PROBABILITY & STATISTICS

with R

for Engineers and Scientists

MICHAEL AKRITAS

Probability & Statistics with R

for Engineers and Scientists

Michael Akritas

The Pennsylvania State University

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

Editor in Chief: Deirdre Lynch
Acquisitions Editor: Patrick Barbera
Editorial Assistants: Salena Casha and Justin Billing
Program Manager: Tatiana Anacki
Project Manager: Mary Sanger
Program Management Team Lead: Marianne Stepanian
Project Management Team Lead: Christina Lepre
Media Producer: Jean Choe
Marketing Manager: Jeff Weidenaar
Marketing Assistant: Brooke Smith
Senior Author Support/Technology Specialist: Joe Vetere
Rights and Permissions Project Manager: Julia Tomasella
Procurement Specialist: Carol Melville
Associate Director of Design: Andrea Nix
Program Design Team Lead: Beth Paquin
Composition and Full-Service Project Management: Integra Software Services Pvt Ltd.
Cover Design: Cenveo Publisher Services
Cover Images: (photo) Derrrek/iStock/Getty Images; (R logo) Courtesy of the The R Foundation for Statistical Computing, http://www.r-project.org

Image credit: p. 446, The Art Gallery Collection/Alamy

Library of Congress Cataloging-in-Publication Data
Akritas, Michael G.
Probability and statistics with R for engineers and scientists / Michael G. Akritas, The Pennsylvania State University.—1st edition.
pages cm
ISBN 978-0-321-85299-1 – ISBN 0-321-85299-0
1. Engineering—Statistical methods—Data processing. 2. Science—Statistical methods—Data processing. 3. R (Computer program language) I. Title.
TA340.A38 2016
519.50285′5133–dc23 2014022502

1 2 3 4 5 6 7 8 9 10—EBM—18 17 16 15 14

PEARSON

www.pearsonhighered.com

ISBN-13: 978-0-321-85299-1
ISBN-10: 0-321-85299-0

To Martha, Niki, George, and Sophia

This page intentionally left blank

Contents

PREFACE

All men by nature desire knowledge

ARISTOTLE
Metaphysics, Book I

Knowledge, however meager, is usable
if we know the amount of
uncertainty in it

CR Rao

Statistics has become an integral part of scientific investigations in virtually all disciplines and is used extensively in industry and government organizations. *Probability & Statistics with R for Engineers and Scientists* offers a comprehensive introduction to the most commonly used statistical ideas and methods.

This book evolved from lecture notes for a one-semester course aimed mainly at undergraduate students in engineering and the natural sciences, as well as mathematics education majors and graduate students from various disciplines. The choice of examples, exercises, and data sets reflects the diversity of this audience.

The mathematical level has been kept relatively modest. Students who have completed one semester of differential and integral calculus should find almost all the exposition accessible. In particular, substantial use of calculus is made only in Chapters 3 and 4 and the third section of Chapter 6. Matrix algebra is used only in Chapter 12, which is usually not taught in a one-semester course.

THE R SOFTWARE PACKAGE

The widespread use of statistics is supported by a number of statistical software packages. Thus, modern courses on statistical methodology familiarize students with reading and interpreting software output. In sharp contrast to other books with the same intended audience, this book emphasizes not only the interpretation of software output, but also the *generation* of this output.

I decided to emphasize the software R (launched in 1984), which is sponsored by the Free Software Foundation. R is now used by the vast majority of statistics graduate students for thesis research, is a leader in new software development,[1] and is increasingly accepted in industry.[2] Moreover, R can be downloaded for free so students do not have to go to computer labs for their assignments. (To download R, go to the site http://www.R-project.org/ and follow the instructions.)

[1] See, e.g., http://www.r-bloggers.com/r-and-the-journal-of-computational-and-graphical-statistics.

[2] See the *New York Times* article “Data Analysts Captivated by R’s Power,” by Ashlee Vance, January 6, 2009.

TEACHING INNOVATIONS AND CHAPTER CONTENT

In addition to the use of a software package as an integral part of teaching probability and statistics, this book contains a number of other innovative approaches, reflecting the teaching philosophy that: (a) students should be intellectually challenged and (b) major concepts should be introduced as early as possible.

This text's major innovations occur in Chapters 1 and 4. Chapter 1 covers most of the important statistical concepts including sampling concepts, random variables, the population mean and variance for finite populations, the corresponding sample statistics, and basic graphics (histograms, stem and leaf plots, scatterplots, matrix scatterplots, pie charts and bar graphs). It goes on to introduce the notions of statistical experiments, comparative studies, and corresponding comparative graphics. The concepts and ideas underlying comparative studies, including main effects and interactions, are interesting in themselves, and their early introduction helps engage students in "statistical thinking."

Chapter 4, which deals with joint (mainly bivariate) distributions, covers the standard topics (marginal and conditional distributions, and independence of random variables), but also introduces the important concepts of covariance and correlation, along with the notion of a regression function. The simple linear regression model is discussed extensively, as it arises in the hierarchical model approach for defining the bivariate normal distribution.

Additional innovations are scattered throughout the rest of the chapters. Chapter 2 is devoted to the definition and basic calculus of probability. Except for the use of R to illustrate some concepts and the early introduction of probability mass function, this material is fairly standard. Chapter 3 gives a more general definition of the mean value and variance of a random variable and connects it to the simple definition given in Chapter 1. The common probability models for discrete and continuous random variables are discussed. Additional models commonly used in reliability studies are presented in the exercises. Chapter 5 discusses the distribution of sums and the Central Limit Theorem. The method of least squares, method of moments, and method of maximum likelihood are discussed in Chapter 6. Chapters 7 and 8 cover interval estimation and hypothesis testing, respectively, for the mean, median, and variance as well as the parameters of the simple linear regression model. Chapters 9 and 10 cover inference procedures for two and $k > 2$ samples, respectively, including paired data and randomized block designs. Nonparametric, or rank-based, inference is discussed alongside traditional methods of inference in Chapters 7 through 10. Chapter 11 is devoted to the analysis of two-factor, three-factor, and fractional factorial designs. Polynomial and multiple regression, and related topics such as weighted least squares, variable selection, multicollinearity, and logistic regression are presented in Chapter 12. The final chapter, Chapter 13, develops procedures used in statistical process control.

DATA SETS

This book contains both real life data sets, with identified sources, and simulated data sets. They can all be found at

www.pearsonhighered.com/akritas

Clicking on the name of a particular data set links to the corresponding data file. Importing data sets into R from the URL is easy when using the *read.table* command. As an example, you can import the data set BearsData.txt into the R data frame *br* by copying and pasting its URL into a read.table command:

```
br=read.table(”http://media.pearsoncmg.com/cmg/pmmg_mml_shared/
  mathstatsresources/Akritas/BearsData.txt”, header=T)
```

The data sets can also be downloaded to your computer and then imported into R from there.

