Probability & Statistics
for Engineers & Scientists

NINTH EDITION

Walpole • Myers • Myers • Ye

ALWAYS LEARNING

PEARSON

Probability & Statistics for Engineers & Scientists

NINTH EDITION
GLOBAL EDITION

Ronald E. Walpole
Roanoke College

Raymond H. Myers
Virginia Tech

Sharon L. Myers
Radford University

Keying Ye
University of Texas at San Antonio

PEARSON

Boston Columbus Hoboken Indianapolis New York San Francisco
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montréal Toronto
Delhi Mexico City São Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

Editorial Director: *Chris Hoag*
Editor in Chief: *Deirdre Lynch*
Acquisitions Editor: *Patrick Barbera*
Assistant Acquisitions Editor, Global Edition: *Aditee Agarwal*
Editorial Assistant: *Justin Billing*
Program Manager: *Chere Bemelmans*
Project Manager: *Christine Whitlock*
Program Management Team Lead: *Karen Wernholm*
Project Management Team Lead: *Peter Silvia*
Project Editor, Global Edition: *K.K. Neelakantan*
Senior Manufacturing Controller, Global Edition: *Trudy Kimber*
Media Production Manager, Global Edition: *Vikram Kumar*
Media Producer: *Aimee Thorne*
MathXL Content Manager: *Bob Carroll*
Product Marketing Manager: *Tiffany Bitzel*
Field Marketing Manager: *Evan St. Cyr*
Marketing Coordinator: *Brooke Smith*
Senior Author Support/Technology Specialist: *Joe Vetere*
Rights and Permissions Project Manager: *Gina Cheselka*
Procurement Specialist: *Carol Melville*
Associate Director of Design USHE EMSS/HSC/EDU: *Andrea Nix*
Program Design Lead: *Heather Scott*
Cover Image: Chill Chillz / Shutterstock.com

Pearson Education Limited
Edinburgh Gate
Harlow
Essex CM20 2JE
England

and Associated Companies throughout the world

Visit us on the World Wide Web at: www.pearsonglobaleditions.com

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

10 9 8 7 6 5 4 3 2 1

ISBN 10: 1292161361
ISBN 13: 9781292161365

Typeset by Aptara
Printed and bound in Italy by LEGO

This book is dedicated to

Billy and Julie
R.H.M. and S.L.M.

Limin, Carolyn and Emily
K.Y.

This page intentionally left blank

Contents

Preface

General Approach and Mathematical Level

Our emphasis in creating this edition is less on adding new material and more on providing clarity and deeper understanding. This objective was accomplished in part by including new end-of-chapter material that adds connective tissue between chapters. We affectionately call these comments at the end of the chapter "Pot Holes." They are very useful to remind students of the big picture and how each chapter fits into that picture, and they aid the student in learning about limitations and pitfalls that may result if procedures are misused. A deeper understanding of real-world use of statistics is made available through class projects, which were added in several chapters. These projects provide the opportunity for students alone, or in groups, to gather their own experimental data and draw inferences. In some cases, the work involves a problem whose solution will illustrate the meaning of a concept or provide an empirical understanding of an important statistical result. Some existing examples were expanded and new ones were introduced to create "case studies," in which commentary is provided to give the student a clear understanding of a statistical concept in the context of a practical situation.

In this edition, we continue to emphasize a balance between theory and applications. Calculus and other types of mathematical support (e.g., linear algebra) are used at about the same level as in previous editions. The coverage of analytical tools in statistics is enhanced with the use of calculus when discussion centers on rules and concepts in probability. Probability distributions and statistical inference are highlighted in Chapters 2 through 10. Linear algebra and matrices are very lightly applied in Chapters 11 through 15, where linear regression and analysis of variance are covered. Students using this text should have had the equivalent of one semester of differential and integral calculus. Linear algebra is helpful but not necessary so long as the section in Chapter 12 on multiple linear regression using matrix algebra is not covered by the instructor. As in previous editions, a large number of exercises that deal with real-life scientific and engineering applications are available to challenge the student. The many data sets associated with the exercises are available for download from the website http://www.pearsonglobaleditions.com/Walpole or in MyStatLab.

