Principles of Statistics:

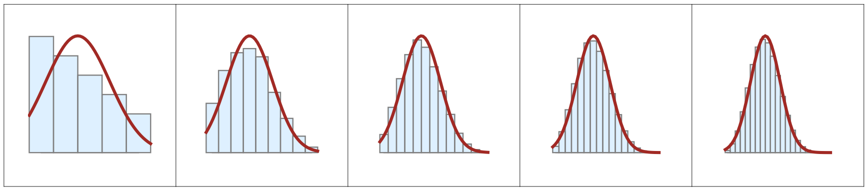
Emphasizing Applications in the Health Sciences

by Jenny A. Baglivo*

Mathematics Professor, Boston College

(email: baglivo@bc.edu)

November 2019



This project grew out of the need to develop a core course for students in the Connell School of Nursing (CSON) at Boston College. The course was designed to satisfy the principles for mathematics core courses offered at the university and to introduce students to the statistical language and principles needed to understand articles discussed in research-oriented courses given by CSON faculty members. The course has been given each year since 2007, and serves more than 100 students per year.

The project consists of a text plus four workbooks. The first draft of the text was used in Spring 2007, and the workbooks were assembled during Summer 2007 from lecture notes and handouts. Materials were updated each year based on feedback from students and instructors.

*Professor Baglivo retired from teaching in December 2018. She currently holds the position of Research Professor of Mathematics at Boston College.

Overview of the text. The first three chapters of the text include introductory material on study design and on techniques for summarizing data; material in these chapters will be referenced often in later chapters. The next two chapters introduce probability theory and the families of probability distributions that will be used extensively in applications. The next six chapters introduce statistical inference and applications of statistical inference.

Complete references for the textbooks used to develop the materials for this project are given in the References Chapter (beginning on page 205).

Overview of the workbooks. Because of time restrictions in semester courses at Boston College, we can thoroughly cover chapters 1 through 6, plus selected topics from the remaining chapters. The workbooks included with this project are organized as follows:

- *Workbook 1: Introductory Concepts*

  Corresponds to material in Chapters 1–3, about 8 lecture periods.
- *Workbook 2: Principles of Probability*

  Corresponds to material in Chapters 4–5, about 12 lecture periods.
- *Workbook 3: Principles of Statistics*

  Corresponds to material in Chapter 6, about 11 lecture periods.
- *Workbook 4: Additional Statistical Methods*

  Includes selected topics from Chapters 7–8, about 7 lecture periods.

Notes written in the margins of the workbooks refer students to specific text readings.

Acknowledgements. I am grateful for the support I received through the Boston College Teaching and Mentoring Grant program during the 2006–07 academic year to develop the first draft of the book and workbooks. Dr. Wei Tao, Principal Psychometrician at American College Testing (ACT), Inc., provided expert assistance during the 2006-07 academic year when she was a doctoral candidate in the Educational Research, Measurement & Evaluation program in the Lynch School of Education at Boston College.

CSON Associate Deans Cathy Read and Sean Clarke (now at NYU Rory Meyers College of Nursing) provided logistical support and guidance over the years.

I gratefully acknowledge the feedback I received from instructors who used the course materials at various stages of development: John Baldwin, Juliana Belding, Tom Crawford, Maria Fox, Iulia Gheorghita, Ross Goluboff, Rob Gross, Kyle Hayden, Siddhi Krishna, Charlie Landraitis, Spencer Leslie, Beth Romano, Ned Rosen and Jamison Wolf.

Finally, I enjoyed working with students from the Connell School of Nursing at Boston College. They are an energetic and dedicated group of young people, and I am proud to have been their professor.

Jenny A. Baglivo
November 2019

Table of Contents

Tables of Cumulative Probabilities and Critical Values

1 Introduction

> *"Statistics is the most important science in the whole world, for upon it depends the practical application of every other science and of every art: the one science essential to all political and social administration, all education, for it only gives exact results of our experience. . . .*
>
> *"To understand God's thoughts, we must study statistics, for these are the measure of his purpose."*
>
> Florence Nightingale (1820-1910)[1]

The study of statistics explores the collection, organization, analysis and interpretation of numerical data. We study statistics because the use of data has become ever more common in a growing number of professions, including the health sciences. Applied properly, statistical methods can be used to answer hard questions.