Throughout the book, the *read.table* command will include only the name of the particular data set to be imported into R. For example, the command for importing the bear data into R will be given as

```
br=read.table(”BearsData.txt”, header=T)
```

SUGGESTED COVERAGE

This book has enough material for a year-long course, but can also be adapted for courses of one semester or two quarters. In a one-semester course, meeting three times a week, I cover selected topics from Chapters 1 through 10 and, recalling briefly the concepts of main effects and interaction (first introduced in Chapter 1), I finish the course by explaining the R commands and output for two-way analysis of variance. I typically deemphasize joint continuous distributions in Chapter 4 and may skip one or more of the following topics: multinomial distribution (Section 4.6.4), the method of maximum likelihood (Section 6.3.2), sign confidence intervals for the median (Section 7.3.4), the comparison of two variances (Section 9.4), the paired T test for proportions (Section 9.5.3), the Wilcoxon signed-rank test (Section 9.5.4), and the chi-square test for proportions (Section 10.2.3). It is possible to include material from Chapter 13 on statistical process control (for example after Chapter 8) by omitting additional material. One suggestion is to omit the section on comparing estimators (Section 6.4), confidence intervals and tests for a normal variance (Sections 7.3.5 and 8.3.6), and randomized block designs (Section 10.4).

ACKNOWLEDGMENTS

I greatly appreciate the support of the Department of Statistics at Penn State University and express my sincere thanks to colleagues, instructors, and graduate students who used various editions of the lecture notes and provided many suggestions for improvement over the years. I also thank all the people at Pearson for a highly professional and cordial collaboration through the various stages of production of the book. Special thanks go to Mary Sanger who supervised the last stages of production with an exceptionally high level of care and professionalism.

I am very grateful for numerous edits and substantive suggestions I received by the following reviewers: Keith Friedman, University of Texas at Austin; Steven T. Garren, James Madison University; Songfeng Zheng, Missouri State University; Roger Johnson, South Dakota School of Mines & Technology; Subhash Kochar, Portland State University; Michael Levine, Purdue University; Karin Reinhold, SUNY at Albany; Kingsley A. Reeves, Jr., University of South Florida; Katarina Jegdic, University of Houston Downtown; Lianming Wang, University of South Carolina; Lynne Butler, Haverford College; John Callister, Cornell University.

This page intentionally left blank

Chapter 1

Basic Statistical Concepts

1.1 Why Statistics?

Statistics deals with collecting, processing, summarizing, analyzing, and interpreting data. On the other hand, scientists and engineers deal with such diverse issues as the development of new products, effective use of materials and labor, solving production problems, quality improvement and reliability, and, of course, basic research. The usefulness of statistics as a tool for dealing with the above problems is best seen through some specific case studies mentioned in the following example.

Example 1.1-1

Examples of specific case studies arising in the sciences and engineering include

1. estimating the coefficient of thermal expansion of a metal;
2. comparing two methods of cloud seeding for hail and fog suppression at international airports;
3. comparing two or more methods of cement preparation in terms of compressive strength;
4. comparing the effectiveness of three cleaning products in removing four different types of stains;
5. predicting the failure time of a beam on the basis of stress applied;
6. assessing the effectiveness of a new traffic regulatory measure in reducing the weekly rate of accidents;
7. testing a manufacturer's claim regarding the quality of its product;
8. studying the relation between salary increases and employee productivity in a large corporation;
9. estimating the proportion of US citizens age 18 and over who are in favor of expanding solar energy sources; and
10. determining whether the content of lead in the water of a certain lake is within the safety limit. ■

The reason why tasks like the above require statistics is **variability**. Thus, if all cement prepared according to the same method had the same compressive strength, the task of comparing the different methods in case study 3 would not require statistics; it would suffice to compare the compressive strength of one cement specimen prepared from each method. However, the strength of different cement

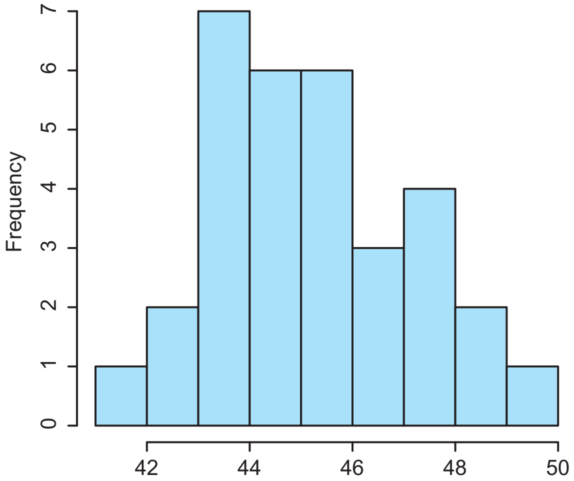


Figure 1-1 Histogram of 32 compressive strength measurements.

specimens prepared by the same method will, in general, differ. Figure 1-1 shows the histogram for 32 compressive strength measurements.[1] (See Section 1.5 for a discussion about histograms.) Similarly, if all beams fail at the same time under a given stress level, the prediction problem in case study 5 would not require statistics. A similar comment applies to all the case studies mentioned in Example 1.1-1.

An appreciation of the complications caused by variability begins by realizing that the problem of case study 3, as stated, is ambiguous. Indeed, if the hardness differs among preparations of the same cement mixture, then what does it mean to compare the hardness of different cement mixtures? A more precise statement of the problem would be to compare the *average* (or *mean*) hardness of the different cement mixtures. Similarly, the estimation problem in case study 1 is stated more precisely by referring to the average (or mean) thermal expansion.

It should also be mentioned that, due to variability, the familiar words *average* and *mean* have a technical meaning in statistics that can be made clear through the concepts of *population* and *sample*. These concepts are discussed in the next section.

1.2 Populations and Samples

As the examples of case studies mentioned in Example 1.1-1 indicate, statistics becomes relevant whenever the study involves the investigation of certain characteristic(s) of members (objects or subjects) in a certain **population** or populations. In statistics the word population is used to denote the set of all objects or subjects relevant to the particular study that are exposed to the same treatment or method. The members of a population are called **population units**.

Example 1.2-1

(a) In Example 1.1-1, case study 1, the characteristic under investigation is the thermal expansion of a metal in the population of all specimens of the particular metal.

(b) In Example 1.1-1, case study 3, we have two or more populations, one for each type of cement mixture, and the characteristic under investigation is compressive strength. Population units are the cement preparations.

(c) In Example 1.1-1, case study 5, the characteristic of interest is time to failure of a beam under a given stress level. Each stress level used in the study

[1] Compressive strength, in MPa (megapascal units), of test cylinders 6 in. in diameter by 12 in. high, using water/cement ratio of 0.4, measured on the 28th day after they were made.

corresponds to a separate population that consists of all beams that will be exposed to that stress level.

(d) In Example 1.1-1, case study 8, we have two characteristics, salary increase and productivity, for each subject in the population of employees of a large corporation. ■

In Example 1.2-1, part (c), we see that all populations consist of the same type of beams but are distinguished by the fact that beams of different populations will be exposed to different stress levels. Similarly, in Example 1.1-1, case study 2, the two populations consist of the same type of clouds distinguished by the fact that they will be seeded by different methods.

