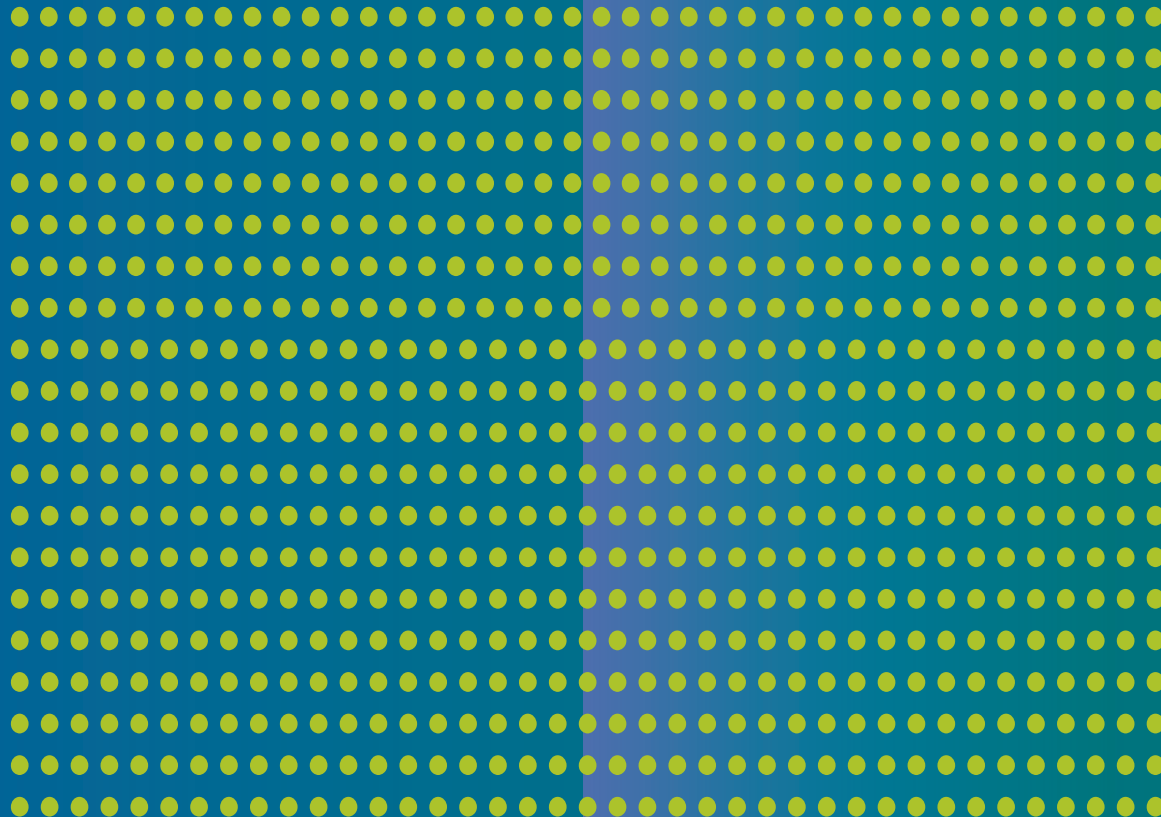


PEARSON NEW INTERNATIONAL EDITION

A Second Course in Statistics
Regression Analysis
William Mendenhall Terry Sincich
Seventh Edition



Pearson New International Edition

A Second Course in Statistics
Regression Analysis
William Mendenhall Terry Sincich
Seventh Edition

PEARSON®

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.pearsoned.co.uk

© Pearson Education Limited 2014

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without either the prior written permission of the publisher or a licence permitting restricted copying in the United Kingdom issued by the Copyright Licensing Agency Ltd, Saffron House, 6–10 Kirby Street, London EC1N 8TS.

All trademarks used herein are the property of their respective owners. The use of any trademark in this text does not vest in the author or publisher any trademark ownership rights in such trademarks, nor does the use of such trademarks imply any affiliation with or endorsement of this book by such owners.

PEARSON®

ISBN 10: 1-292-04290-7
ISBN 13: 978-1-292-04290-9

British Library Cataloguing-in-Publication Data

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

Printed in the United States of America

Table of Contents

1. A Review of Basic Concepts William Mendenhall/Terry Sincich	1
2. Introduction to Regression Analysis William Mendenhall/Terry Sincich	83
3. Simple Linear Regression William Mendenhall/Terry Sincich	95
4. Multiple Regression Models William Mendenhall/Terry Sincich	167
5. Principles of Model Building William Mendenhall/Terry Sincich	251
6. Variable Screening Methods William Mendenhall/Terry Sincich	317
7. Some Regression Pitfalls William Mendenhall/Terry Sincich	339
8. Residual Analysis William Mendenhall/Terry Sincich	369
9. Special Topics in Regression William Mendenhall/Terry Sincich	425
10. Introduction to Time Series Modeling and Forecasting William Mendenhall/Terry Sincich	481
11. Principles of Experimental Design William Mendenhall/Terry Sincich	537
12. The Analysis of Variance for Designed Experiments William Mendenhall/Terry Sincich	561

Appendix: Derivation of the Least Squares Estimates of ss_0 and ss_1 in Simple Linear Regression	
William Mendenhall/Terry Sincich	671
Appendix: The Mechanics of a Multiple Regression Analysis	
William Mendenhall/Terry Sincich	675
Appendix: A Procedure for Inverting a Matrix	
William Mendenhall/Terry Sincich	705
Appendix: Useful Statistical Tables	
William Mendenhall/Terry Sincich	711
Normal Curve Areas Table	
William Mendenhall/Terry Sincich	735
Critical Values for Student's t	
William Mendenhall/Terry Sincich	737
Index	739

A REVIEW OF BASIC CONCEPTS (OPTIONAL)

Contents

- | | | | |
|---|---|----|--|
| 1 | Statistics and Data | 7 | Sampling Distributions and the Central Limit Theorem |
| 2 | Populations, Samples, and Random Sampling | 8 | Estimating a Population Mean |
| 3 | Describing Qualitative Data | 9 | Testing a Hypothesis About a Population Mean |
| 4 | Describing Quantitative Data Graphically | 10 | Inferences About the Difference Between Two Population Means |
| 5 | Describing Quantitative Data Numerically | 11 | Comparing Two Population Variances |
| 6 | The Normal Probability Distribution | | |

Objectives

1. Review some basic concepts of sampling.
2. Review methods for describing both qualitative and quantitative data.
3. Review inferential statistical methods: confidence intervals and hypothesis tests.


Although we assume students have had a prerequisite introductory course in statistics, courses vary somewhat in content and in the manner in which they present statistical concepts. To be certain that we are starting with a common background, we use this chapter to review some basic definitions and concepts. Coverage is optional.

I Statistics and Data

According to *The Random House College Dictionary* (2001 ed.), statistics is “the science that deals with the collection, classification, analysis, and interpretation of numerical facts or data.” In short, statistics is the **science of data**—a science that will enable you to be proficient data producers and efficient data users.

Definition 1 **Statistics** is the science of data. This involves collecting, classifying, summarizing, organizing, analyzing, and interpreting data.

Data are obtained by measuring some characteristic or property of the objects (usually people or things) of interest to us. These objects upon which the measurements (or observations) are made are called **experimental units**, and the properties being measured are called **variables** (since, in virtually all studies of interest, the property varies from one observation to another).

Items within the text that are accompanied by  can be found on the companion website, www.pearsonhighered.com/mathstatsresources. The credit line below identifies the chapter under which the relevant files reside.