Summary of Changes

- We've added MyStatLab, a course management systems that delivers proven results in helping individual students succeed. MyStatLab provides engaging experiences that personalize, stimulate, and measure learning for each student.

To learn more about how MyStatLab combines proven learning applications with powerful assessment, visit www.mystatlab.com or contact your Pearson representative.

- Class projects were added in several chapters to provide a deeper understanding of the real-world use of statistics. Students are asked to produce or gather their own experimental data and draw inferences from these data.
- More case studies were added and others expanded to help students understand the statistical methods being presented in the context of a real-life situation.
- "Pot Holes" were added at the end of some chapters and expanded in others. These comments are intended to present each chapter in the context of the big picture and discuss how the chapters relate to one another. They also provide cautions about the possible misuse of statistical techniques MSL bullet.
- Chapter 1 has been enhanced to include more on single-number statistics as well as graphical techniques. New fundamental material on sampling and experimental design is presented.
- Examples added to Chapter 8 on sampling distributions are intended to motivate P-values and hypothesis testing. This prepares the student for the more challenging material on these topics that will be presented in Chapter 10.
- Chapter 12 contains additional development regarding the effect of a single regression variable in a model in which collinearity with other variables is severe.
- Chapter 15 now introduces material on the important topic of response surface methodology (RSM). The use of noise variables in RSM allows the illustration of mean and variance (dual response surface) modeling.
- The central composite design (CCD) is introduced in Chapter 15.
- More examples are given in Chapter 18, and the discussion of using Bayesian methods for statistical decision making has been enhanced.

Content and Course Planning

This text is designed for either a one- or a two-semester course. A reasonable plan for a one-semester course might include Chapters 1 through 10. This would result in a curriculum that concluded with the fundamentals of both estimation and hypothesis testing. Instructors who desire that students be exposed to simple linear regression may wish to include a portion of Chapter 11. For instructors who desire to have analysis of variance included rather than regression, the one-semester course may include Chapter 13 rather than Chapters 11 and 12. Chapter 13 features one-factor analysis of variance. Another option is to eliminate portions of Chapters 5 and/or 6 as well as Chapter 7. With this option, one or more of the discrete or continuous distributions in Chapters 5 and 6 may be eliminated. These distributions include the negative binomial, geometric, gamma, Weibull, beta, and log normal distributions. Other features that one might consider removing from a one-semester curriculum include maximum likelihood estimation, prediction, and/or tolerance limits in Chapter 9. A one-semester curriculum has built-in flexibility, depending on the relative interest of the instructor in regression, analysis of variance, experimental design, and response surface methods (Chapter 15). There are several discrete and continuous distributions (Chapters 5 and 6) that have applications in a variety of engineering and scientific areas.

Chapters 11 through 18 contain substantial material that can be added for the second semester of a two-semester course. The material on simple and multiple linear regression is in Chapters 11 and 12, respectively. Chapter 12 alone offers a substantial amount of flexibility. Multiple linear regression includes such "special topics" as categorical or indicator variables, sequential methods of model selection such as stepwise regression, the study of residuals for the detection of violations of assumptions, cross validation and the use of the PRESS statistic as well as C_p, and logistic regression. The use of orthogonal regressors, a precursor to the experimental design in Chapter 15, is highlighted. Chapters 13 and 14 offer a relatively large amount of material on analysis of variance (ANOVA) with fixed, random, and mixed models. Chapter 15 highlights the application of two-level designs in the context of full and fractional factorial experiments (2^k). Special screening designs are illustrated. Chapter 15 also features a new section on response surface methodology (RSM) to illustrate the use of experimental design for finding optimal process conditions. The fitting of a second order model through the use of a central composite design is discussed. RSM is expanded to cover the analysis of robust parameter design type problems. Noise variables are used to accommodate dual response surface models. Chapters 16, 17, and 18 contain a moderate amount of material on nonparametric statistics, quality control, and Bayesian inference.