This text introduces the principles of probability and statistics, and applies the principles to problems of interest in the health sciences. The order and selection of topics have been chosen to give you a solid understanding of the principles.

Major topics are study design principles (Chapter 2), numerical and graphical summaries of data (Chapter 3), introduction to probability theory (Chapter 4), families of probability distributions (Chapter 5), introduction to statistical inference (Chapter 6), large sample analysis of two samples (Chapter 7), small sample analyses of means (Chapter 8), introduction to nonparametric analysis (Chapter 9) and introduction to association analysis (Chapter 10).

Chapters 2 and 3 are concerned with data production and data summary. Here you will learn about different sampling schemes, sources of bias, the differences between observational and experimental studies, the importance of randomization in experiments, ethical considerations in the gathering and use of data and in the design and analysis of experiments, the types of numerical data, numerical and graphical summaries of data, data distributions, and relationships between factors. These ideas will be used throughout the text.

Chapters 4 and 5 focus on probability theory. Probability theory is the study of random phenomena and is the foundation for statistical inference. Here you will learn basic probability operations, conditional probability, Bayes' rule and its application to the analysis of diagnostic tests, families of probability distributions, numerical summaries of probability models, conditional expectation and regression to the mean. In addition, you will be introduced to maximum likelihood estimation and to different approaches researchers use when applying probability concepts to problems in statistics.

[1] These quotes appeared in an article in *The American Statistician* (1990, 44:74-80) celebrating the contributions of women to the field of statistics. Florence Nightingale, who is often called the "passionate statistician," is known for her contributions to nursing, mathematics and statistics.

Following in the footsteps of Florence Nightingale, a grassroots nurse-inspired movement known as the Nightingale Initiative for Global Health (NIGH) was established in 2004 "to increase global public concern for and commitment to the priority of human health." The NIGH has been celebrated by many organizations, including the World Health Organization, for their contributions to global health.

See `https://www.biography.com/scientist/florence-nightingale` to learn more about the life of Florence Nightingale. See page 57 of this text for an example of Nightingale's famous rose diagram.

Chapter 6 focuses on the principles of statistical inference, including concepts from estimation theory and hypothesis testing theory. You will learn about sampling distributions, how to construct and interpret confidence intervals for unknown parameters, how to conduct and interpret two-sided hypothesis tests about unknown parameters, and about the limitations of statistical inference. In addition, you will be introduced to methods for determining the appropriate sample size for a given study. Statistical inference is developed in the "large samples" setting in this chapter.

Chapters 7 through 10 develop further methods for statistical inference, and include applications of these methods to health sciences studies. You will learn methods for analyzing two samples in the large samples setting, methods for "small sample" analyses of means, nonparametric methods for two sample analyses, and methods for analyzing the association between two factors. In addition, you will be introduced to different models of statistical inference.

The last section of each chapter gives a brief summary of the concepts learned in that chapter, provides additional examples to expand on those concepts, ties ideas from that chapter to ideas in previous chapters, and previews ideas from later chapters. Be sure to read each chapter, including the important brief summary section, with care.

The companion workbooks provide additional reinforcement of concepts learned in the text. The workbooks were designed to be used as day-to-day lecture notes, with room to write solutions to each exercise.

The text and companion workbooks focus on first principles, and the application of these principles in a selection of settings. The following texts further reinforce the principles and applications learned in this text, and present applications of the principles in new settings:

1. *Principles of Biostatistics* by Pagano & Gauvreau.
2. *Statistics: The Art & Science of Learning From Data* by Agresti & Franklin.
3. *The Practice of Statistics in the Life Sciences* by Baldi & Moore.

Complete references for these texts are given in the References Chapter (beginning on page 205).

2 Study Designs

Statistics has been described as the art and science of learning from data. The first and most important step is data gathering. This chapter discusses good study designs (that is, methods for producing data that give clear answers to research questions) and gives examples of well-designed and poorly-designed studies. References for this chapter include the texts by Agresti & Franklin (2007, Chapters 3 and 4), Baldi & Moore (2009, Chapters 7 and 8), Freedman et al (1991, Parts I and VI), Moore & McCabe (1999, Chapter 3), Moore & Notz (2006, Part I) and Pagano & Gauvreau (2000, Chapter 22).

2.1 Individuals, Samples and Populations

An *individual* (or *sampling unit* or *subject*) is the entity we measure in a study. Individuals are not necessarily persons. For example, if we were interested in comparing success rates for a given surgical procedure at hospitals in different parts of the country, the sampling unit would be a hospital and not a person.