From Chapter 1 of *A Second Course in Statistics: Regression Analysis*, Seventh Edition, William Mendenhall, Terry Sincich. Copyright © 2012 by Pearson Education, Inc. Published by Pearson Prentice Hall. All rights reserved.

Definition 2 An **experimental unit** is an object (person or thing) upon which we collect data.

Definition 3 A **variable** is a characteristic (property) of the experimental unit with outcomes (data) that vary from one observation to the next.

All data (and consequently, the variables we measure) are either **quantitative** or **qualitative** in nature. Quantitative data are data that can be measured on a naturally occurring numerical scale. In general, qualitative data take values that are nonnumerical; they can only be classified into categories. The statistical tools that we use to analyze data depend on whether the data are quantitative or qualitative. Thus, it is important to be able to distinguish between the two types of data.

Definition 4 **Quantitative data** are observations measured on a naturally occurring numerical scale.

Definition 5 Nonnumerical data that can only be classified into one of a group of categories are said to be **qualitative data**.

Example
I

Chemical and manufacturing plants often discharge toxic waste materials such as DDT into nearby rivers and streams. These toxins can adversely affect the plants and animals inhabiting the river and the riverbank. The U.S. Army Corps of Engineers conducted a study of fish in the Tennessee River (in Alabama) and its three tributary creeks: Flint Creek, Limestone Creek, and Spring Creek. A total of 144 fish were captured, and the following variables were measured for each:

1. River/creek where each fish was captured
2. Number of miles upstream where the fish was captured
3. Species (channel catfish, largemouth bass, or smallmouth buffalofish)
4. Length (centimeters)
5. Weight (grams)
6. DDT concentration (parts per million)

The data are saved in the FISHDDT file. Data for 10 of the 144 captured fish are shown in Table 1.

- (a) Identify the experimental units.
- (b) Classify each of the five variables measured as quantitative or qualitative.

Solution

- (a) Because the measurements are made for each fish captured in the Tennessee River and its tributaries, the experimental units are the 144 captured fish.
- (b) The variables upstream that capture location, length, weight, and DDT concentration are quantitative because each is measured on a natural numerical scale: upstream in miles from the mouth of the river, length in centimeters, weight in grams, and DDT in parts per million. In contrast, river/creek and species cannot be measured quantitatively; they can only be classified into categories (e.g., channel catfish, largemouth bass, and smallmouth buffalofish for species). Consequently, data on river/creek and species are qualitative. ■



Table 1 Data collected by U.S. Army Corps of Engineers (selected observations)

River/Creek	Upstream	Species	Length	Weight	DDT
FLINT	5	CHANNELCATFISH	42.5	732	10.00
FLINT	5	CHANNELCATFISH	44.0	795	16.00
SPRING	1	CHANNELCATFISH	44.5	1133	2.60
TENNESSEE	275	CHANNELCATFISH	48.0	986	8.40
TENNESSEE	275	CHANNELCATFISH	45.0	1023	15.00
TENNESSEE	280	SMALLMOUTHBUFF	49.0	1763	4.50
TENNESSEE	280	SMALLMOUTHBUFF	46.0	1459	4.20
TENNESSEE	285	LARGEMOUTHBASS	25.0	544	0.11
TENNESSEE	285	LARGEMOUTHBASS	23.0	393	0.22
TENNESSEE	285	LARGEMOUTHBASS	28.0	733	0.80

I Exercises

- College application data.** Colleges and universities are requiring an increasing amount of information about applicants before making acceptance and financial aid decisions. Classify each of the following types of data required on a college application as quantitative or qualitative.

 - High school GPA
 - Country of citizenship
 - Applicant's score on the SAT or ACT
 - Gender of applicant
 - Parents' income
 - Age of applicant
- Fuel Economy Guide.** The data in the accompanying table were obtained from the *Model Year 2009 Fuel Economy Guide* for new automobiles.

 - Identify the experimental units.
 - State whether each of the variables measured is quantitative or qualitative.
- Ground motion of earthquakes.** In the *Journal of Earthquake Engineering* (November 2004), a team of civil and environmental engineers studied the ground motion characteristics of 15 earthquakes that occurred around the world between 1940 and 1995. Three (of many) variables measured on each earthquake were the type of ground motion (short, long, or forward directive), earthquake magnitude (Richter scale), and peak ground acceleration (feet per second). One of the goals of the study was to estimate the inelastic spectra of any ground motion cycle.

 - Identify the experimental units for this study.
 - Identify the variables measured as quantitative or qualitative.
- Use of herbal medicines.** *The American Association of Nurse Anesthetists Journal* (February 2000) published the results of a study on the use of herbal medicines before surgery. Each of 500

MODEL NAME	MFG	TRANSMISSION TYPE	ENGINE SIZE (LITERS)	NUMBER OF CYLINDERS	EST. CITY MILEAGE (MPG)	EST. HIGHWAY MILEAGE (MPG)
TSX	Acura	Automatic	2.4	4	21	30
Jetta	VW	Automatic	2.0	4	29	40
528i	BMW	Manual	3.0	6	18	28
Fusion	Ford	Automatic	3.0	6	17	25
Camry	Toyota	Manual	2.4	4	21	31
Escalade	Cadillac	Automatic	6.2	8	12	19

Source: *Model Year 2009 Fuel Economy Guide*, U.S. Dept. of Energy, U.S. Environmental Protection Agency (www.fueleconomy.gov).

surgical patients was asked whether they used herbal or alternative medicines (e.g., garlic, ginkgo, kava, fish oil) against their doctor's advice before surgery. Surprisingly, 51% answered "yes."

- (a) Identify the experimental unit for the study.
 - (b) Identify the variable measured for each experimental unit.
 - (c) Is the data collected quantitative or qualitative?
- 5 Drinking-water quality study.** *Disasters* (Vol. 28, 2004) published a study of the effects of a tropical cyclone on the quality of drinking water on a remote Pacific island. Water samples (size 500 milliliters) were collected approximately 4 weeks after Cyclone Ami hit the island. The following variables were recorded for each water sample. Identify each variable as quantitative or qualitative.
- (a) Town where sample was collected
 - (b) Type of water supply (river intake, stream, or borehole)

- (c) Acidic level (pH scale, 1–14)
- (d) Turbidity level (nephelometric turbidity units [NTUs])
- (e) Temperature (degrees Centigrade)
- (f) Number of fecal coliforms per 100 milliliters
- (g) Free chlorine-residual (milligrams per liter)
- (h) Presence of hydrogen sulphide (yes or no)

- 6 Accounting and Machiavellianism.** *Behavioral Research in Accounting* (January 2008) published a study of Machiavellian traits in accountants. *Machiavellian* describes negative character traits that include manipulation, cunning, duplicity, deception, and bad faith. A questionnaire was administered to a random sample of 700 accounting alumni of a large southwestern university. Several variables were measured, including age, gender, level of education, income, job satisfaction score, and Machiavellian ("Mach") rating score. What type of data (quantitative or qualitative) is produced by each of the variables measured?

2 Populations, Samples, and Random Sampling

When you examine a data set in the course of your study, you will be doing so because the data characterize a group of experimental units of interest to you. In statistics, the data set that is collected for all experimental units of interest is called a **population**. This data set, which is typically large, either exists in fact or is part of an ongoing operation and hence is conceptual. Some examples of statistical populations are given in Table 2.

Definition 6 A **population data set** is a collection (or set) of data measured on all experimental units of interest to you.

