion of one semester of differential and integral calculus. Linear algebra would be helpful but not necessary if the instructor chooses not to include Section 7.11 on multiple linear regression using matrix algebra.

Class projects and case studies are presented throughout the text to give the student a deeper understanding of real-world usage of statistics. Class projects provide the opportunity for students to work alone or in groups to gather their own experimental data and draw inferences using the data. In some cases, the work conducted by the student involves a problem whose solution will illustrate the meaning of a concept and/or will provide an empirical understanding of an important statistical result. Case studies provide commentary to give the student a clear understanding in the context of a practical situation. The comments we affectionately call "Pot Holes" at the end of each chapter present the big picture and show how the chapters relate to one another. They also provide warnings about the possible misuse of statistical techniques presented in the chapter. A large number of exercises are available to challenge the student. These exercises deal with real-life scientific and engineering applications. The many data sets associated with the exercises are available for download from the website http://www.pearsonhighered.com/mathstatsresources.

Content and Course Planning

This textbook contains nine chapters. The first two chapters introduce the notion of random variables and their properties, including their role in characterizing data sets. Fundamental to this discussion is the distinction, in a practical sense, between populations and samples.

In Chapter 3, both discrete and continuous random variables are illustrated with examples. The binomial, Poisson, hypergeometric, and other useful discrete distributions are discussed. In addition, continuous distributions include the nor-

mal, gamma, and exponential. In all cases, real-life scenarios are given to reveal how these distributions are used in practical engineering problems.

The material on specific distributions in Chapter 3 is followed in Chapter 4 by practical topics such as random sampling and the types of descriptive statistics that convey the center of location and variability of a sample. Examples involving the sample mean and sample variance are included. Following the introduction of central tendency and variability is a substantial amount of material dealing with the importance of sampling distributions. Real-life illustrations highlight how sampling distributions are used in basic statistical inference. Central Limit type methodology is accompanied by the mechanics and purpose behind the use of the normal, Student t, χ^2, and f distributions, as well as examples that illustrate their use. Students are exposed to methodology that will be brought out again in later chapters in the discussions of estimation and hypothesis testing. This fundamental methodology is accompanied by illustration of certain important graphical methods, such as stem-and-leaf and box-and-whisker plots. Chapter 4 presents the first of several case studies involving real data.

Chapters 5 and 6 complement each other, providing a foundation for the solution of practical problems in which estimation and hypothesis testing are employed. Statistical inference involving a single mean and two means, as well as one and two proportions, is covered. Confidence intervals are displayed and thoroughly discussed; prediction intervals and tolerance intervals are touched upon. Problems with paired observations are covered in detail.

In Chapter 7, the basics of simple linear regression (SLR) and multiple linear regression (MLR) are covered in a depth suitable for a one-semester course. Chapters 8 and 9 use a similar approach to expose students to the standard methodology associated with analysis of variance (ANOVA). Although regression and ANOVA are challenging topics, the clarity of presentation, along with case studies, class projects, examples, and exercises, allows students to gain an understanding of the essentials of both.

In the discussion of rules and concepts in probability, the coverage of analytical tools is enhanced through the use of calculus. Though the material on multiple linear regression in Chapter 7 covers the essential methodology, students are not burdened with the level of matrix algebra and relevant manipulations that they would confront in a text designed for a two-semester course.

Computer Software

Case studies, beginning in Chapter 4, feature computer printout and graphical material generated using both SAS® and MINITAB®. The inclusion of the computer reflects our belief that students should have the experience of reading and interpreting computer printout and graphics, even if the software in the text is not that which is used by the instructor. Exposure to more than one type of software can broaden the experience base for the student. There is no reason to believe that the software used in the course will be that which the student will be called upon to use in a professional setting.

Supplements

Instructor's Solutions Manual. This resource contains worked-out solutions to all text exercises and is available for download from Pearson's Instructor Resource Center at www.pearsonhighered.com/irc.