A *population* (or *target population*) is the set of all individuals of interest. After accounting for practical constraints, the set from which we can actually sample is known as the *study population.* A list of the elements in the study population is called the *sampling frame.*

A *sample* is a subset of the (study) population. For example, if we were interested in comparing treatments for reducing serum cholesterol in women over 50 with high cholesterol, then our target population would be women over 50 with high cholesterol. We would use measurements made on a sample of these women to gain information about the population.

2.2 Variables, Statistics and Parameters

A *variable* is a characteristic of an individual. Different individuals will have different values of the characteristic of interest. A *statistic* is a numerical summary of values based on a sample. In general, different samples produce different numerical summaries. A *parameter* is a numerical summary of values for the entire population. A parameter is a fixed number, whose value is usually not known exactly.

Example 2.1 (Nursing Home Residents) Table 2.1 (page 4) gives information on nursing home residents in each state in the United States plus the District of Columbia (DC). If we consider DC to be a state, then the population of states has 51 individuals.

For each state, the numbers of nursing home residents at least 65 years old is reported as a rate per 1000 population in this age range. In Massachusetts, for example, 54 of every 1000 people at least 65 years old lives in a nursing home. Rates vary by state from a minimum of 13.6 (Hawaii) to a maximum of 74.9 (South Dakota).

The parameter of interest in this example is the mean rate across all 51 states,

$$43.8941 = (35.7 + 22.5 + \cdots + 37.8)/51.$$

The mean rate is a fixed number. By contrast, mean rates computed using samples from the population would vary from sample to sample. For example, the mean rate for the sample

Table 2.1: *Numbers of nursing home residents at least 65 years old per 1000 population 65 years of age and over for each state in the United States plus the District of Columbia.*

State	Rate	State	Rate	State	Rate
Alabama	35.7	Kentucky	47.8	North Dakota	59.5
Alaska	22.5	Louisiana	59.7	Ohio	49.3
Arizona	21.7	Maine	51.4	Oklahoma	59.1
Arkansas	50.5	Maryland	44.6	Oregon	30.2
California	26.7	Massachusetts	54.0	Pennsylvania	40.6
Colorado	40.0	Michigan	36.1	Rhode Island	60.3
Connecticut	54.6	Minnesota	67.6	South Carolina	28.8
Delaware	43.2	Mississippi	40.6	South Dakota	74.9
District of Columbia	34.4	Missouri	57.6	Tennessee	45.8
Florida	23.1	Montana	44.2	Texas	47.7
Georgia	45.5	Nebraska	67.4	Utah	29.1
Hawaii	13.6	Nevada	18.6	Vermont	46.2
Idaho	31.3	New Hampshire	52.7	Virginia	33.8
Illinois	52.1	New Jersey	33.4	Washington	37.8
Indiana	59.2	New Mexico	30.2	West Virginia	32.3
Iowa	70.5	New York	37.0	Wisconsin	62.1
Kansas	65.4	North Carolina	30.4	Wyoming	37.8

(*Source*: Pagano & Gauvreau, 2000, Appendix B, Table B.3)

{Delaware, Hawaii, Michigan, New York, Ohio} is

$$(43.2 + 13.6 + 36.1 + 37.0 + 49.3)/5 = 35.84.$$

The numerical summary 35.84 is an example of a statistic based on a sample of size 5 taken from the population of states.

2.3 Experiments and Observational Studies

In an *experiment*, researchers assign subjects to certain experimental conditions (often called *treatments* or *interventions*) and then measure the variables of interest. In an *observational study*, researchers merely observe the values of variables of interest.

The purpose of an experiment is to determine (if possible) a causal link between assigned treatments and measured responses. For example, in a study comparing the effectiveness of regular followup phone calls to check on the status of children with asthma (and suggest changes to treatment if needed), each child in a given sample would be assigned to one of two treatment groups: the child's caretaker would either receive regular followup calls or would not receive the calls. The variable of interest may be the number of emergency room visits during a one-year period. If the average number of emergency room visits is strikingly lower in the subsample of children whose caretakers received regular followup calls, then the results would suggest that followup phone calls are effective.