Many populations are too large to measure (because of time and cost); others cannot be measured because they are partly conceptual, such as the set of quality

Variable	Experimental Units	Population Data Set	Type
a. Starting salary of a graduating Ph.D. biologist	All Ph.D. biologists graduating this year	Set of starting salaries of all Ph.D. biologists who graduated this year	Existing
b. Breaking strength of water pipe in Philadelphia	All water pipe sections in Philadelphia	Set of breakage rates for all water pipe sections in Philadelphia	Existing
c. Quality of an item produced on an assembly line	All manufactured items	Set of quality measurements for all items manufactured over the recent past and in the future	Part existing, part conceptual
d. Sanitation inspection level of a cruise ship	All cruise ships	Set of sanitation inspection levels for all cruise ships	Existing

measurements (population c in Table 2). Thus, we are often required to select a subset of values from a population and to make **inferences** about the population based on information contained in a **sample**. This is one of the major objectives of modern statistics.

Definition 7 A **sample** is a subset of data selected from a population.

Definition 8 A **statistical inference** is an estimate, prediction, or some other generalization about a population based on information contained in a sample.

Example
2

According to the research firm Magnum Global (2008), the average age of viewers of the major networks' television news programming is 50 years. Suppose a cable network executive hypothesizes that the average age of cable TV news viewers is less than 50. To test her hypothesis, she samples 500 cable TV news viewers and determines the age of each.

- (a) Describe the population.
- (b) Describe the variable of interest.
- (c) Describe the sample.
- (d) Describe the inference.

Solution

- (a) The population is the set of units of interest to the cable executive, which is the set of all cable TV news viewers.
- (b) The age (in years) of each viewer is the variable of interest.
- (c) The sample must be a subset of the population. In this case, it is the 500 cable TV viewers selected by the executive.
- (d) The inference of interest involves the *generalization* of the information contained in the sample of 500 viewers to the population of all cable news viewers. In particular, the executive wants to estimate the average age of the viewers in order to determine whether it is less than 50 years. She might accomplish this by calculating the average age in the sample and using the sample average to estimate the population average. ■

Whenever we make an inference about a population using sample information, we introduce an element of uncertainty into our inference. Consequently, it is important to report the **reliability** of each inference we make. Typically, this is accomplished by using a probability statement that gives us a high level of confidence that the inference is true. In Example 2, we could support the inference about the average age of all cable TV news viewers by stating that the population average falls within 2 years of the calculated sample average with "95% confidence." (Throughout the text, we demonstrate how to obtain this measure of reliability—and its meaning—for each inference we make.)

Definition 9 A **measure of reliability** is a statement (usually quantified with a probability value) about the degree of uncertainty associated with a statistical inference.

The level of confidence we have in our inference, however, will depend on how **representative** our sample is of the population. Consequently, the sampling procedure plays an important role in statistical inference.

Definition 10 A **representative sample** exhibits characteristics typical of those possessed by the population.

The most common type of sampling procedure is one that gives every different sample of fixed size in the population an equal probability (chance) of selection. Such a sample—called a **random sample**—is likely to be representative of the population.

Definition 11 A **random sample** of n experimental units is one selected from the population in such a way that every different sample of size n has an equal probability (chance) of selection.

How can a random sample be generated? If the population is not too large, each observation may be recorded on a piece of paper and placed in a suitable container. After the collection of papers is thoroughly mixed, the researcher can remove n pieces of paper from the container; the elements named on these n pieces of paper are the ones to be included in the sample. Lottery officials utilize such a technique in generating the winning numbers for Florida's weekly 6/52 Lotto game. Fifty-two white ping-pong balls (the population), each identified from 1 to 52 in black numerals, are placed into a clear plastic drum and mixed by blowing air into the container. The ping-pong balls bounce at random until a total of six balls "pop" into a tube attached to the drum. The numbers on the six balls (the random sample) are the winning Lotto numbers.

This method of random sampling is fairly easy to implement if the population is relatively small. It is not feasible, however, when the population consists of a large number of observations. Since it is also very difficult to achieve a thorough mixing, the procedure only approximates random sampling. Most scientific studies, however, rely on computer software (with built-in random-number generators) to automatically generate the random sample. Almost all of the popular statistical software packages available (e.g., SAS, SPSS, MINITAB) have procedures for generating random samples.

2 Exercises

7 Guilt in decision making. The effect of guilt emotion on how a decision-maker focuses on the problem was investigated in the *Journal of Behavioral Decision Making* (January 2007). A total of 155 volunteer students participated in the experiment, where each was randomly assigned to one of three emotional states (guilt, anger, or neutral) through a reading/writing task. Immediately after the task, the students were presented with a decision problem (e.g., whether or not to spend money on repairing a very old car). The researchers found

that a higher proportion of students in the guilty-state group chose not to repair the car than those in the neutral-state and anger-state groups.

(a) Identify the population, sample, and variables measured for this study.

(b) What inference was made by the researcher?

8 Use of herbal medicines. Refer to the *American Association of Nurse Anesthetists Journal* (February 2000) study on the use of herbal medicines before surgery, Exercise 4. The 500 surgical patients

that participated in the study were randomly selected from surgical patients at several metropolitan hospitals across the country.

- (a) Do the 500 surgical patients represent a population or a sample? Explain.
 - (b) If your answer was sample in part a, is the sample likely to be representative of the population? If you answered population in part a, explain how to obtain a representative sample from the population.
- 9 Massage therapy for athletes.** Does a massage enable the muscles of tired athletes to recover from exertion faster than usual? To answer this question, researchers recruited eight amateur boxers to participate in an experiment (*British Journal of Sports Medicine*, April 2000). After a 10-minute workout in which each boxer threw 400 punches, half the boxers were given a 20-minute massage and half just rested for 20 minutes. Before returning to the ring for a second workout, the heart rate (beats per minute) and blood lactate level (micromoles) were recorded for each boxer. The researchers found no difference in the means of the two groups of boxers for either variable.
- (a) Identify the experimental units of the study.
 - (b) Identify the variables measured and their type (quantitative or qualitative).
 - (c) What is the inference drawn from the analysis?
 - (d) Comment on whether this inference can be made about all athletes.
- 10 Gallup Youth Poll.** A Gallup Youth Poll was conducted to determine the topics that teenagers most want to discuss with their parents. The findings show that 46% would like more discussion about the family's financial situation, 37% would like to talk about school, and 30% would like to talk about religion. The survey was based on a national sampling of 505 teenagers, selected at random from all U.S. teenagers.
- (a) Describe the sample.
 - (b) Describe the population from which the sample was selected.
 - (c) Is the sample representative of the population?
 - (d) What is the variable of interest?
 - (e) How is the inference expressed?
 - (f) Newspaper accounts of most polls usually give a *margin of error* (e.g., plus or minus 3%) for the survey result. What is the purpose of the margin of error and what is its interpretation?
- 11 Insomnia and education.** Is insomnia related to education status? Researchers at the Universities of Memphis, Alabama at Birmingham, and Tennessee investigated this question in the *Journal of Abnormal Psychology* (February 2005). Adults living in Tennessee were selected to participate in the study using a random-digit telephone dialing procedure. Two of the many variables measured for each of the 575 study participants were number of years of education and insomnia status (normal sleeper or chronic insomnia). The researchers discovered that the fewer the years of education, the more likely the person was to have chronic insomnia.
- (a) Identify the population and sample of interest to the researchers.
 - (b) Describe the variables measured in the study as quantitative or qualitative.
 - (c) What inference did the researchers make?
- 12 Accounting and Machiavellianism.** Refer to the *Behavioral Research in Accounting* (January 2008) study of Machiavellian traits in accountants, Exercise 6. Recall that a questionnaire was administered to a random sample of 700 accounting alumni of a large southwestern university; however, due to nonresponse and incomplete answers, only 198 questionnaires could be analyzed. Based on this information, the researchers concluded that Machiavellian behavior is not required to achieve success in the accounting profession.
- (a) What is the population of interest to the researcher?
 - (b) Identify the sample.
 - (c) What inference was made by the researcher?
 - (d) How might the nonresponses impact the inference?