Student's Solutions Manual. ISBN-10: 0-321-78399-9; ISBN-13: 978-0-321-78399-8. This resource contains complete solutions to selected exercises. It is available for purchase from MyPearsonStore at www.mypearsonstore.com, or ask your local representative for value pack options.

PowerPoint® Lecture Slides. These slides include most of the figures and tables from the text. Slides are available for download from Pearson's Instructor Resource Center at www.pearsonhighered.com/irc.

Looking for more comprehensive coverage for a two-semester course? See the more comprehensive book *Probability and Statistics for Engineers and Scientists*, 9th edition, by Walpole, Myers, Myers, and Ye (ISBN-10: 0-321-62911-6; ISBN-13: 978-0-321-62911-1).

Acknowledgments

We are indebted to those colleagues who provided many helpful suggestions for this text. They are David Groggel, *Miami University*; Lance Hemlow, *Raritan Valley Community College*; Ying Ji, *University of Texas at San Antonio*; Thomas Kline, *University of Northern Iowa*; Sheila Lawrence, *Rutgers University*; Luis Moreno, *Broome County Community College*; Donald Waldman, *University of Colorado—Boulder*; and Marlene Will, *Spalding University.* We would also like to thank Delray Schultz, *Millersville University*, and Keith Friedman, *University of Texas — Austin*, for ensuring the accuracy of this text.

We would like to thank the editorial and production services provided by numerous people from Pearson/Prentice Hall, especially the editor in chief Deirdre Lynch, acquisitions editor Chris Cummings, sponsoring editor Christina Lepre, associate content editor Dana Bettez, editorial assistant Sonia Ashraf, production project manager Tracy Patruno, and copyeditor Sally Lifland. We thank the Virginia Tech Statistical Consulting Center, which was the source of many real-life data sets.

R.H.M.
S.L.M.
K.Y.

Contents

Chapter 1

Introduction to Statistics and Probability

1.1 Overview: Statistical Inference, Samples, Populations, and the Role of Probability

Beginning in the 1980s and continuing into the 21st century, a great deal of attention has been focused on *improvement of quality* in American industry. Much has been said and written about the Japanese "industrial miracle," which began in the middle of the 20th century. The Japanese were able to succeed where we and other countries had failed—namely, to create an atmosphere that allows the production of high-quality products. Much of the success of the Japanese has been attributed to the use of *statistical methods* and statistical thinking among management personnel.

Use of Scientific Data

The use of statistical methods in manufacturing, development of food products, computer software, energy sources, pharmaceuticals, and many other areas involves the gathering of information or **scientific data**. Of course, the gathering of data is nothing new. It has been done for well over a thousand years. Data have been collected, summarized, reported, and stored for perusal. However, there is a profound distinction between collection of scientific information and **inferential statistics**. It is the latter that has received rightful attention in recent decades.

The offspring of inferential statistics has been a large "toolbox" of statistical methods employed by statistical practitioners. These statistical methods are designed to contribute to the process of making scientific judgments in the face of **uncertainty** and **variation**. The product density of a particular material from a manufacturing process will not always be the same. Indeed, if the process involved is a batch process rather than continuous, there will be not only variation in material density among the batches that come off the line (batch-to-batch variation), but also within-batch variation. Statistical methods are used to analyze data from a process such as this one in order to gain more sense of where in the process changes may be made to improve the **quality** of the process. In this process, qual-