As mentioned in the previous section, the characteristic of interest varies among members of the same population. This is called the **inherent** or **intrinsic variability** of a population. A consequence of intrinsic variability is that complete, or *population-level*, understanding of characteristic(s) of interest requires a **census**, that is, examination of all members of the population. For example, full understanding of the relation between salary and productivity, as it applies to the population of employees of a large corporation (Example 1.1-1, case study 8), requires obtaining information on these two characteristics for all employees of the particular large corporation. Typically, however, census is not conducted due to cost and time considerations.

Example 1.2-2

(a) Cost and time considerations make it impractical to conduct a census of all US citizens age 18 and over in order to determine the proportion of these citizens who are in favor of expanding solar energy sources.

(b) Cost and time considerations make it impractical to analyze all the water in a lake in order to determine the lake's content of lead. ■

Moreover, census is often not feasible because the population is **hypothetical** or **conceptual**, in the sense that not all members of the population are available for examination.

Example 1.2-3

(a) If the objective is to study the quality of a product (as in Example 1.1-1, case studies 7 and 4), the relevant population consists not only of the available supply of this product, but also that which will be produced in the future. Thus, the relevant population is hypothetical.

(b) In a study aimed at reducing the weekly rate of accidents (Example 1.1-1, case study 6) the relevant population consists not only of the one-week time periods on which records have been kept, but also of future one-week periods. Thus, the relevant population is hypothetical. ■

In studies where it is either impractical or infeasible to conduct a census (which is the vast majority of cases), answers to questions regarding population-level properties/attributes of characteristic(s) under investigation are obtained by **sampling** the population. Sampling refers to the process of selecting a number of population units and recording their characteristic(s). For example, determination of the proportion of US citizens age 18 and over who are in favor of expanding solar energy sources is based on a sample of such citizens. Similarly, the determination of whether or not the content of lead in the water of a certain lake is within the safety limit must be based on water samples. The good news is that if the sample is suitably drawn from the population, then the **sample properties/attributes** of the characteristic of interest resemble (though they are not identical to) the **population properties/attributes**.

Example 1.2-4

(a) A *sample proportion* (i.e., the proportion in a chosen sample) of US citizens who favor expanding the use of solar energy approximates (but is, in general, different from) the *population proportion*. (Precise definitions of sample proportion and population proportion are given in Section 1.6.1.)

(b) The average concentration of lead in water samples (*sample average*) approximates (but is, in general, different from) the average concentration in the entire lake (*population average*). (Precise definitions of sample average and population average are given in Section 1.6.2.)

(c) The relation between salary and productivity manifested in a sample of employees approximates (but is, in general, different from) the relation in the entire population of employees of a large corporation. ■

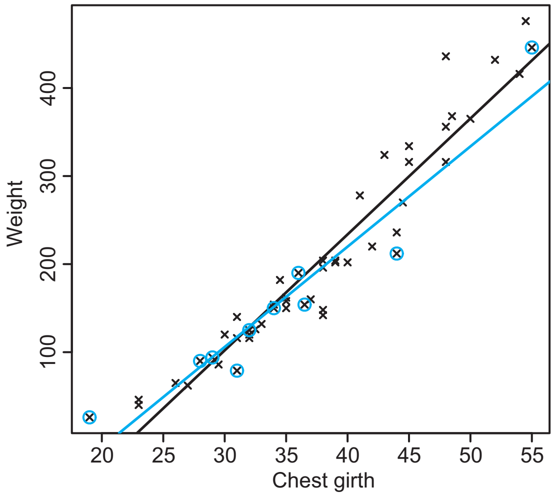
Example 1.2-5

The easier-to-measure chest girth of bears is often used to estimate the harder-to-measure weight. Chest girth and weight measurements for 50 bears residing in a given forested area are marked with "x" in Figure 1-2. The colored circles indicate the chest girth and weight measurements of the bears in a sample of size 10.[2] The black line captures the roughly linear relationship between chest girth and weight in the population of 50 black bears, while the colored line does the same for the sample.[3] It is seen that the relationship between chest girth and weight suggested by the sample is similar but not identical to that of the population.

■

Sample properties of the characteristic of interest also differ from sample to sample. This is another consequence of the intrinsic variability of the population from which samples are drawn. For example, the number of US citizens, in a sample of size 20, who favor expanding solar energy will (most likely) be different from the corresponding number in a different sample of 20 US citizens. (See also the examples in Section 1.6.2.) The term **sampling variability** is used to describe such differences in the characteristic of interest from sample to sample.

Figure 1-2 Population and sample relationships between chest girth (in) and weight (lb) of black bears.



[2] The sample was obtained by the method of *simple random sampling* described in Section 1.3.

[3] The lines were fitted by the method of *least squares* described in Chapter 6.

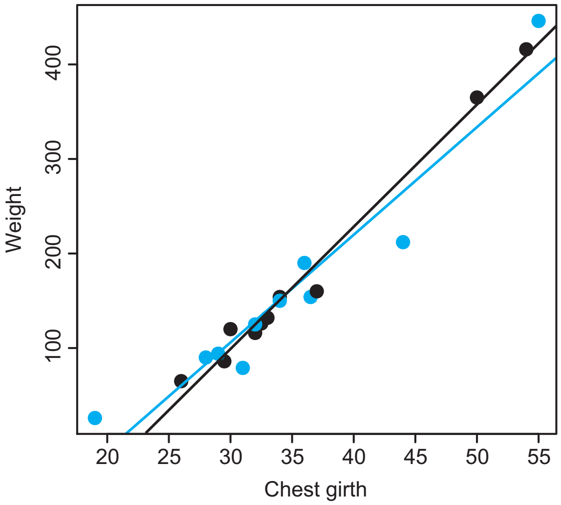


Figure 1-3 Variability in the relationships between chest girth and weight of black bears suggested by two different samples of size 10.

Example 1.2-6

As an illustration of sampling variability, a second sample of size 10 was taken from the population of 50 black bears described in Example 1.2-5. Figure 1-3 shows the chest girth and weight measurements for the original sample in colored dots while those for the second sample are shown in black dots. The sampling variability is demonstrated by the colored and black lines, which suggest somewhat different relationships between chest girth and weight, although both lines approximate the population relationship. ■

One must never lose sight of the fact that all scientific investigations aim at discovering the population-level properties/attributes of the characteristic(s) of interest. In particular, the problems in all the case studies mentioned in Example 1.1-1 refer to population-level properties. Thus, the technical meaning of the familiar word *average* (or *mean*), which was alluded to at the end of Section 1.1, is that of the population average (or mean); see Section 1.6.2 for a precise definition.