3 Describing Qualitative Data

Consider a study of aphasia published in the *Journal of Communication Disorders* (March 1995). Aphasia is the "impairment or loss of the faculty of using or understanding spoken or written language." Three types of aphasia have been identified by researchers: Broca's, conduction, and anomic. They wanted to determine whether one type of aphasia occurs more often than any other, and, if so, how often. Consequently, they measured aphasia type for a sample of 22 adult aphasiacs. Table 3 gives the type of aphasia diagnosed for each aphasiac in the sample.

 APHASIA

Subject	Type of Aphasia
1	Broca's
2	Anomic
3	Anomic
4	Conduction
5	Broca's
6	Conduction
7	Conduction
8	Anomic
9	Conduction
10	Anomic
11	Conduction
12	Broca's
13	Anomic
14	Broca's
15	Anomic
16	Anomic
17	Anomic
18	Conduction
19	Broca's
20	Anomic
21	Conduction
22	Anomic

Source: Reprinted from *Journal of Communication Disorders*, Mar. 1995, Vol. 28, No. 1, E. C. Li, S. E. Williams, and R. D. Volpe, "The effects of topic and listener familiarity of discourse variables in procedural and narrative discourse tasks," p. 44 (Table 1) Copyright © 1995, with permission from Elsevier.

For this study, the variable of interest, aphasia type, is qualitative in nature. Qualitative data are nonnumerical in nature; thus, the value of a qualitative variable can only be classified into categories called *classes*. The possible aphasia types—Broca's, conduction, and anomic—represent the classes for this qualitative variable. We can summarize such data numerically in two ways: (1) by computing the *class frequency*—the number of observations in the data set that fall into each class; or (2) by computing the *class relative frequency*—the proportion of the total number of observations falling into each class.

Definition 12 A **class** is one of the categories into which qualitative data can be classified.

Definition 13 The **class frequency** is the number of observations in the data set falling in a particular class.

Definition 14 The **class relative frequency** is the class frequency divided by the total number of observations in the data set, i.e.,

$$\text{class relative frequency} = \frac{\text{class frequency}}{n}$$

Examining Table 3, we observe that 5 aphasiacs in the study were diagnosed as suffering from Broca’s aphasia, 7 from conduction aphasia, and 10 from anomic aphasia. These numbers—5, 7, and 10—represent the class frequencies for the three classes and are shown in the summary table, Table 4.

Table 4 also gives the relative frequency of each of the three aphasia classes. From Definition 14, we know that we calculate the relative frequency by dividing the class frequency by the total number of observations in the data set. Thus, the relative frequencies for the three types of aphasia are

$$\text{Broca's: } \frac{5}{22} = .227$$

$$\text{Conduction: } \frac{7}{22} = .318$$

$$\text{Anomic: } \frac{10}{22} = .455$$

From these relative frequencies we observe that nearly half (45.5%) of the 22 subjects in the study are suffering from anomic aphasia.

Although the summary table in Table 4 adequately describes the data in Table 3, we often want a graphical presentation as well. Figures 1 and 2 show two of the most widely used graphical methods for describing qualitative data—bar graphs and pie charts. Figure 1 shows the frequencies of aphasia types in a **bar graph** produced with SAS. Note that the height of the rectangle, or “bar,” over each class is equal to the class frequency. (Optionally, the bar heights can be proportional to class relative frequencies.)

Table 4 Summary table for data on 22 adult aphasiacs

Class	Frequency	Relative Frequency
(Type of Aphasia)	(Number of Subjects)	(Proportion)
Broca’s	5	.227
Conduction	7	.318
Anomic	10	.455
Totals	22	1.000

A Review of Basic Concepts (Optional)

Figure 1 SAS bar graph for data on 22 aphasiacs

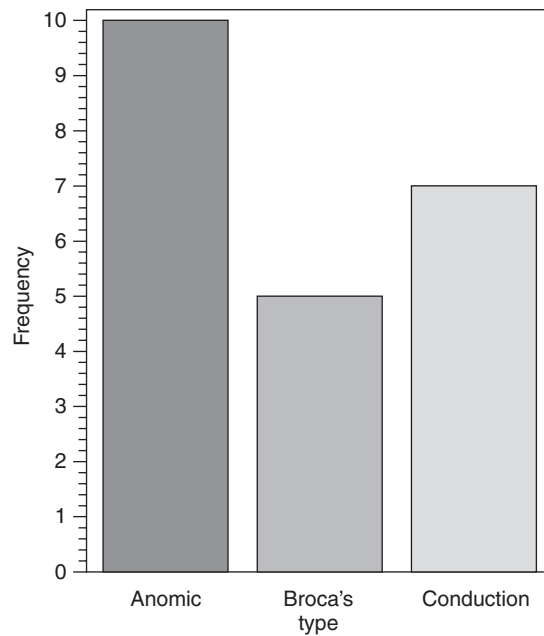
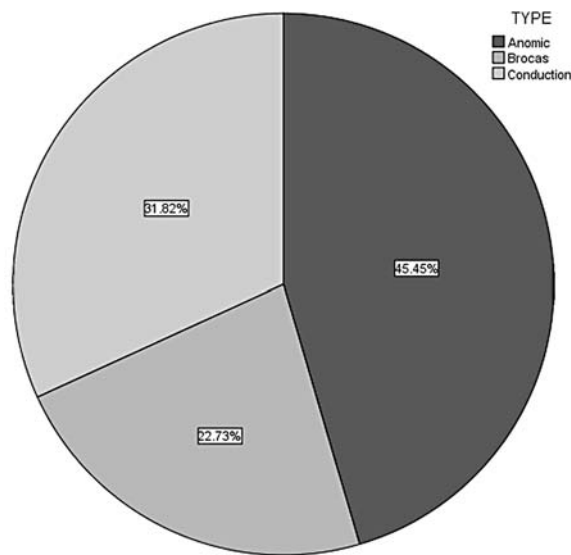


Figure 2 SPSS pie chart for data on 22 aphasiacs



In contrast, Figure 2 shows the relative frequencies of the three types of aphasia in a **pie chart** generated with SPSS. Note that the pie is a circle (spanning 360°) and the size (angle) of the “pie slice” assigned to each class is proportional to the class relative frequency. For example, the slice assigned to anomic aphasia is 45.5% of 360° , or $(.455)(360^\circ) = 163.8^\circ$.

3 Exercises

- 13 Estimating the rhino population.** The International Rhino Federation estimates that there are 17,800 rhinoceroses living in the wild in Africa and Asia. A breakdown of the number of rhinos of each species is reported in the accompanying table.

RHINO SPECIES	POPULATION ESTIMATE
African Black	3,610
African White	11,330
(Asian) Sumatran	300
(Asian) Javan	60
(Asian) Indian	2,500
Total	17,800

Source: International Rhino Federation, March 2007.

- (a) Construct a relative frequency table for the data.
 (b) Display the relative frequencies in a bar graph.
 (c) What proportion of the 17,800 rhinos are African rhinos? Asian?
- 14 Blogs for Fortune 500 firms.** Website communication through blogs and forums is becoming a key marketing tool for companies. The *Journal of Relationship Marketing* (Vol. 7, 2008) investigated the prevalence of blogs and forums at Fortune 500 firms with both English and Chinese websites. Of the firms that provided blogs/forums as a marketing tool, the accompanying table gives a breakdown on the entity responsible for creating the blogs/forums. Use a graphical method to describe the data summarized in the table. Interpret the graph.