ity may well be defined in relation to closeness to a target density value in harmony with *what portion of the time* this closeness criterion is met. An engineer may be concerned with a specific instrument that is used to measure sulfur monoxide in the air during pollution studies. If the engineer has doubts about the effectiveness of the instrument, there are two **sources of variation** that must be dealt with. The first is the variation in sulfur monoxide values that are found at the same locale on the same day. The second is the variation between values observed and the **true** amount of sulfur monoxide that is in the air at the time. If either of these two sources of variation is exceedingly large (according to some standard set by the engineer), the instrument may need to be replaced. In a biomedical study of a new drug that reduces hypertension, 85% of patients experienced relief, while it is generally recognized that the current drug, or "old" drug, brings relief to 80% of patients that have chronic hypertension. However, the new drug is more expensive to make and may result in certain side effects. Should the new drug be adopted? This is a problem that is encountered (often with much more complexity) frequently by pharmaceutical firms in conjunction with the FDA (Federal Drug Administration). Again, the consideration of variation needs to be taken into account. The "85%" value is based on a certain number of patients chosen for the study. Perhaps if the study were repeated with new patients the observed number of "successes" would be 75%! It is the natural variation from study to study that must be taken into account in the decision process. Clearly this variation is important, since variation from patient to patient is endemic to the problem.

Variability in Scientific Data

In the problems discussed above the statistical methods used involve dealing with variability, and in each case the variability to be studied is that encountered in scientific data. If the observed product density in the process were always the same and were always on target, there would be no need for statistical methods. If the device for measuring sulfur monoxide always gives the same value and the value is accurate (i.e., it is correct), no statistical analysis is needed. If there were no patient-to-patient variability inherent in the response to the drug (i.e., it either always brings relief or not), life would be simple for scientists in the pharmaceutical firms and FDA and no statistician would be needed in the decision process. Statistics researchers have produced an enormous number of analytical methods that allow for analysis of data from systems like those described above. This reflects the true nature of the science that we call inferential statistics, namely, using techniques that allow us to go beyond merely reporting data to drawing conclusions (or inferences) about the scientific system. Statisticians make use of fundamental laws of probability and statistical inference to draw conclusions about scientific systems. Information is gathered in the form of **samples**, or collections of **observations**. The process of sampling will be introduced in this chapter, and the discussion continues throughout the entire book.

Samples are collected from **populations**, which are collections of all individuals or individual items of a particular type. At times a population signifies a scientific system. For example, a manufacturer of computer boards may wish to eliminate defects. A sampling process may involve collecting information on 50 computer boards sampled randomly from the process. Here, the population is all

computer boards manufactured by the firm over a specific period of time. If an improvement is made in the computer board process and a second sample of boards is collected, any conclusions drawn regarding the effectiveness of the change in process should extend to the entire population of computer boards produced under the "improved process." In a drug experiment, a sample of patients is taken and each is given a specific drug to reduce blood pressure. The interest is focused on drawing conclusions about the population of those who suffer from hypertension.

Often, it is very important to collect scientific data in a systematic way, with planning being high on the agenda. At times the planning is, by necessity, quite limited. We often focus only on certain properties or characteristics of the items or objects in the population. Each characteristic has particular engineering or, say, biological importance to the "customer," the scientist or engineer who seeks to learn about the population. For example, in one of the illustrations above the quality of the process had to do with the product density of the output of a process. An engineer may need to study the effect of process conditions, temperature, humidity, amount of a particular ingredient, and so on. He or she can systematically move these **factors** to whatever levels are suggested according to whatever prescription or **experimental design** is desired. However, a forest scientist who is interested in a study of factors that influence wood density in a certain kind of tree cannot necessarily design an experiment. This case may require an **observational study** in which data are collected in the field but **factor levels** can not be preselected. Both of these types of studies lend themselves to methods of statistical inference. In the former, the quality of the inferences will depend on proper planning of the experiment. In the latter, the scientist is at the mercy of what can be gathered. For example, it is sad if an agronomist is interested in studying the effect of rainfall on plant yield and the data are gathered during a drought.

The importance of statistical thinking by managers and the use of statistical inference by scientific personnel is widely acknowledged. Research scientists gain much from scientific data. Data provide understanding of scientific phenomena. Product and process engineers learn a great deal in their off-line efforts to improve the process. They also gain valuable insight by gathering production data (on-line monitoring) on a regular basis. This allows them to determine necessary modifications in order to keep the process at a desired level of quality.