Population-level properties/attributes of characteristic(s) are called **population parameters**. Examples include the population mean (or average) and the population proportion that were referred to in Example 1.2-4. These and some additional examples of population parameters are defined in Sections 1.6 and 1.7. Further examples of population parameters, to be discussed in later chapters, include the *correlation coefficient* between two characteristics, e.g., between salary increase and productivity or between chest girth and weight. The corresponding sample properties/attributes of characteristics are called **statistics**, which is a familiar term because of its use in *sports statistics*. The sample mean (or average), sample proportion, and some additional statistics are defined in Sections 1.6 and 1.7, while further statistics are introduced in later chapters.

A sample can be thought of as a window that provides a glimpse into the population. However, due to sampling variability, a sample cannot yield accurate information regarding the population properties/attributes of interest. Using the new terminology introduced in the previous paragraph, this can be restated as: statistics approximate corresponding population parameters but are, in general, not equal to them.

Because only sample information is available, population parameters remain unknown. **Statistical inference** is the branch of statistics dealing with the uncertainty issues that arise in extrapolating to the population the information contained in the sample. Statistical inference helps decision makers choose actions in the absence of accurate knowledge about the population by

- assessing the accuracy with which statistics approximate corresponding population parameters; and
- providing an appraisal of the probability of making the wrong decision, or incorrect prediction.

For example, city officials might want to know whether a new industrial plant is pushing the average air pollution beyond the acceptable limits. Air samples are taken and the air pollution is measured in each sample. The sample average, or sample mean, of the air pollution measurements must then be used to decide if the overall (i.e., population-level) average air pollution is elevated enough to justify taking corrective action. In the absence of accurate knowledge, there is a risk that city officials might decide that the average air pollution exceeds the acceptable limit, when in fact it does not, or, conversely, that the average air pollution does not exceed the acceptable limit, when in fact it does.

As we will see in later chapters, statistical inference mainly takes the form of **estimation** (both **point** and, the more useful, **interval** estimation) of the population parameter(s) of interest, and of **testing** various **hypotheses** regarding the value of the population parameter(s) of interest. For example, estimation would be used in the task of estimating the average coefficient of thermal expansion of a metal (Example 1.1-1, case study 1), while the task of testing a manufacturer's claim regarding the quality of its product (Example 1.1-1, case study 7) involves hypothesis testing. Finally, the principles of statistical inference are also used in the problem of **prediction**, which arises, for example, if we would like to predict the failure time of a particular beam on the basis of the stress to which it will be exposed (Example 1.1-1, case study 5). The majority of the statistical methods presented in this book fall under the umbrella of statistical inference.

Exercises

1. A car manufacturer wants to assess customer satisfaction for cars sold during the previous year.

(a) Describe the population involved.

(b) Is the population involved hypothetical or not?

2. A field experiment is conducted to compare the yield of three varieties of corn used for biofuel. Each variety will be planted on 10 randomly selected plots and the yield will be measured at the time of harvest.

(a) Describe the population(s) involved.

(b) What is the characteristic of interest?

(c) Describe the sample(s).

3. An automobile assembly line is manned by two shifts a day. The first shift accounts for two-thirds of the overall production. Quality control engineers want to compare the average number of nonconformances per car in each of the two shifts.

(a) Describe the population(s) involved.

(b) Is (are) the population(s) involved hypothetical or not?

(c) What is the characteristic of interest?

4. A consumer magazine article titled "How Safe Is the Air in Airplanes" reports that the air quality, as quantified by the degree of staleness, was measured on 175 domestic flights.

(a) Identify the population of interest.

(b) Identify the sample.

(c) What is the characteristic of interest?

5. In an effort to determine the didactic benefits of computer activities when used as an integral part of a statistics course for engineers, one section is taught using the traditional method, while another is taught with computer activities. At the end of the semester, each student's score on the same test is recorded. To eliminate unnecessary variability, both sections were taught by the same professor.

(a) Is there one or two populations involved in the study?

(b) Describe the population(s) involved.

(c) Is (are) the population(s) involved hypothetical or not?

(d) What is (are) the sample(s) in this study?

1.3 Some Sampling Concepts

1.3.1 REPRESENTATIVE SAMPLES

Proper extrapolation of sample information to the population, that is, valid statistical inference, requires that the sample be **representative** of the population. For example, extrapolation of the information from a sample that consists of those who work in the oil industry to the population of US citizens will unavoidably lead to wrong conclusions about the prevailing public opinion regarding the use of solar energy.

A famous (or infamous) example that demonstrates what can go wrong when a non-representative sample is used is the *Literary Digest* poll of 1936. The magazine *Literary Digest* had been extremely successful in predicting the results in US presidential elections, but in 1936 it predicted a 3-to-2 victory for Republican Alf Landon over the Democratic incumbent Franklin Delano Roosevelt. The blunder was due to the use of a non-representative sample, which is discussed further in Section 1.3.4. It is worth mentioning that the prediction of the *Literary Digest* magazine was wrong even though it was based on 2.3 million responses (out of 10 million questionnaires sent). On the other hand, Gallup correctly predicted the outcome of that election by surveying only 50,000 people.

The notion of representativeness of a sample, though intuitive, is hard to pin down because there is no way to tell just by looking at a sample whether or not it is representative. Thus we adopt an indirect definition and say that a sample is representative if it leads to valid statistical inference. The only assurance that the sample will be representative comes from the method used to select the sample. Some of these sampling methods are discussed below.

1.3.2 SIMPLE RANDOM SAMPLING AND STRATIFIED SAMPLING

The most straightforward method for obtaining a representative sample is called **simple random sampling**. A sample of size n, selected from some population, is a simple random sample if the selection process ensures that every sample of size n has an equal chance of being selected. In particular, every member of the population has the same chance of being included in the sample.

A common way to select a simple random sample of size n from a finite population consisting of N units is to number the population units from $1, \ldots, N$, use a **random number generator** to randomly select n of these numbers, and form the sample from the units that correspond to the n selected numbers. A random number generator for selecting a simple random sample simulates the process of writing each number from $1, \ldots, N$ on slips of paper, putting the slips in a box, mixing them thoroughly, selecting one slip at random, and recording the number on the slip. The process is repeated (without replacing the selected slips in the box) until n distinct numbers from $1, \ldots, N$ have been selected.

Example 1.3-1

Sixty KitchenAid professional grade mixers are manufactured per day. Prior to shipping, a simple random sample of 12 must be selected from each day's production and carefully rechecked for possible defects.

(a) Describe a procedure for obtaining a simple random sample of 12 mixers from a day's production of 60 mixers.

(b) Use R to implement the procedure described in part (a).

Solution

As a first step we identify each mixer with a number from 1 to 60. Next, we write each number from 1 to 60 on separate, identical slips of paper, put all 60 slips of paper in a box, and mix them thoroughly. Finally, we select 12 slips from the box, one at a time and without replacement. The 12 numbers selected specify the desired sample of size $n = 12$ mixers from a day's production of 60. This process can be implemented in R with the command

Simple Random Sampling in R

```
y = sample(seq(1, 60), size = 12)        (1.3.1)
```

The command without the $y=$, that is, *sample(seq(1, 60), size = 12)*, will result in the 12 random numbers being typed in the R console; with the command as stated the random numbers are stored in the object y and can be seen by typing the letter "y." A set of 12 numbers thus obtained is 6, 8, 57, 53, 31, 35, 2, 4, 16, 7, 49, 41.