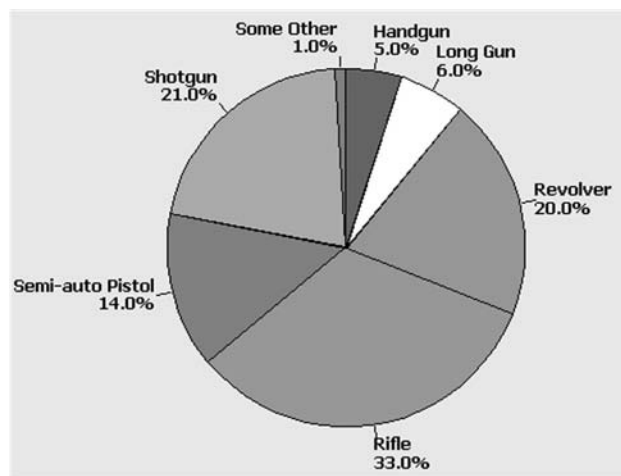
BLOG/FORUM	PERCENTAGE OF FIRMS
Created by company	38.5
Created by employees	34.6
Created by third party	11.5
Creator not identified	15.4

Source: "Relationship Marketing in Fortune 500 U.S. and Chinese Web Sites," Karen E. Mishra and Li Cong, *Journal of Relationship Marketing*, Vol. 7, No. 1, 2008, reprinted by permission of the publisher (Taylor and Francis, Inc.)

- 15 National Firearms Survey.** In the journal *Injury Prevention* (January 2007), researchers from the

Harvard School of Public Health reported on the size and composition of privately held firearm stock in the United States. In a representative household telephone survey of 2,770 adults, 26% reported that they own at least one gun. The accompanying graphic summarizes the types of firearms owned.

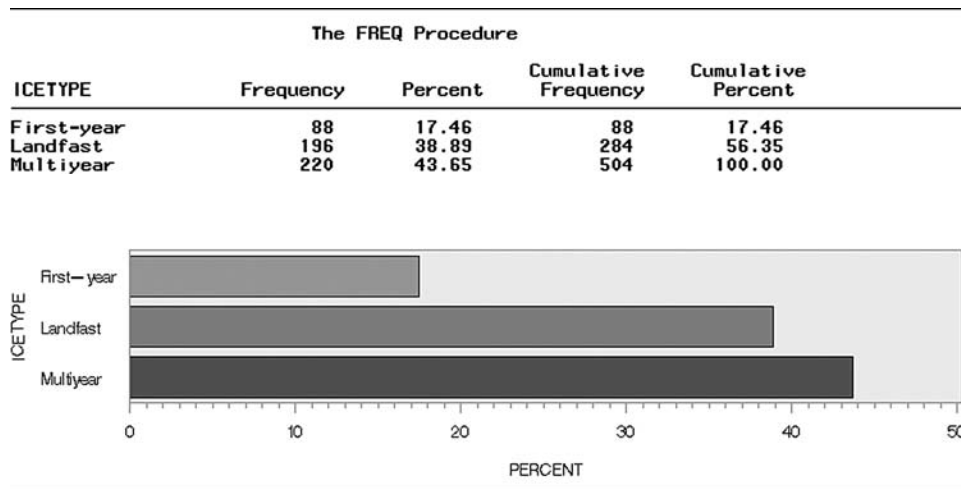
- (a) What type of graph is shown?
 (b) Identify the qualitative variable described in the graph.
 (c) From the graph, identify the most common type of firearms.



 PONDICE

- 16 Characteristics of ice melt ponds.** The National Snow and Ice Data Center (NSIDC) collects data on the albedo, depth, and physical characteristics of ice melt ponds in the Canadian arctic. Environmental engineers at the University of Colorado are using these data to study how climate impacts the sea ice. Data for 504 ice melt ponds located in the Barrow Strait in the Canadian arctic are saved in the PONDICE file. One variable of interest is the type of ice observed for each pond. Ice type is classified as first-year ice, multiyear ice, or landfast ice. A SAS summary table and horizontal bar graph that describe the ice types of the 504 melt ponds are shown at the top of the next page.

- (a) Of the 504 melt ponds, what proportion had landfast ice?



- (b) The University of Colorado researchers estimated that about 17% of melt ponds in the Canadian arctic have first-year ice. Do you agree?
- (c) Interpret the horizontal bar graph.

- (c) Apply a graphical method to all 223 wells to describe the detectible level of MTBE distribution.
- (d) Use two bar charts, placed side by side, to compare the proportions of contaminated wells for private and public well classes. What do you infer?

17 Groundwater contamination in wells. In New Hampshire, about half the counties mandate the use of reformulated gasoline. This has led to an increase in the contamination of groundwater with methyl *tert*-butyl ether (MTBE). *Environmental Science and Technology* (January 2005) reported on the factors related to MTBE contamination in private and public New Hampshire wells. Data were collected for a sample of 223 wells. These data are saved in the MTBE file. Three of the variables are qualitative in nature: well class (public or private), aquifer (bedrock or unconsolidated), and detectible level of MTBE (below limit or detect). [Note: A detectible level of MTBE occurs if the MTBE value exceeds .2 micrograms per liter.] The data for 10 selected wells are shown in the accompanying table.

- (a) Apply a graphical method to all 223 wells to describe the well class distribution.
- (b) Apply a graphical method to all 223 wells to describe the aquifer distribution.

MTBE (selected observations)

WELL CLASS	AQUIFER	DETECT MTBE
Private	Bedrock	Below Limit
Private	Bedrock	Below Limit
Public	Unconsolidated	Detect
Public	Unconsolidated	Below Limit
Public	Unconsolidated	Below Limit
Public	Unconsolidated	Below Limit
Public	Unconsolidated	Detect
Public	Unconsolidated	Below Limit
Public	Unconsolidated	Below Limit
Public	Bedrock	Detect
Public	Bedrock	Detect

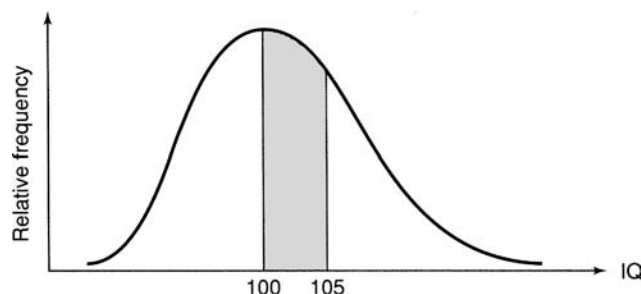
Source: Ayotte, J. D., Argue, D. M., and McGarry, F. J. “Methyl *tert*-butyl ether occurrence and related factors in public and private wells in southeast New Hampshire,” *Environmental Science and Technology*, Vol. 39, No. 1, Jan. 2005. Reprinted with permission.