There are times when a scientific practitioner wishes only to gain some sort of summary of a set of data represented in the sample. In other words, inferential statistics is not required. Rather, a set of single-number statistics or **descriptive statistics** is helpful. These numbers give a sense of center of the location of the data, variability in the data, and the general nature of the distribution of observations in the sample. Though no specific statistical methods leading to **statistical inference** are incorporated, much can be learned. At times, descriptive statistics are accompanied by graphics. Modern statistical software packages allow for computation of **means**, **medians**, **standard deviations**, and other single-number statistics as well as production of graphs that show a "footprint" of the nature of the sample, including histograms, stem-and-leaf plots, scatter plots, dot plots, and box plots.

The Role of Probability

From this chapter to Chapter 3, we deal with fundamental notions of probability. A thorough grounding in these concepts allows the reader to have a better understanding of statistical inference. Without some formalism of probability theory, the student cannot appreciate the true interpretation from data analysis through modern statistical methods. It is quite natural to study probability prior to studying statistical inference. Elements of probability allow us to quantify the strength or "confidence" in our conclusions. In this sense, concepts in probability form a major component that supplements statistical methods and helps us gauge the strength of the statistical inference. The discipline of probability, then, provides the transition between descriptive statistics and inferential methods. Elements of probability allow the conclusion to be put into the language that the science or engineering practitioners require. An example follows that will enable the reader to understand the notion of a P-value, which often provides the "bottom line" in the interpretation of results from the use of statistical methods.

Example 1.1: Suppose that an engineer encounters data from a manufacturing process in which 100 items are sampled and 10 are found to be defective. It is expected and anticipated that occasionally there will be defective items. Obviously these 100 items represent the sample. However, it has been determined that in the long run, the company can only tolerate 5% defective in the process. Now, the elements of probability allow the engineer to determine how conclusive the sample information is regarding the nature of the process. In this case, the **population** conceptually represents all possible items from the process. Suppose we learn that *if the process is acceptable*, that is, if it does produce items no more than 5% of which are defective, there is a probability of 0.0282 of obtaining 10 or more defective items in a random sample of 100 items from the process. This small probability suggests that the process does, indeed, have a long-run rate of defective items that exceeds 5%. In other words, under the condition of an acceptable process, the sample information obtained would rarely occur. However, it did occur! Clearly, though, it would occur with a much higher probability if the process defective rate exceeded 5% by a significant amount. ∎

From this example it becomes clear that the elements of probability aid in the translation of sample information into something conclusive or inconclusive about the scientific system. In fact, what was learned likely is alarming information to the engineer or manager. Statistical methods, which we will actually detail in Chapter 6, produced a P-value of 0.0282. The result suggests that the process **very likely is not acceptable**. The concept of a ***P*-value** is dealt with at length in succeeding chapters. The example that follows provides a second illustration.

Example 1.2: Often the nature of the scientific study will dictate the role that probability and deductive reasoning play in statistical inference. Exercise 5.28 on page 221 provides data associated with a study conducted at Virginia Tech on the development of a relationship between the roots of trees and the action of a fungus. Minerals are transferred from the fungus to the trees and sugars from the trees to the fungus. Two samples of 10 northern red oak seedlings were planted in a greenhouse, one containing seedlings treated with nitrogen and the other containing seedlings with

no nitrogen. All other environmental conditions were held constant. All seedlings contained the fungus *Pisolithus tinctorus*. More details are supplied in Chapter 5. The stem weights in grams were recorded after the end of 140 days. The data are given in Table 1.1.