Clearly, the above technique cannot be used with hypothetical/infinite populations. However, measurements taken according to a set of well-defined instructions can assure that the essential properties of simple random sampling hold. For example, in comparing the compressive strength of cement mixtures, guidelines can be established for the mixture preparations and the measurement process to assure that the sample of measurements taken is representative.

As already mentioned, simple random sampling guarantees that every population unit has the same chance of being included in the sample. However, the mere fact that every population unit has the same chance of being included in the sample does not guarantee that the sampling process is simple random. This is illustrated in the following example.

Example 1.3-2

In order to select a representative sample of 10 from a group of 100 undergraduate students consisting of 50 male and 50 female students, the following sampling method is implemented: (a) assign numbers 1–50 to the male students and use a random number generator to select five of them; (b) repeat the same for the female students. Does this method yield a simple random sample of 10 students?

Solution

First note that the sampling method described guarantees that every student has the same chance (1 out of 10) of being selected. However, this sampling excludes all samples with unequal numbers of male and female students. For example, samples consisting of 4 male and 6 female students are excluded, that is, have zero chance of being selected. Hence, the condition for simple random sampling, namely, that each sample of size 10 has equal chance of being selected, is violated. It follows that the method described does not yield a simple random sample.

The sampling method of Example 1.3-2 is an example of what is called **stratified sampling**. Stratified sampling can be used whenever the population of interest consists of well-defined subgroups, or sub-populations, which are called **strata**. Examples of strata are ethnic groups, types of cars, age of equipment, different labs where water samples are sent for analysis, and so forth. Essentially, a stratified sample consists of simple random samples from each of the strata. A common method of choosing the within-strata sample sizes is to make the sample representation of each stratum equal to its population representation. This method of *proportionate allocation* is used in Example 1.3-2. Stratified samples are also representative, that is, they allow for valid statistical inference. In fact, if population units belonging to the same stratum tend to be more homogenous (i.e., similar) than population units belonging in different strata, then stratified sampling provides more accurate information regarding the entire population, and thus it is preferable.

1.3.3 SAMPLING WITH AND WITHOUT REPLACEMENT

In sampling from a finite population, one can choose to do the sampling **with replacement** or **without replacement**. Sampling with replacement means that after a unit is selected and its characteristic is recorded, it is replaced back into the population and may therefore be selected again. Tossing a fair coin can be thought of as sampling with replacement from the population {*Heads, Tails*}. In sampling without replacement, each unit can be included only once in the sample. Hence, simple random sampling is sampling without replacement.

It is easier to analyze the properties of a sample drawn with replacement because each selected unit is drawn from the same (the original) population of N units. (Whereas, in sampling without replacement, the second selection is drawn from a reduced population of $N - 1$ units, the third is drawn from a further reduced population of $N - 2$ units, and so forth.) On the other hand, including population unit(s) more than once (which is possible when sampling with replacement) clearly does not enhance the representativeness of the sample. Hence, the conceptual convenience of sampling with replacement comes with a cost, and, for this reason, it is typically avoided (but see the next paragraph). However, the cost is negligible when the population size is much larger than the sample size. This is because the likelihood of a unit being included twice in the sample is negligible, so that sampling with and without replacement are essentially equivalent. In such cases, we can pretend that a sample obtained by simple random sampling (i.e., without replacement) has the same properties as a sample obtained with replacement.

A major application of sampling with replacement occurs in the statistical method known by the name of **bootstrap**. Typically, however, this useful and widely used tool for statistical inference is not included in introductory textbooks.

1.3.4 NON-REPRESENTATIVE SAMPLING

Non-representative samples arise whenever the sampling plan is such that a part, or parts, of the population of interest are either excluded from, or systematically under-represented in, the sample.

Typical non-representative samples are the so-called **self-selected** and **convenience** samples. As an example of a self-selected sample, consider a magazine that conducts a reply-card survey of its readers, then uses information from cards that were returned to make statements like “80% of readers have purchased cellphones with digital camera capabilities.” In this case, readers who like to update and try new technology are more likely to respond indicating their purchases. Thus, the proportion of purchasers of cellphones with digital camera capabilities in the sample of returned cards will likely be much higher than it is amongst all readers. As an example of a convenience sample, consider using the students in your statistics class as a sample of students at your university. Note that this sampling plan excludes students from majors that do not require a statistics course. Moreover, most students take statistics in their sophomore or junior year and thus freshmen and seniors will be under-represented.

Perhaps the most famous historical example of a sampling blunder is the 1936 pre-election poll by the *Literary Digest* magazine. For its poll, the *Literary Digest* used a sample of 10 million people selected mainly from magazine subscribers, car owners, and telephone directories. In 1936, those who owned telephones or cars, or subscribed to magazines, were more likely to be wealthy individuals who were not happy with the Democratic incumbent. Thus, it was a convenience sample that excluded (or severely under-represented) parts of the population. Moveover, only 2.3 million responses were returned from the 10 million questionnaires that were sent. Obviously, those who felt strongly about the election were more likely to respond, and a majority of them wanted change. Thus, the *Literary Digest* sample was self-selected, in addition to being a sample of convenience. (The *Literary Digest* went bankrupt, while Gallup survived to make another blunder another day [in the 1948 Dewey-Truman contest].)

The term *selection bias* refers to the systematic exclusion or under-representation of some part(s) of the population of interest. Selection bias, which is inherent in self-selected and convenience samples, is the typical cause of non-representative samples. Simple random sampling and stratified sampling avoid selection bias. Other sampling methods that avoid selection bias do exist, and in some situations they can be less costly or easier to implement. But in this book we will mainly assume that the samples are simple random samples, with occasional passing reference to stratified sampling.

Exercises

1. The person designing the study of Exercise 5 in Section 1.2, aimed at determining the didactic benefits of computer activities, can make one of the two choices: (i) make sure that the students know which of the two sections will be taught with computer activities, so they can make an informed choice, or (ii) not make available any information regarding the teaching method of the two sections. Which of these two choices provides a closer approximation to simple random sampling?

2. A type of universal remote for home theater systems is manufactured in three distinct locations. Twenty percent of the remotes are manufactured in location A, 50% in location B, and 30% in location C. The quality control team (QCT) wants to inspect a simple random sample (SRS) of 100 remotes to see if a recently reported problem with the menu feature has been corrected. The QCT requests that each location send to the QC Inspection Facility a SRS of remotes from their recent production as follows: 20 from location A, 50 from B, and 30 from C.

(a) Does the sampling scheme described produce a simple random sample of size 100 from the recent production of remotes?

(b) Justify your answer in part (a). If you answer no, then what kind of sampling is it?