The purpose of an observational study is to describe a situation. For example, in a study comparing levels of smoking among hospital patients with lung cancer and with other respiratory problems, the variable of interest may be the average number of cigarettes smoked daily in the ten-year period preceding the onset of disease. Here the "treatments" are set (the hospital patient either has lung cancer or has other respiratory problems). If the average level of smoking is strikingly higher in the sample of patients with lung cancer, then the results would suggest a causal link between smoking and lung cancer.

Footnotes. The asthma study scenario above briefly describes a study published in 2006 suggesting that followup calls improves the quality of care for inner-city children with asthma (*Pediatrics* 117:1095-1103). The smoking study scenario briefly describes a famous early study, published in 1952, suggesting the link between smoking and lung cancer (*British Journal of Medicine* 2:1271-1286); for further details, see page 28.

2.3.1 Comparative Studies

The two study scenarios above are examples of *comparative studies*. In the asthma study scenario, the numbers of emergency room visits for subjects whose caretakers received followup phone calls is compared to the numbers for subjects whose caretakers did not receive the calls. In the smoking study scenario, average daily smoking levels in subjects with lung cancer is compared to levels in subjects with other respiratory diseases.

2.3.2 Response and Explanatory Variables

A *response variable* is one that measures an outcome or result of interest in a study. In the asthma study scenario above, for example, the response variable is the number of emergency room visits in a one-year period. In the smoking study scenario, the response variable is whether or not the patient had lung cancer (the response variable takes two values).

An *explanatory variable* is one that we think will help explain changes in a response variable. In the asthma study scenario, for example, the explanatory variable of interest is whether or not the subject's caretaker received followup phone calls (the explanatory variable takes two values). In the lung cancer study scenario, the explanatory variable of interest is the average daily smoking level in the ten-year period prior to onset of disease.

2.3.3 Prospective and Retrospective Studies

A *prospective study* is one that follows its subjects into the future. The asthma study scenario describes a prospective study since subjects are assigned to treatment groups and the response is measured at a later time.

A *retrospective study* is one that looks backward in time at events that happened in the past. The smoking study scenario describes a retrospective study since researchers are interested in determining if previous cigarette use can be used to help explain the development of lung cancer in a certain population of hospital patients. Retrospective observational studies are often used in medicine and public health.

2.3.4 Assessing Cause and Effect

In each study design scenario above (the asthma study scenario and the smoking study scenario) the researchers were interested in examining the possible causal link between the explanatory and response variables. A well-designed comparative experiment can be used to establish cause and effect. By contrast, a well-designed comparative observational study cannot be used to establish cause and effect definitively, although it can be used to give useful information about cause and effect. Thus, in general, experiments are preferred to observational studies.

Observational studies are used when experiments are not practical or not ethical. For example, it would be unethical to assign individuals to "smoking" and "nonsmoking" groups and to require them to follow their assigned treatments for a certain period of time in order to examine if smoking and lung cancer are linked.

2.4 Focus on Sampling

Researchers use information gathered from a sample of individuals to gain information about a population. The process of generalizing from the sample to the population is known as *inference* (or *statistical inference*). In order to make valid inferences from sample to population we need to focus on the methods used to select the individuals in the sample.

2.4.1 Sampling Procedures and Bias

A *sampling procedure* (or *selection procedure*) is a method for choosing individuals for a sample. *Sampling bias* (or *selection bias*) occurs if there is a systematic tendency to exclude one or another type of individual.

When conducting surveys, a *nonresponse bias* occurs when a chosen individual does not respond to one or more questions. Similarly, a *response bias* occurs when an individual responds incorrectly (for example, by lying or by choosing the wrong response to a poorly worded question).

These ideas are illustrated in the following famous example.

Example 2.2 (The *Literary Digest* Poll) (Freedman et al, 1991, pages 306-8)

> "In 1936, Franklin Delano Roosevelt was completing his first term of office as president of the United States. It was an election year, and the Republican candidate was Governor Alfred Landon of Kansas. The country was struggling to recover from the Great Depression. There were still nine million unemployed; real income had dropped by one third in the period 1929-1933, and was just beginning to turn upward. But Landon was campaigning on a program of economy in government, and Roosevelt was defensive about his deficit financing.

Landon. The spenders must go.
Roosevelt. We had to balance the budget of the American people before we could balance the budget of the national government. That makes common sense, doesn't it?

"The Nazis were rearming Germany, and the Civil War in Spain was moving to its hopeless climax. These issues dominated the headlines in the *New York Times*, but were ignored by both candidates.