4 Describing Quantitative Data Graphically

A useful graphical method for describing quantitative data is provided by a relative frequency distribution. Like a bar graph for qualitative data, this type of graph shows the proportions of the total set of measurements that fall in various intervals on the scale of measurement. For example, Figure 3 shows the intelligence quotients (IQs) of identical twins. The area over a particular interval under a relative frequency distribution curve is proportional to the fraction of the total number

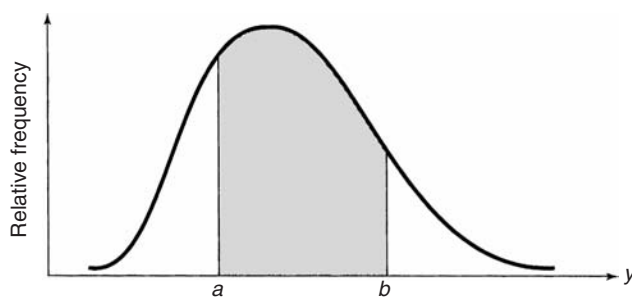
of measurements that fall in that interval. In Figure 3, the fraction of the total number of identical twins with IQs that fall between 100 and 105 is proportional to the shaded area. **If we take the total area under the distribution curve as equal to 1, then the shaded area is equal to the fraction of IQs that fall between 100 and 105.**

Figure 3 Relative frequency distribution: IQs of identical twins



Throughout this text we denote the quantitative variable measured by the symbol y . Observing a single value of y is equivalent to selecting a single measurement from the population. The probability that it will assume a value in an interval, say, a to b , is given by its relative frequency or **probability distribution**. The total area under a probability distribution curve is always assumed to equal 1. Hence, the probability that a measurement on y will fall in the interval between a and b is equal to the shaded area shown in Figure 4.

Figure 4 Probability distribution for a quantitative variable



Since the theoretical probability distribution for a quantitative variable is usually unknown, we resort to obtaining a sample from the population: Our objective is to describe the sample and use this information to make inferences about the probability distribution of the population. **Stem-and-leaf plots** and **histograms** are two of the most popular graphical methods for describing quantitative data. Both display the frequency (or relative frequency) of observations that fall into specified intervals (or classes) of the variable's values.

For small data sets (say, 30 or fewer observations) with measurements with only a few digits, stem-and-leaf plots can be constructed easily by hand. Histograms, on the other hand, are better suited to the description of larger data sets, and they permit greater flexibility in the choice of classes. Both, however, can be generated using the computer, as illustrated in the following examples.

Example
3

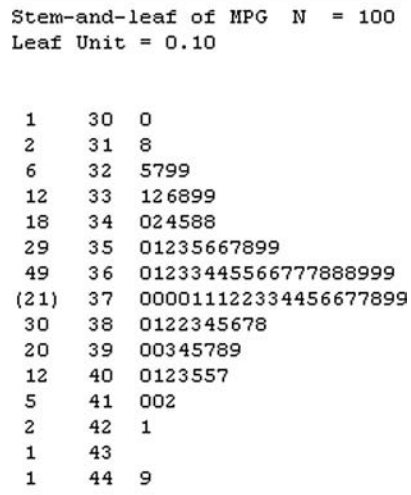
The Environmental Protection Agency (EPA) performs extensive tests on all new car models to determine their highway mileage ratings. The 100 measurements in Table 5 represent the results of such tests on a certain new car model.

A visual inspection of the data indicates some obvious facts. For example, most of the mileages are in the 30s, with a smaller fraction in the 40s. But it is difficult to provide much additional information without resorting to a graphical method of summarizing the data. A stem-and-leaf plot for the 100 mileage ratings, produced using MINITAB, is shown in Figure 5. Interpret the figure.

 EPAGAS

36.3	41.0	36.9	37.1	44.9	36.8	30.0	37.2	42.1	36.7
32.7	37.3	41.2	36.6	32.9	36.5	33.2	37.4	37.5	33.6
40.5	36.5	37.6	33.9	40.2	36.4	37.7	37.7	40.0	34.2
36.2	37.9	36.0	37.9	35.9	38.2	38.3	35.7	35.6	35.1
38.5	39.0	35.5	34.8	38.6	39.4	35.3	34.4	38.8	39.7
36.3	36.8	32.5	36.4	40.5	36.6	36.1	38.2	38.4	39.3
41.0	31.8	37.3	33.1	37.0	37.6	37.0	38.7	39.0	35.8
37.0	37.2	40.7	37.4	37.1	37.8	35.9	35.6	36.7	34.5
37.1	40.3	36.7	37.0	33.9	40.1	38.0	35.2	34.8	39.5
39.9	36.9	32.9	33.8	39.8	34.0	36.8	35.0	38.1	36.9

Figure 5 MINITAB stem-and-leaf plot for EPA gas mileages



Solution

In a stem-and-leaf plot, each measurement (mpg) is partitioned into two portions, a *stem* and a *leaf*. MINITAB has selected the digit to the right of the decimal point to represent the leaf and the digits to the left of the decimal point to represent the stem. For example, the value 36.3 mpg is partitioned into a stem of 36 and a leaf of 3, as illustrated below:

Stem		Leaf
36		3

The stems are listed in order in the second column of the MINITAB plot, Figure 5, starting with the smallest stem of 30 and ending with the largest stem of 44.

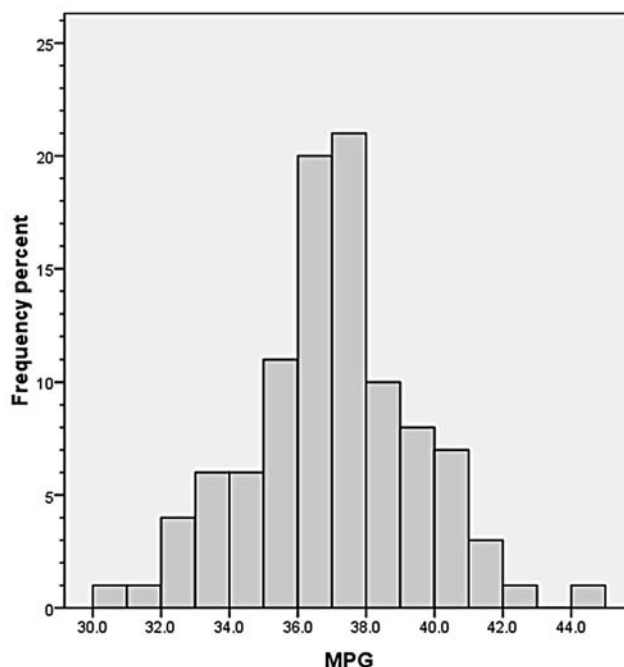
The respective leaves are then placed to the right of the appropriate stem row in increasing order.* For example, the stem row of 32 in Figure 5 has four leaves—5, 7, 9, and 9—representing the mileage ratings of 32.5, 32.7, 32.9, and 32.9, respectively. Notice that the stem row of 37 (representing MPGs in the 37's) has the most leaves (21). Thus, 21 of the 100 mileage ratings (or 21%) have values in the 37's. If you examine stem rows 35, 36, 37, 38, and 39 in Figure 5 carefully, you will also find that 70 of the 100 mileage ratings (70%) fall between 35.0 and 39.9 mpg. ■

Example 4

Refer to Example 3. Figure 6 is a relative frequency histogram for the 100 EPA gas gas mileages (Table 5) produced using SPSS.

- (a) Interpret the graph.
- (b) Visually estimate the proportion of mileage ratings in the data set between 36 and 38 MPG.

Figure 6 SPSS histogram for 100 EPA gas mileages



Solution

- (a) In constructing a histogram, the values of the mileages are divided into the intervals of equal length (1 MPG), called **classes**. The endpoints of these classes are shown on the horizontal axis of Figure 6. The relative frequency (or percentage) of gas mileages falling in each class interval is represented by the vertical bars over the class. You can see from Figure 6 that the mileages tend to pile up near 37 MPG; in fact, the class interval from 37 to 38 MPG has the greatest relative frequency (represented by the highest bar).