3. A civil engineering student, working on his thesis, plans a survey to determine the proportion of all current drivers in his university town that regularly use their seat belt. He decides to interview his classmates in the three classes he is currently enrolled.

(a) What is the population of interest?

(b) Do the student's classmates constitute a simple random sample from the population of interest?

(c) What name have we given to the sample that the student collected?

(d) Do you think that this sample proportion is likely to overestimate or underestimate the true proportion of all drivers who regularly use their seat belt?

4. In the Macworld Conference Expo Keynote Address on January 9, 2007, Steve Jobs announced a new product, the iPhone. A technology consultant for a consumer magazine wants to select 15 devices from the pilot lot of 70 iPhones to inspect feature coordination. Describe a method for obtaining a simple random sample of 15 from the lot of 70 iPhones. Use R to select a sample of 15. Give the R commands and the sample you obtained.

5. A distributor has just received a shipment of 90 drain pipes from a major manufacturer of such pipes. The distributor wishes to select a sample of size 5 to carefully inspect for defects. Describe a method for obtaining a simple random sample of 5 pipes from the shipment of 90 pipes. Use R to implement the method. Give the R commands and the sample you obtained.

6. A service agency wishes to assess its clients' views on quality of service over the past year. Computer records identify 1000 clients over the past 12 months, and a decision is made to select 100 clients to survey.

(a) Describe a procedure for selecting a simple random sample of 100 clients from last year's population of 1000 clients.

(b) The population of 1000 clients consists of 800 Caucasian-Americans, 150 African-Americans and 50 Hispanic-Americans. Describe an alternative procedure for selecting a representative random sample of size 100 from the population of 1000 clients.

(c) Give the R commands for implementing the sampling procedures described in parts (a) and (b).

7. A car manufacturer wants information about customer satisfaction for cars sold during the previous year. The particular manufacturer makes three different types of cars. Describe and discuss two different random sampling methods that might be employed.

8. A particular product is manufactured in two facilities, A and B. Facility B is more modern and accounts for 70% of the total production. A quality control engineer wishes to obtain a simple random sample of 50 from the entire production during the past hour. A coin is flipped and each time the flip results in heads, the engineer selects an item at random from those produced in facility A, and each time the flip results in tails, the engineer selects an item at random from those produced in facility B. Does this sampling scheme result in simple random sampling? Explain your answer.

9. An automobile assembly line operates for two shifts a day. The first shift accounts for two-thirds of the overall production. The task of quality control engineers is to monitor the number of nonconformances per car. Each day a simple random sample of 6 cars from the first shift, and a simple random sample of 3 cars from the second shift is taken, and the number of nonconformances per car is recorded. Does this sampling scheme produce a simple random sample of size 9 from the day's production? Justify your answer.

1.4 Random Variables and Statistical Populations

The characteristics of interest in all study examples given in Section 1.1 can be **quantitative** in the sense that they can be measured and thus can be expressed as numbers. Though quantitative characteristics are more common, **categorical**, including **qualitative**, characteristics also arise. Two examples of qualitative characteristics are gender and type of car, while strength of opinion is (ordinal) categorical. Since statistical procedures are applied on numerical data sets, numbers are assigned for expressing categorical characteristics. For example, −1 can be used to denote that a subject is male, and +1 to denote a female subject.

A characteristic of any type expressed as a number is called a **variable**. Categorical variables are a particular kind of **discrete** variables. Quantitative variables can also be discrete. For example, all variables expressing counts, such as the number in favor of a certain proposition, are discrete. Quantitative variables expressing measurements on a continuous scale, such as measurements of length, strength, weight, or time to failure, are examples of **continuous** variables. Finally, variables can be **univariate**, **bivariate**, or **multivariate** depending on whether one or two or more characteristics are measured, or recorded, on each population unit.

Example 1.4-1

(a) In a study aimed at determining the relation between productivity and salary increase, two characteristics are recorded on each population unit (productivity and salary increase), resulting in a bivariate variable.

(b) Consider the study that surveys US citizens age 18 and over regarding their opinion on solar energy. If an additional objective of the study is to determine how this opinion varies among different age groups, then the age of each individual in the sample is also recorded, resulting in a bivariate variable. If, in addition, the study aims to investigate how this opinion varies between genders, then the gender of each individual in the sample is also recorded, resulting in a multivariate variable.

(c) Consider the environmental study that measures the content of lead in water samples from a lake in order to determine if the concentration of lead exceeds the safe limits. If other contaminants are also of concern, then the content of these other contaminants is also measured in each water sample, resulting in a multivariate variable. ■

Due to the intrinsic variability, the value of the (possibly multivariate) variable varies among population units. It follows that when a population unit is randomly sampled from a population, its value is not known a priori. The value of the variable of a population unit that will be randomly sampled will be denoted by a capital letter, such as X. The fact that X is not known a priori justifies the term **random variable** for X.

> A random variable, X, denotes the value of the variable of a population unit that will be sampled.

The population from which a random variable was drawn will be called the **underlying population** of the random variable. Such terminology is particularly helpful in studies involving several populations, as are all studies that compare the performance of two or more methods or products; see, for example, case study 3 of Example 1.1-1.

Finally, we need a term for the entire collection of values that the variable under investigation takes among the units in the population. Stated differently, suppose that each unit in the population is labeled by the value of the variable under investigation, and the values in all labels are collected. This collection of values is called the **statistical population**. Note that if two (or more) population units have the same value of the variable, then this value appears two (or more) times in the statistical population.

Example 1.4-2

Consider the study that surveys US citizens age 18 and over regarding their opinion on solar energy. Suppose that the opinion is rated on the scale $0, 1, \ldots, 10$, and imagine each member of the population labeled by the value of their opinion. The statistical population contains as many 0's as there are people with opinion rated 0, as many 1's as there are people whose opinion is rated 1, and so forth. ■

The word "population" will be used to refer either to the population of units or to the statistical population. The context, or an explanation, will make clear which is the case.

In the above discussion, a random variable was introduced as the numerical outcome of random sampling from a (statistical) population. More generally, the concept of a random variable applies to the outcome of any action or process that generates a random numerical outcome. For example, the process of taking the arithmetic average of a simple random sample (see Section 1.6 for details) generates a random numerical outcome which, therefore, is a random variable.

Exercises

1. In a population of 500 tin plates, the number of plates with 0, 1, and 2 scratches is $N_0 = 190$, $N_1 = 160$, and $N_2 = 150$.

(a) Identify the variable of interest and the statistical population.

(b) Is the variable of interest quantitative or qualitative?

(c) Is the variable of interest univariate, bivariate, or multivariate?

2. Consider the following examples of populations, together with the variable/characteristic measured on each population unit.

(a) All undergraduate students currently enrolled at PSU. Variable: major.

(b) All campus restaurants. Variable: seating capacity.

(c) All books in Penn State libraries. Variable: frequency of check-out.

(d) All steel cylinders made in a given month. Variable: diameter.

For each of the above examples, describe the statistical population, state whether the variable of interest is quantitative or qualitative, and specify another variable that could be measured on the population units.