Chapter 1 is an overview of statistical inference presented on a mathematically simple level. It has been expanded from the eighth edition to more thoroughly cover single-number statistics and graphical techniques. It is designed to give students a preliminary presentation of elementary concepts that will allow them to understand more involved details that follow. Elementary concepts in sampling, data collection, and experimental design are presented, and rudimentary aspects of graphical tools are introduced, as well as a sense of what is garnered from a data set. Stem-and-leaf plots and box-and-whisker plots have been added. Graphs are better organized and labeled. The discussion of uncertainty and variation in a system is thorough and well illustrated. There are examples of how to sort out the important characteristics of a scientific process or system, and these ideas are illustrated in practical settings such as manufacturing processes, biomedical studies, and studies of biological and other scientific systems. A contrast is made between the use of discrete and continuous data. Emphasis is placed on the use of models and the information concerning statistical models that can be obtained from graphical tools.

Chapters 2, 3, and 4 deal with basic probability as well as discrete and continuous random variables. Chapters 5 and 6 focus on specific discrete and continuous distributions as well as relationships among them. These chapters also highlight examples of applications of the distributions in real-life scientific and engineering studies. Examples, case studies, and a large number of exercises edify the student concerning the use of these distributions. Projects bring the practical use of these distributions to life through group work. Chapter 7 is the most theoretical chapter in the text. It deals with transformation of random variables and will likely not be used unless the instructor wishes to teach a relatively theoretical course. Chapter 8 contains graphical material, expanding on the more elementary set of graphical tools presented and illustrated in Chapter 1. Probability plotting is discussed and illustrated with examples. The very important concept of sampling distributions is presented thoroughly, and illustrations are given that involve the central limit theorem and the distribution of a sample variance under normal, independent (i.i.d.) sampling. The t and F distributions are introduced to motivate their use in chapters to follow. New material in Chapter 8 helps the student to visualize the importance of hypothesis testing, motivating the concept of a P-value.

Chapter 9 contains material on one- and two-sample point and interval estimation. A thorough discussion with examples points out the contrast between the different types of intervals—confidence intervals, prediction intervals, and tolerance intervals. A case study illustrates the three types of statistical intervals in the context of a manufacturing situation. This case study highlights the differences among the intervals, their sources, and the assumptions made in their development, as well as what type of scientific study or question requires the use of each one. A new approximation method has been added for the inference concerning a proportion. Chapter 10 begins with a basic presentation on the pragmatic meaning of hypothesis testing, with emphasis on such fundamental concepts as null and alternative hypotheses, the role of probability and the P-value, and the power of a test. Following this, illustrations are given of tests concerning one and two samples under standard conditions. The two-sample t-test with paired observations is also described. A case study helps the student to develop a clear picture of what interaction among factors really means as well as the dangers that can arise when interaction between treatments and experimental units exists. At the end of Chapter 10 is a very important section that relates Chapters 9 and 10 (estimation and hypothesis testing) to Chapters 11 through 16, where statistical modeling is prominent. It is important that the student be aware of the strong connection.