Figure 6 also exhibits **symmetry** around the center of the data—that is, a tendency for a class interval to the right of center to have about the same relative frequency as the corresponding class interval to the left of center. This

* The first column in the MINITAB stem-and-leaf plot gives the cumulative number of measurements in the nearest “tail” of the distribution beginning with the stem row.

is in contrast to **positively skewed** distributions (which show a tendency for the data to tail out to the right due to a few extremely large measurements) or to **negatively skewed** distributions (which show a tendency for the data to tail out to the left due to a few extremely small measurements).

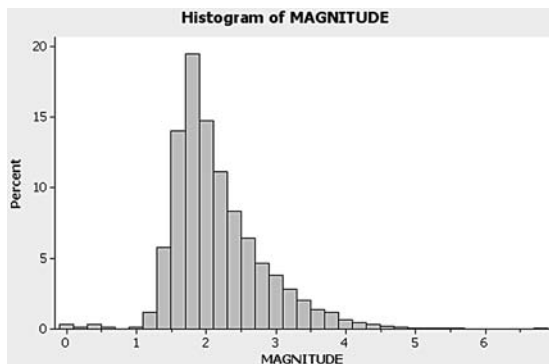
- (b) The interval 36–38 MPG spans two mileage classes: 36–37 and 37–38. The proportion of mileages between 36 and 38 MPG is equal to the sum of the relative frequencies associated with these two classes. From Figure 6 you can see that these two class relative frequencies are .20 and .21, respectively. Consequently, the proportion of gas mileage ratings between 36 and 38 MPG is $(.20 + .21) = .41$, or 41%. ■

4 Exercises

EARTHQUAKE

18 Earthquake aftershock magnitudes. Seismologists use the term “aftershock” to describe the smaller earthquakes that follow a main earthquake. Following the Northridge earthquake on January 17, 1994, the Los Angeles area experienced 2,929 aftershocks in a three-week period. The magnitudes (measured on the Richter scale) for these aftershocks were recorded by the U.S. Geological Survey and are saved in the EARTHQUAKE file. A MINITAB relative frequency histogram for these magnitudes is shown below.

- (a) Estimate the percentage of the 2,929 aftershocks measuring between 1.5 and 2.5 on the Richter scale.
 (b) Estimate the percentage of the 2,929 aftershocks measuring greater than 3.0 on the Richter scale.



- (c) Is the aftershock data distribution skewed right, skewed left, or symmetric?

19 Eating disorder study. Data from a psychology experiment were reported and analyzed in *American Statistician* (May 2001). Two samples of female students participated in the experiment. One sample consisted of 11 students known to suffer from the eating disorder bulimia; the other sample consisted of 14 students with normal eating habits. Each student completed a questionnaire from which a “fear of negative evaluation” (FNE) score was produced. (The higher the score, the greater the fear of negative evaluation.) The data are displayed in the table at the bottom of the page.

- (a) Construct a stem-and-leaf display for the FNE scores of all 25 female students.
 (b) Highlight the bulimic students on the graph, part a. Does it appear that bulimics tend to have a greater fear of negative evaluation? Explain.
 (c) Why is it important to attach a measure of reliability to the inference made in part b?

20 Data on postmortem intervals. *Postmortem interval* (PMI) is defined as the elapsed time between death and an autopsy. Knowledge of PMI is considered essential when conducting medical research on human cadavers. The data in the table are the PMIs of 22 human brain specimens obtained at autopsy in a recent study (*Brain and Language*, June 1995). Graphically describe the PMI data with a stem-and-leaf plot. Based on the plot, make a summary statement about the PMI of the 22 human brain specimens.

BULIMIA

Bulimic students:	21	13	10	20	25	19	16	21	24	13	14			
Normal students:	13	6	16	13	8	19	23	18	11	19	7	10	15	20

Source: Randles, R. H. “On neutral responses (zeros) in the sign test and ties in the Wilcoxon-Mann-Whitney test,” *American Statistician*, Vol. 55, No. 2, May 2001 (Figure 3).

 BRAINPMI

Postmortem intervals for 22 human brain specimens

5.5	14.5	6.0	5.5	5.3	5.8	11.0	6.1
7.0	14.5	10.4	4.6	4.3	7.2	10.5	6.5
3.3	7.0	4.1	6.2	10.4	4.9		

Source: Reprinted from *Brain and Language*, Vol. 49, Issue 3, T. L. Hayes and D. A. Lewis, “Anatomical Specialization of the Anterior Motor Speech Area: Hemispheric Differences in Magnopyramidal Neurons,” p. 292 (Table 1), Copyright © 1995, with permission of Elsevier.

21 Is honey a cough remedy? Coughing at night is a common symptom of an upper respiratory tract infection, yet there is no accepted therapeutic cure. Does a teaspoon of honey before bed really calm a child’s cough? To test the folk remedy, pediatric researchers at Pennsylvania State University carried out a designed study conducted over two nights (*Archives of Pediatrics and Adolescent Medicine*, December 2007.) A sample of 105 children who were ill with an upper respiratory tract infection and their parents participated in the study. On the first night, the parents rated their children’s cough symptoms on a scale from 0 (no problems at all) to 6 (extremely severe) in five different areas. The total symptoms score (ranging from 0 to 30 points) was the variable of interest for the 105 patients. On the second night, the parents were instructed to give their sick child a dosage of liquid “medicine” prior to bedtime. Unknown to the parents, some were given a dosage of dextromethorphan (DM)—an over-the-counter cough medicine—while others were given a similar dose of honey. Also, a third group of parents (the control group) gave their sick children no dosage at all. Again, the parents rated their children’s cough symptoms, and the improvement in total cough symptoms score was determined for each child. The data (improvement scores) for the study are shown in the accompanying

table, followed by a MINITAB stem-and-leaf plot of the data. Shade the leaves for the honey dosage group on the stem-and-leaf plot. What conclusions can pediatric researchers draw from the graph? Do you agree with the statement (extracted from the article), “honey may be a preferable treatment for the cough and sleep difficulty associated with childhood upper respiratory tract infection”?

Stem-and-leaf of TotalScore N = 105
Leaf Unit = 0.10

1	0	0
4	1	000
4	2	
7	3	000
16	4	000000000
20	5	0000
28	6	00000000
41	7	0000000000000
52	8	0000000000
(13)	9	0000000000000
40	10	0000000000
30	11	000000
24	12	0000000000000
11	13	0000
7	14	0
6	15	00000
1	16	0

22 Comparing voltage readings. A Harris Corporation/University of Florida study was undertaken to determine whether a manufacturing process performed at a remote location could be established locally. Test devices (pilots) were setup at both the old and new locations, and voltage readings on the process were obtained. A “good” process was considered to be one with voltage readings of at least 9.2 volts (with larger readings better than smaller readings). The first table on the next page contains voltage readings for 30 production runs at each location.

 HONEYCOUGH

Honey Dosage:	12	11	15	11	10	13	10	4	15	16	9	14	10	6	10	8	11	12	12	8			
	12	9	11	15	10	15	9	13	8	12	10	8	9	5	12								
DM Dosage:	4	6	9	4	7	7	7	9	12	10	11	6	3	4	9	12	7	6	8	12	12	4	12
	13	7	10	13	9	4	4	10	15	9													
No Dosage (Control):	5	8	6	1	0	8	12	8	7	7	1	6	7	7	12	7	9	7	9	5	11	9	5
	6	8	8	6	7	10	9	4	8	7	3	1	4	3									

Source: Paul, I. M., et al. “Effect of honey, dextromethorphan, and no treatment on nocturnal cough and sleep quality for coughing children and their parents,” *Archives of Pediatrics and Adolescent Medicine*, Vol. 161, No. 12, Dec. 2007 (data simulated).