3. At the final assembly point of BMW cars in Graz, Austria, the car's engine and transmission arrive from Germany and France, respectively. A quality control inspector, visiting for the day, selects a simple random sample of *n* cars from the *N* cars available for inspection, and records the total number of engine and transmission nonconformances for each of the *n* cars.

(a) Is the variable of interest univariate, bivariate or multivariate?

(b) Is the variable of interest quantitative or qualitative?

(c) Describe the statistical population.

(d) Suppose the number of nonconformances in the engine and transmission are recorded separately for each car. Is the new variable univariate, bivariate, or multivariate?

4. In Exercise 4 in Section 1.2, a consumer magazine article reports that the air quality, as quantified by the degree of staleness, was measured on 175 domestic flights.

(a) Identify the variable of interest and the statistical population.

(b) Is the variable of interest quantitative or qualitative?

(c) Is the variable of interest univariate or multivariate?

5. A car manufacturing company that makes three different types of cars wants information about customer satisfaction for cars sold during the previous year. Each customer is asked for the type of car he or she bought last year and to rate his or her level of satisfaction on a scale from 1–6.

(a) Identify the variable recorded and the statistical population.

(b) Is the variable of interest univariate?

(c) Is the variable of interest quantitative or categorical?

1.5 Basic Graphics for Data Visualization

This section describes some of the most common graphics for data presentation and visualization. Additional graphics are introduced throughout this book.

1.5.1 HISTOGRAMS AND STEM AND LEAF PLOTS

Histograms and stem and leaf plots offer ways of organizing and displaying data. Histograms consist of dividing the range of the data into consecutive intervals, or *bins*, and constructing a box, or vertical bar, above each bin. The height of each box represents the bin's *frequency*, which is the number of observations that fall in the bin. Alternatively, the heights can be adjusted so the histogram's area (i.e., the total area defined by the boxes) equals one.

R will automatically choose the number of bins but it also allows user-specified intervals. Moreover, R offers the option of constructing a *smooth histogram*. Figure 1-4 shows a histogram (with area adjusted to one) of the Old Faithful geyser's eruption durations with a smooth histogram superimposed. (The data are from the R data frame *faithful*.)

Stem and leaf plots offer a somewhat different way for organizing and displaying data. They retain more information about the original data than histograms but do not offer as much flexibility in selecting the bins. The basic idea is to think of each observation as the *stem*, which consists of the beginning digit(s), and the *leaf*, which consists of the first of the remaining digits. In spite of different grouping of the observations, the stem and leaf display of the Old Faithful geyser's eruption durations shown in Figure 1-5 reveals a similar bimodal (i.e., having two modes or peaks) shape.

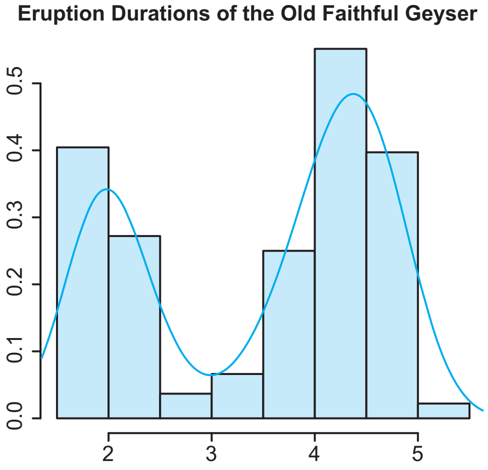


Figure 1-4 Histogram and smooth histogram for 272 eruption durations (min).

Figure 1-5 Stem and leaf plot for the 272 eruption durations.

```
16 | 070355555588
18 | 000022233333335577777777888822335777888
20 | 00002223378800035778
22 | 0002335578023578
24 | 00228
26 | 23
28 | 080
30 | 7
32 | 2337
34 | 250077
36 | 0000823577
38 | 2333335582225577
40 | 0000003357788888002233555577778
42 | 03335555778800233333555577778
44 | 02222335557780000000023333357778888
46 | 0000233357700000023578
48 | 00000022335800333
50 | 0370
```

With the R object x containing the data (e.g., $x =$ *faithful$eruptions*), the R commands for histograms and the stem and leaf plot are [# is the comment character]

R Commands for Histograms, Smooth Histograms, and Stem and Leaf Plots

```
hist(x) # basic frequency histogram

hist(x, freq=FALSE) # histogram area = 1

plot(density(x)) # basic smooth histogram

hist(x, freq=F); lines(density(x)) # superimposes
  the two

stem(x) # basic stem and leaf plot

stem(x, scale=1) # equivalent to the above
```

(1.5.1)

REMARK 1.5-1

1. The main label of a figure and the labels for the axes are controlled by *main* = ” ”, *xlab* = ” ”, *ylab* = ” ”, respectively; leaving a blank space between the quotes results in no labels. The color can also be specified. For example, the commands used for Figure 1-4 are *x* = *faithful$eruptions; hist(x, freq* = *F, main* = *”Eruption Durations of the Old Faithful Geyser”, xlab* = *” ”, col* = *”grey”); lines(density(x), col* = *”red”)*.
2. To override the automatic selection of bins one can either specify the number of bins (for example *breaks* = *6*), or specify explicitly the break points of the bins. Try *hist(faithful$eruptions, breaks* = *seq(1.2, 5.3, 0.41))*.
3. For additional control parameters type *?hist*, *?density*, or *?stem* on the R console. ◁

As an illustration of the role of the *scale* parameter in the stem command (whose default value is 1), consider the data on US beer production (in millions of barrels)

```
3 | 566699
4 | 11122444444
4 | 6678899
5 | 022334
5 | 5
```

for different quarters during the period 1975–1982. Entering the data in the R object *x* through *x* = *c(35, 36, 36, 36, 39, 39, 41, 41, 41, 42, 42, 44, 44, 44, 44, 44, 44, 46, 46, 47, 48, 48, 49, 49, 50, 52, 52, 53, 53, 54, 55)*, the command *stem(x, scale* = *0.5)* results in the above stem and leaf display. Note that leaves within each stem have been split into the low half (integers from 0 through 4) and the upper half (integers from 5 through 9).

1.5.2 SCATTERPLOTS

Scatterplots are useful for exploring the relationship between two and three variables. For example, Figures 1-2 and 1-3 show such scatterplots for the variables bear chest girth and bear weight for a population of black bears and a sample drawn from that population. These scatterplots suggested a fairly strong *positive association* between chest girth and weight (i.e., bigger chest girth suggests a heavier bear), so that chest girth can be used for predicting a bear’s weight. In this section we will see some enhanced versions of the basic scatterplot and a three-dimensional (3D) scatterplot.

Scatterplots with Subclass Identification The scatterplot in Figure 1-6 is similar to the scatterplot of Figure 1-2 but uses colors to distinguish between male and female bears. The additional insight gained from Figure 1-6 is that the relationship between the variables chest girth and weight is similar for both genders in that population of black bears.