Chapters 11 and 12 contain material on simple and multiple linear regression, respectively. Considerably more attention is given in this edition to the effect that collinearity among the regression variables plays. A situation is presented that shows how the role of a single regression variable can depend in large part on what regressors are in the model with it. The sequential model selection procedures (forward, backward, stepwise, etc.) are then revisited in regard to this concept, and the rationale for using certain P-values with these procedures is provided. Chapter 12 offers material on nonlinear modeling with a special presentation of logistic regression, which has applications in engineering and the biological sciences. The material on multiple regression is quite extensive and thus provides considerable flexibility for the instructor, as indicated earlier. At the end of Chapter 12 is commentary relating that chapter to Chapters 14 and 15. Several features were added that provide a better understanding of the material in general. For example, the end-of-chapter material deals with cautions and difficulties one might encounter. It is pointed out that there are types of responses that occur naturally in practice (e.g. proportion responses, count responses, and several others) with which standard least squares regression should not be used because standard assumptions do not hold and violation of assumptions may induce serious errors. The suggestion is made that data transformation on the response may alleviate the problem in some cases. Flexibility is again available in Chapters 13 and 14, on the topic of analysis of variance. Chapter 13 covers one-factor ANOVA in the context of a completely randomized design. Complementary topics include tests on variances and multiple comparisons. Comparisons of treatments in blocks are highlighted, along with the topic of randomized complete blocks. Graphical methods are extended to ANOVA to aid the student in supplementing the formal inference with a pictorial type of inference that can aid scientists and engineers in presenting material. A new project is given in which students incorporate the appropriate randomization into each plan and use graphical techniques and P-values in reporting the results. Chapter 14 extends the material in Chapter 13 to accommodate two or more factors that are in a factorial structure. The ANOVA presentation in Chapter 14 includes work in both random and fixed effects models. Chapter 15 offers material associated with 2^k factorial designs; examples and case studies present the use of screening designs and special higher fractions of the 2^k. Two new and special features are the presentations of response surface methodology (RSM) and robust parameter design. These topics are linked in a case study that describes and illustrates a dual response surface design and analysis featuring the use of process mean and variance response surfaces.

Computer Software

Case studies, beginning in Chapter 8, 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 practice following graduation. Examples and case studies in the text are supplemented, where appropriate, by various types of residual plots, quantile plots, normal probability plots, and other plots. Such plots are particularly prevalent in Chapters 11 through 15.

Acknowledgments

We are indebted to those colleagues who reviewed the previous editions of this book and provided many helpful suggestions for this edition. 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 Schulz, *Millersville University*; Roxane Burrows, *Hocking College*; and Frank Chmely for ensuring the accuracy of this text.

We would like to thank the editorial and production services provided by numerous people from Pearson, especially editor in chief Deirdre Lynch, acquisitions editor Patrick Barbera, Project Manager Christine Whitlock, Editorial Assistant Justin Billing, and copyeditor Sally Lifland. Many useful comments and suggestions by proofreader Gail Magin are greatly appreciated. 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.

Acknowledgments for the Global Edition

Pearson would like to thank and acknowledge Neelesh Upadhye, *Indian Institute of Technology Madras*, Aneesh Kumar K., *Mahatma Gandhi College*, and Bindu P. P., *Government Arts and Science College*, for contributing to the Global Edition, and Abhishek K. Umrawal, *University of Delhi*, Olivia T.K. Choi, *The University of Hong Kong*, Mani Sankar, *East Point College of Engineering and Technology*, and Shalabh, *Indian Institute of Technology Kanpur*, for reviewing the Global Edition.

BREAKTHROUGH
To improving results

Get the Most Out of MyStatLab™

MyStatLab is the world's leading online resource for teaching and learning statistics. MyStatLab helps students and instructors improve results and provides engaging experiences and personalized learning for each student so learning can happen in any environment. Plus, it offers flexible and time-saving course management features to allow instructors to easily manage their classes while remaining in complete control, regardless of course format.

Personalized Support for Students

- MyStatLab comes with many learning resources–eText, applets, videos, and more–all designed to support your students as they progress through their course.

- The Adaptive Study Plan acts as a personal tutor, updating in real time based on student performance to provide personalized recommendations on what to work on next. With the new Companion Study Plan assignments, instructors can now assign the Study Plan as a prerequisite to a test or quiz, helping to guide students through concepts they need to master.

- Personalized Homework allows instructors to create homework assignments tailored to each student's specific needs, focused on just the topics they have not yet mastered.

Used by nearly 4 million students each year, the MyStatLab and MyMathLab family of products delivers consistent, measurable gains in student learning outcomes, retention, and subsequent course success.

Resources for Success

Instructor's Solutions Manual

The Instructor's Solutions Manual contains worked-out solutions to all text exercises and is available for download from Pearson Education's Instructor's Resource Center **(www.pearsonglobaleditions.com/walpole)** and in MyStatLab.