Table 1.1: Data Set for Example 1.2

No Nitrogen	Nitrogen
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

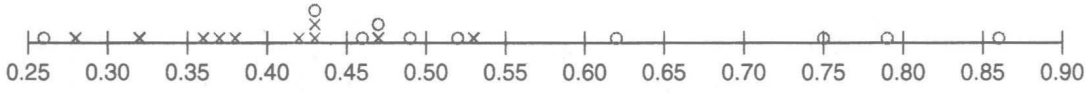


Figure 1.1: A dot plot of stem weight data.

In this example there are two samples from two **separate populations**. The purpose of the experiment is to determine if the use of nitrogen has an influence on the growth of the roots. The study is a comparative study (i.e., we seek to compare the two populations with regard to a certain important characteristic). It is instructive to plot the data as shown in the dot plot of Figure 1.1. The ○ values represent the "nitrogen" data and the × values represent the "no-nitrogen" data.

Notice that the general appearance of the data might suggest to the reader that, on average, the use of nitrogen increases the stem weight. Four nitrogen observations are considerably larger than any of the no-nitrogen observations. Most of the no-nitrogen observations appear to be below the center of the data. The appearance of the data set would seem to indicate that nitrogen is effective. But how can this be quantified? How can all of the apparent visual evidence be summarized in some sense? As in the preceding example, the fundamentals of probability can be used. The conclusions may be summarized in a probability statement or P-value. We will not show here the statistical inference that produces the summary probability. As in Example 1.1, these methods will be discussed in Chapter 6. The issue revolves around the "probability that data like these could be observed" *given that nitrogen has no effect*, in other words, given that both samples were generated from the same population. Suppose that this probability is small, say 0.03. That would certainly be strong evidence that the use of nitrogen does indeed influence (apparently increases) average stem weight of the red oak seedlings. ∎

How Do Probability and Statistical Inference Work Together?

It is important for the reader to understand the clear distinction between the discipline of probability, a science in its own right, and the discipline of inferential statistics. As we have already indicated, the use or application of concepts in probability allows real-life interpretation of the results of statistical inference. As a result, it can be said that statistical inference makes use of concepts in probability. One can glean from the two examples above that the sample information is made available to the analyst and, with the aid of statistical methods and elements of probability, conclusions are drawn about some feature of the population (the process does not appear to be acceptable in Example 1.1, and nitrogen does appear to influence average stem weights in Example 1.2). Thus for a statistical problem, **the sample along with inferential statistics allows us to draw conclusions about the population, with inferential statistics making clear use of elements of probability**. This reasoning is *inductive* in nature. Now as we move into Section 1.4 and beyond, the reader will note that, unlike what we do in our two examples here, we will not focus on solving statistical problems. Many examples will be given in which no sample is involved. There will be a population clearly described with all features of the population known. Then questions of importance will focus on the nature of data that might hypothetically be drawn from the population. Thus, one can say that **elements in probability allow us to draw conclusions about characteristics of hypothetical data taken from the population, based on known features of the population**. This type of reasoning is *deductive* in nature. Figure 1.2 shows the fundamental relationship between probability and inferential statistics.

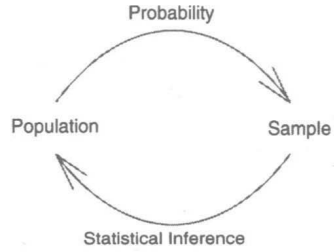


Figure 1.2: Fundamental relationship between probability and inferential statistics.

Now, in the grand scheme of things, which is more important, the field of probability or the field of statistics? They are both very important and clearly are complementary. The only certainty concerning the pedagogy of the two disciplines lies in the fact that if statistics is to be taught at more than merely a "cookbook" level, then the discipline of probability must be taught first. This rule stems from the fact that nothing can be learned about a population from a sample until the analyst learns the rudiments of uncertainty in that sample. For example, consider Example 1.1. The question centers around whether or not the population, defined by the process, is no more than 5% defective. In other words, the conjecture is that **on the average** 5 out of 100 items are defective. Now, the sample contains 100 items and 10 are defective. Does this support the conjecture or refute it? On the