Scatterplot Matrix As the name suggests, a scatterplot matrix is a matrix of scatterplots for all pairs of variables in a data set. In fact, two scatterplots are produced for every pair of variables, with each variable being plotted once on the *x*-axis and once on the *y*-axis. Figure 1-7 gives the matrix of all pairwise scatterplots between the

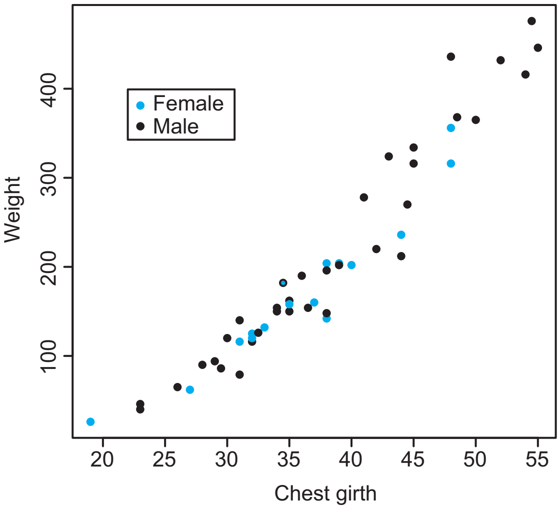


Figure 1-6 Bear weight vs chest girth scatterplot.

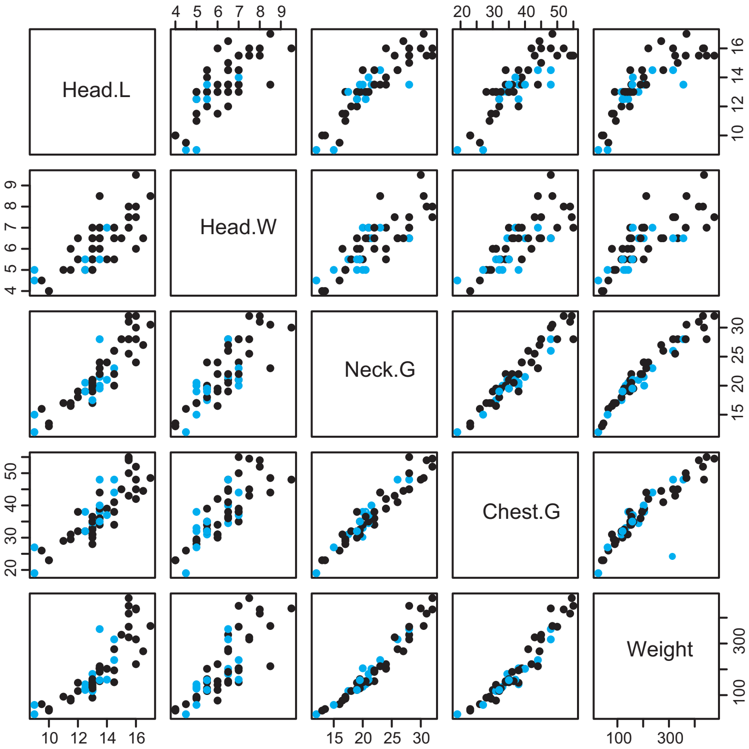


Figure 1-7 Scatterplot matrix for bear measurements.

different measurements taken on the black bears. The scatterplot in location (2,1), that is, in row 2 and column 1, has Head.L (head length) on the *x*-axis and Head.W (head width) on the *y*-axis, while the scatterplot in location (1,2) has Head.W on the *x*-axis and Head.L on the *y*-axis.

Scatterplot matrices are useful for identifying which variable serves as the best predictor for another variable. For example, Figure 1-7 suggests that a bear's chest girth and neck girth are the two best single predictors for a bear's weight.

With the data read into data frame *br* (for example by *br=read.table(”BearsData.txt”, header=T)*), the R commands that generated Figures 1-6 and 1-7 are:[4]

```
R Commands for Figures 1-6 and 1-7

attach(br) # so variables can be referred to by name

plot(Chest.G, Weight, pch=21, bg=c(”red”,
  ”green”)[unclass(Sex)]) # Figure 1-6

legend( x=22, y=400, pch=c(21, 21), col=c(”red”,
  ”green”), legend=c(”Female”, ”Male”)) # add legend in
  Figure 1-6

pairs(br[4:8], pch=21,bg=c(”red”, ”green”)[unclass(Sex)]) #
  Figure 1-7
```

Scatterplots with Marginal Histograms This enhancement of the basic scatterplot shows individual histograms for the two variables used in the scatterplot. Figure 1-8 shows such an enhancement for the scatterplot of Figure 1-6.[5] The term *marginal*, which is justified by the fact the histograms appear on the margins of the scatterplot, is commonly used to refer to the statistical population of individual variables in a multivariate data set; see also Chapter 4.

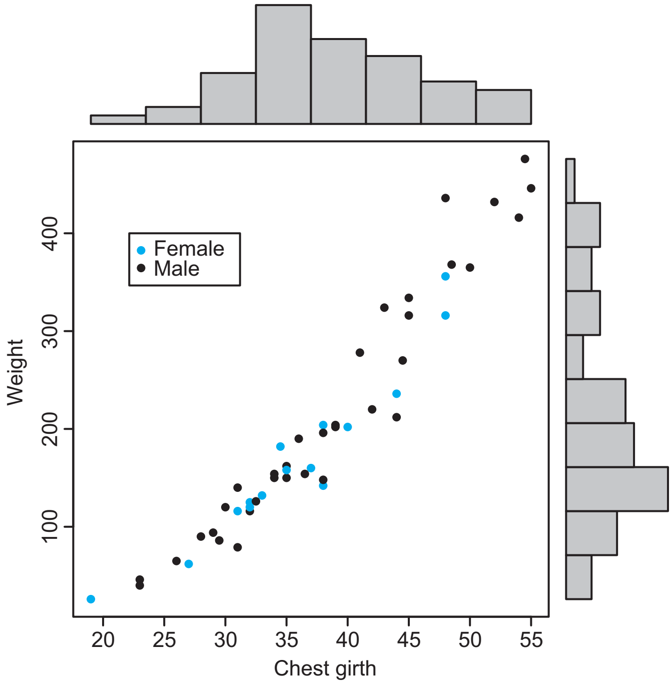


Figure 1-8 Scatterplot of bear weight vs chest girth showing the marginal histograms.

[4] Attempts to estimate a bear’s weight from its chest girth measurements go back to Neil F. Payne (1976). Estimating live weight of black bears from chest girth measurements, *The Journal of Wildlife Management*, 40(1): 167–169. The data used in Figure 1-7 is a subset of a data set contributed to Minitab by Dr. Gary Alt.

[5] The R commands that generated Figure 1-8 are given at http://www.stat.psu.edu/∼mga/401/fig/ScatterHist.txt; they are a variation of the commands given in an example on http://www.r-bloggers.com/example-8-41-scatterplot-with-marginal-histograms.