PowerPoint Slides

The PowerPoint slides include most of the figures and tables from the text. Slides are available to download from Pearson Education's Instructor Resource Center **(www.pearsonglobaleditions.com/walpole)** and in MyStatLab.

MyStatLab™ Online Course (access code required)

MyStatLab from Pearson is the world's leading online resource for teaching and learning statistics; it integrates interactive homework, assessment, and media in a flexible, easy to use format. MyStatLab is a course management system that helps individual students succeed. It provides engaging experiences that personalize, stimulate, and measure learning for each student. Tools are embedded to make it easy to integrate statistical software into the course. And, it comes from an experienced partner with educational expertise and an eye on the future.

MyStatLab leverages the power of the web-based statistical software, StatCrunch™, and includes access to **www.StatCrunch.com.** To learn more about how MyStatLab combines proven learning applications with powerful assessment, visit **www.mystatlab.com** or contact your Pearson representative.

Chapter 1

Introduction to Statistics and Data Analysis

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

Beginning in the 1980s and continuing into the 21st century, an inordinate amount 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, quality 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 is introduced in Chapter 2, 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. Definitions and illustrations of the single-number statistics and graphs, including histograms, stem-and-leaf plots, scatter plots, dot plots, and box plots, will be given in sections that follow.

The Role of Probability

In this book, Chapters 2 to 6 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 10, 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 9.40 on page 314 provides data associated with a study conducted at the Virginia Polytechnic Institute and State University 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 9. 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

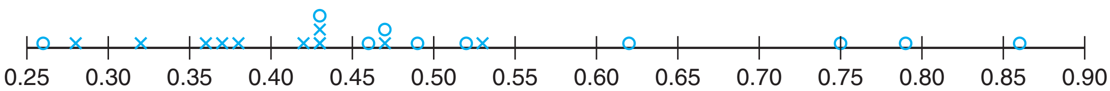

Figure 1.1: A dot plot of stem weight data.

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 $\circ$ values represent the "nitrogen" data and the $\times$ 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 10. 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 Chapter 2 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.

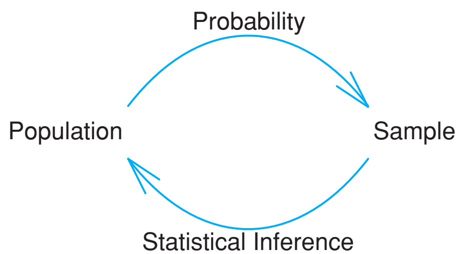

Figure 1.2: Fundamental relationship between probability and inferential statistics.

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 surface it would appear to be a refutation of the conjecture because 10 out of 100 seem to be “a bit much.” But without elements of probability, how do we know? Only through the study of material in future chapters will we learn the conditions under which the process is acceptable (5% defective). The probability of obtaining 10 or more defective items in a sample of 100 is 0.0282.

We have given two examples where the elements of probability provide a summary that the scientist or engineer can use as evidence on which to build a decision. The bridge between the data and the conclusion is, of course, based on foundations of statistical inference, distribution theory, and sampling distributions discussed in future chapters.

1.2 Sampling Procedures; Collection of Data

In Section 1.1 we discussed very briefly the notion of sampling and the sampling process. While sampling appears to be a simple concept, the complexity of the questions that must be answered about the population or populations necessitates that the sampling process be very complex at times. While the notion of sampling is discussed in a technical way in Chapter 8, we shall endeavor here to give some common-sense notions of sampling. This is a natural transition to a discussion of the concept of variability.

Simple Random Sampling

The importance of proper sampling revolves around the degree of confidence with which the analyst is able to answer the questions being asked. Let us assume that only a single population exists in the problem. Recall that in Example 1.2 two populations were involved. **Simple random sampling** implies that any particular sample of a specified *sample size* has the same chance of being selected as any other sample of the same size. The term **sample size** simply means the number of elements in the sample. Obviously, a table of random numbers can be utilized in sample selection in many instances. The virtue of simple random sampling is that it aids in the elimination of the problem of having the sample reflect a different (possibly more confined) population than the one about which inferences need to be made. For example, a sample is to be chosen to answer certain questions regarding political preferences in a certain state in the United States. The sample involves the choice of, say, 1000 families, and a survey is to be conducted. Now, suppose it turns out that random sampling is not used. Rather, all or nearly all of the 1000 families chosen live in an urban setting. It is believed that political preferences in rural areas differ from those in urban areas. In other words, the sample drawn actually confined the population and thus the inferences need to be confined to the “limited population,” and in this case confining may be undesirable. If, indeed, the inferences need to be made about the state as a whole, the sample of size 1000 described here is often referred to as a **biased sample**.