 VOLTAGE

OLD LOCATION			NEW LOCATION		
9.98	10.12	9.84	9.19	10.01	8.82
10.26	10.05	10.15	9.63	8.82	8.65
10.05	9.80	10.02	10.10	9.43	8.51
10.29	10.15	9.80	9.70	10.03	9.14
10.03	10.00	9.73	10.09	9.85	9.75
8.05	9.87	10.01	9.60	9.27	8.78
10.55	9.55	9.98	10.05	8.83	9.35
10.26	9.95	8.72	10.12	9.39	9.54
9.97	9.70	8.80	9.49	9.48	9.36
9.87	8.72	9.84	9.37	9.64	8.68

Source: Harris Corporation, Melbourne, Fla.

 SHIPSANIT (selected observations)

SHIP NAME	SANITATION SCORE
<i>Adventure of the Seas</i>	95
<i>Albatross</i>	96
<i>Amsterdam</i>	98
<i>Arabella</i>	94
<i>Arcadia</i>	98
.	.
.	.
<i>Wind Surf</i>	95
<i>Yorktown Clipper</i>	91
<i>Zaandam</i>	98
<i>Zenith</i>	94
<i>Zuiderdam</i>	94

Source: National Center for Environmental Health, Centers for Disease Control and Prevention, May 24, 2006.

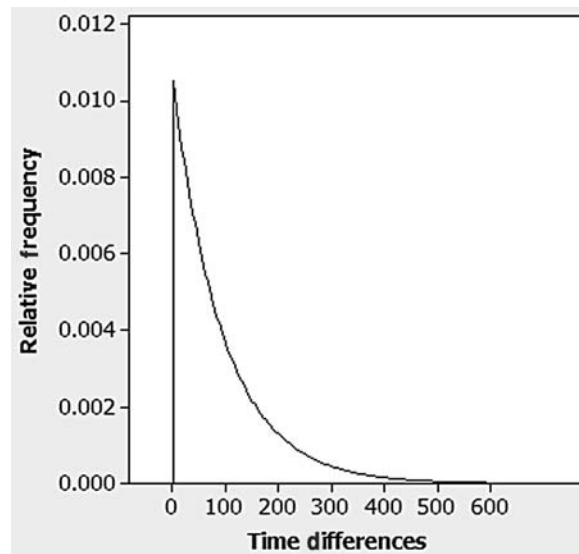
- (a) Construct a relative frequency histogram for the voltage readings of the old process.
- (b) Construct a stem-and-leaf display for the voltage readings of the old process. Which of the two graphs in parts a and b is more informative?
- (c) Construct a frequency histogram for the voltage readings of the new process.
- (d) Compare the two graphs in parts a and c. (You may want to draw the two histograms on the same graph.) Does it appear that the manufacturing process can be established locally (i.e., is the new process as good as or better than the old)?

23 Sanitation inspection of cruise ships. To minimize the potential for gastrointestinal disease outbreaks, all passenger cruise ships arriving at U.S. ports are subject to unannounced sanitation inspections. Ships are rated on a 100-point scale by the Centers for Disease Control and Prevention. A score of 86 or higher indicates that the ship is providing an accepted standard of sanitation. The May 2006 sanitation scores for 169 cruise ships are saved in the SHIPSANIT file. The first five and last five observations in the data set are listed in the accompanying table.

- (a) Generate a stem-and-leaf display of the data. Identify the stems and leaves of the graph.
- (b) Use the stem-and-leaf display to estimate the proportion of ships that have an accepted sanitation standard.
- (c) Locate the inspection score of 84 (*Sea Bird*) on the stem-and-leaf display.
- (d) Generate a histogram for the data.
- (e) Use the histogram to estimate the proportion of ships that have an accepted sanitation standard.

 PHISHING

24 Phishing attacks to email accounts. *Phishing* is the term used to describe an attempt to extract personal/financial information (e.g., PIN numbers, credit card information, bank account numbers) from unsuspecting people through fraudulent email. An article in *Chance* (Summer 2007) demonstrates how statistics can help identify phishing attempts and make e-commerce safer. Data from an actual phishing attack against an organization were used to determine whether the attack may have been an “inside job” that originated within the company. The company setup a publicized email account—called a “fraud box”—that enabled employees to notify them if they suspected an email phishing attack.



The interarrival times, that is, the time differences (in seconds), for 267 fraud box email notifications were recorded. *Chance* showed that if there is minimal or no collaboration or collusion from within the company, the interarrival times would have a frequency distribution similar to the one

shown in the accompanying figure. The 267 interarrival times are saved in the PHISHING file. Construct a frequency histogram for the interarrival times. Is the data skewed to the right? Give your opinion on whether the phishing attack against the organization was an “inside job.”

5 Describing Quantitative Data Numerically

Numerical descriptive measures provide a second (and often more powerful) method for describing a set of quantitative data. These measures, which locate the center of the data set and its spread, actually enable you to construct an approximate mental image of the distribution of the data set.

Note: Most of the formulas used to compute numerical descriptive measures require the summation of numbers. For instance, we may want to sum the observations in a data set, or we may want to square each observation and then sum the squared values. The symbol Σ (sigma) is used to denote a summation operation.

For example, suppose we denote the n sample measurements on a random variable y by the symbols $y_1, y_2, y_3, \dots, y_n$. Then the sum of all n measurements in the sample is represented by the symbol

$$\sum_{i=1}^n y_i$$

This is read “summation y , y_1 to y_n ” and is equal to the value

$$y_1 + y_2 + y_3 + \dots + y_n$$

One of the most common measures of central tendency is the **mean**, or arithmetic average, of a data set. Thus, if we denote the sample measurements by the symbols y_1, y_2, y_3, \dots , the sample mean is defined as follows:

Definition 15 The **mean** of a sample of n measurements y_1, y_2, \dots, y_n is

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

The mean of a population, or equivalently, the expected value of y , $E(y)$, is usually unknown in a practical situation (we will want to infer its value based on the sample data). Most texts use the symbol μ to denote the mean of a population. Thus, we use the following notation:

Notation

Sample mean: \bar{y}
Population mean: $E(y) = \mu$

The spread or variation of a data set is measured by its **range**, its **variance**, or its **standard deviation**.

Definition 16 The **range** of a sample of n measurements y_1, y_2, \dots, y_n is the difference between the largest and smallest measurements in the sample.

Example 5

If a sample consists of measurements 3, 1, 0, 4, 7, find the sample mean and the sample range.

Solution

The sample mean and range are

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{15}{5} = 3$$

$$\text{Range} = 7 - 0 = 7$$

The variance of a set of measurements is defined to be the average of the *squares of the deviations* of the measurements about their mean. Thus, the population variance, which is usually unknown in a practical situation, would be the mean or expected value of $(y - \mu)^2$, or $E[(y - \mu)^2]$. We use the symbol σ^2 to represent the variance of a population:

$$E[(y - \mu)^2] = \sigma^2$$

The quantity usually termed the **sample variance** is defined in the box.

Definition 17 The **variance** of a sample of n measurements y_1, y_2, \dots, y_n is defined to be

$$s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1} = \frac{\sum_{i=1}^n y_i^2 - n\bar{y}^2}{n - 1}$$

Note that the sum of squares of deviations in the sample variance is divided by $(n - 1)$, rather than n . Division by n produces estimates that tend to underestimate σ^2 . Division by $(n - 1)$ corrects this problem.

Example 6

Refer to Example 5. Calculate the sample variance for the sample 3, 1, 0, 4, 7.

Solution

We first calculate

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2 = 75 - 5(3)^2 = 30$$

where $\bar{y} = 3$ from Example 4. Then

$$s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1} = \frac{30}{4} = 7.5$$