As we hinted earlier, simple random sampling is not always appropriate. Which alternative approach is used depends on the complexity of the problem. Often, for example, the sampling units are not homogeneous and naturally divide themselves into nonoverlapping groups that are homogeneous. These groups are called *strata*, and a procedure called *stratified random sampling* involves random selection of a sample *within* each stratum. The purpose is to be sure that each of the strata is neither over- nor underrepresented. For example, suppose a sample survey is conducted in order to gather preliminary opinions regarding a bond referendum that is being considered in a certain city. The city is subdivided into several ethnic groups which represent natural strata. In order not to disregard or overrepresent any group, separate random samples of families could be chosen from each group.

Experimental Design

The concept of randomness or random assignment plays a huge role in the area of **experimental design**, which was introduced very briefly in Section 1.1 and is an important staple in almost any area of engineering or experimental science. This will be discussed at length in Chapters 13 through 15. However, it is instructive to give a brief presentation here in the context of random sampling. A set of so-called **treatments** or **treatment combinations** becomes the populations to be studied or compared in some sense. An example is the nitrogen versus no-nitrogen treatments in Example 1.2. Another simple example would be "placebo" versus "active drug," or in a corrosion fatigue study we might have treatment combinations that involve specimens that are coated or uncoated as well as conditions of low or high humidity to which the specimens are exposed. In fact, there are four treatment or factor combinations (i.e., 4 populations), and many scientific questions may be asked and answered through statistical and inferential methods. Consider first the situation in Example 1.2. There are 20 diseased seedlings involved in the experiment. It is easy to see from the data themselves that the seedlings are different from each other. Within the nitrogen group (or the no-nitrogen group) there is considerable **variability** in the stem weights. This variability is due to what is generally called the **experimental unit**. This is a very important concept in inferential statistics, in fact one whose description will not end in this chapter. The nature of the variability is very important. If it is too large, stemming from a condition of excessive nonhomogeneity in experimental units, the variability will "wash out" any detectable difference between the two populations. Recall that in this case that did not occur.

The dot plot in Figure 1.1 and P-value indicated a clear distinction between these two conditions. What role do those experimental units play in the data-taking process itself? The common-sense and, indeed, quite standard approach is to assign the 20 seedlings or experimental units **randomly to the two treatments or conditions**. In the drug study, we may decide to use a total of 200 available patients, patients that clearly will be different in some sense. They are the experimental units. However, they all may have the same chronic condition for which the drug is a potential treatment. Then in a so-called **completely randomized design**, 100 patients are assigned randomly to the placebo and 100 to the active drug. Again, it is these experimental units within a group or treatment that produce the variability in data results (i.e., variability in the measured result), say blood pressure, or whatever drug efficacy value is important. In the corrosion fatigue study, the experimental units are the specimens that are the subjects of the corrosion.

Why Assign Experimental Units Randomly?

What is the possible negative impact of not randomly assigning experimental units to the treatments or treatment combinations? This is seen most clearly in the case of the drug study. Among the characteristics of the patients that produce variability in the results are age, gender, and weight. Suppose merely by chance the placebo group contains a sample of people that are predominately heavier than those in the treatment group. Perhaps heavier individuals have a tendency to have a higher blood pressure. This clearly biases the result, and indeed, any result obtained through the application of statistical inference may have little to do with the drug and more to do with differences in weights among the two samples of patients.

We should emphasize the attachment of importance to the term **variability**. Excessive variability among experimental units "camouflages" scientific findings. In future sections, we attempt to characterize and quantify measures of variability. In sections that follow, we introduce and discuss specific quantities that can be computed in samples; the quantities give a sense of the nature of the sample with respect to center of location of the data and variability in the data. A discussion of several of these single-number measures serves to provide a preview of what statistical information will be important components of the statistical methods that are used in future chapters. These measures that help characterize the nature of the data set fall into the category of **descriptive statistics**. This material is a prelude to a brief presentation of pictorial and graphical methods that go even further in characterization of the data set. The reader should understand that the statistical methods illustrated here will be used throughout the text. In order to offer the reader a clearer picture of what is involved in experimental design studies, we offer Example 1.3.

Example 1.3: A corrosion study was made in order to determine whether coating an aluminum metal with a corrosion retardation substance reduced the amount of corrosion. The coating is a protectant that is advertised to minimize fatigue damage in this type of material. Also of interest is the influence of humidity on the amount of corrosion. A corrosion measurement can be expressed in thousands of cycles to failure. Two levels of coating, no coating and chemical corrosion coating, were used. In addition, the two relative humidity levels are 20% relative humidity and 80% relative humidity.

The experiment involves four treatment combinations that are listed in the table that follows. There are eight experimental units used, and they are aluminum specimens prepared; two are assigned randomly to each of the four treatment combinations. The data are presented in Table 1.2.

The corrosion data are averages of two specimens. A plot of the averages is pictured in Figure 1.3. A relatively large value of cycles to failure represents a small amount of corrosion. As one might expect, an increase in humidity appears to make the corrosion worse. The use of the chemical corrosion coating procedure appears to reduce corrosion.

In this experimental design illustration, the engineer has systematically selected the four treatment combinations. In order to connect this situation to concepts with which the reader has been exposed to this point, it should be assumed that the

Table 1.2: Data for Example 1.3

Coating	**Humidity**	**Average Corrosion in Thousands of Cycles to Failure**
Uncoated	20%	975
	80%	350
Chemical Corrosion	20%	1750
	80%	1550

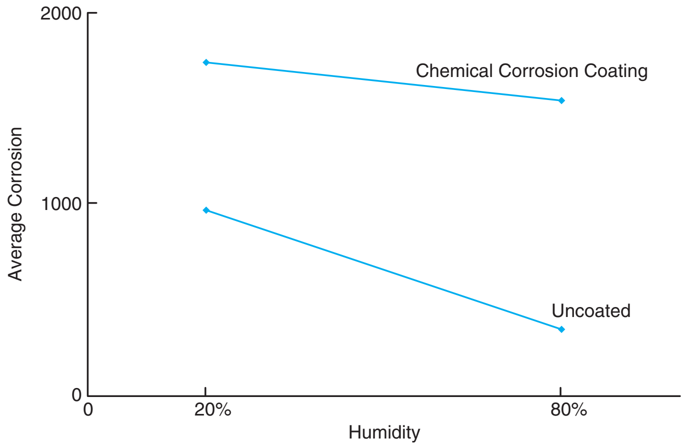

Figure 1.3: Corrosion results for Example 1.3.

conditions representing the four treatment combinations are four separate populations and that the two corrosion values observed for each population are important pieces of information. The importance of the average in capturing and summarizing certain features in the population will be highlighted in Section 1.3. While we might draw conclusions about the role of humidity and the impact of coating the specimens from the figure, we cannot truly evaluate the results from an analytical point of view without taking into account the *variability around* the average. Again, as we indicated earlier, if the two corrosion values for each treatment combination are close together, the picture in Figure 1.3 may be an accurate depiction. But if each corrosion value in the figure is an average of two values that are widely dispersed, then this variability may, indeed, truly "wash away" any information that appears to come through when one observes averages only. The foregoing example illustrates these concepts:

(1) random assignment of treatment combinations (coating, humidity) to experimental units (specimens)

(2) the use of sample averages (average corrosion values) in summarizing sample information

(3) the need for consideration of measures of variability in the analysis of any sample or sets of samples