Landon. We must mind our own business.

"Most observers thought Roosevelt would be an easy winner. Not so the *Literary Digest* magazine, which predicted an overwhelming victory for Landon, by 57% to 43%. This prediction was based on the largest number of people ever replying to a poll – about 2.4 million individuals. It was backed by the enormous prestige of the *Digest*, which had called the winner in every presidential election since 1916. However, Roosevelt won the 1936 election by a landslide; 62% to 38%. (The *Literary Digest* went bankrupt soon after.)

"The magnitude of the *Digest*'s error is staggering. It is the largest ever made by a major poll. Where did it come from? The number of replies was more than big enough. In fact, George Gallup was just setting up his survey organization. Using his methods, he was able to predict what the *Digest* predictions were going to be, well in advance of their publication, with an error of only one percentage point. Using another sample of about 50,000 people, he correctly forecast the Roosevelt victory, although his prediction of Roosevelt's share of the vote was off by quite a bit. Gallup forecast 56% for Roosevelt; the actual percentage was 62%, so the error was 62% − 56% = 6 percentage points. (Survey organizations use "percentage points" as the units for the difference between actual and predicted percents.) The results are summarized in [the following table].

	Roosevelt's percentage
Gallup's prediction of the *Digest* prediction	44
The *Digest* prediction of the election result	43
Gallup's prediction of the election result	56
The election results	62

"To find out where the *Digest* went wrong, you have to ask how they picked their sample . . . The *Digest*'s procedure was to mail questionnaires to 10 million people. The names and addresses of these people came from sources like telephone books and club membership lists. That tended to screen out the poor, who were unlikely to belong to clubs or have telephones. (At the time, for example, only one household in four had a telephone.) So there was a very strong selection bias against the poor in the *Digest*'s sampling procedure. Prior to 1936, this bias may not have affected the predictions very much, because rich and poor voted along similar lines. But in 1936, the political split followed economic lines more closely: the poor voted overwhelmingly for Roosevelt, the rich were for Landon . . .

"So the *Digest* did very badly at the first step in sampling. But there is also a second step. After deciding which people ought to be in the sample, a survey

> organization still has to get their opinions. This is harder than it looks . . . In the main *Digest* poll, only 2.4 million people bothered to reply, out of the 10 million who got the questionnaire. These 2.4 million respondents do not even represent the 10 million people who were polled, let alone the population of all voters. The *Digest* poll was spoiled both by selection bias and nonresponse bias."

Footnotes. It is easy to be impressed by the large sample size in the *Literary Digest* example above. But a large sample size does not erase the problem of selection bias. (Selection bias is not reduced by increasing the sample size.) By using a sampling procedure based on the rules of probability theory, and working with only 3,000 individuals from the same lists the *Digest* used, George Gallup was able to accurately predict the *Digest*'s results to within one percentage point.

Further, Gallup's use of probability methods in his choice of the sample of size 50,000 from the population of voters (and the attention he paid to how the survey was conducted once the sample was chosen) allowed him to estimate the actual election results more accurately.

2.4.2 Probability and Nonprobability Samples

A *probability sample* is one that is chosen using a probability method. A *nonprobability sample* (or a *convenience sample*) is one that is not chosen by a probability method. Three common types of probability samples are simple random samples, stratified random samples and cluster random samples.

Simple random samples. Let N denote the total number of individuals in a study population and n the number of individuals in the sample. A *simple random sample* of size n is a sample of n individuals chosen from the study population in such a way that every collection of n individuals has the same chance of being chosen. Computer-generated random numbers are often used to select simple random samples. Specifically,

1. The individuals in the study population are listed (the list is called the *sampling frame*) and each is assigned a whole number between 1 and N.
2. The computer is used to construct a random subset of size n from the set $\{1, 2, \ldots, N\}$.
3. Individuals with numbers in the chosen subset are part of the simple random sample.

Stratified random samples. In a *stratified random sample*, the study population is first divided into H distinct subpopulations (known as *strata*), where each individual is in one and only one stratum. A simple random sample of size n_i is then chosen from the i^{th} subpopulation. The total sample size is

$$n = n_1 + n_2 + \ldots + n_H.$$

Stratified random samples allow researchers to take into consideration any information that is known about individuals and that might affect the characteristic(s) of interest. For example, political pollsters might subdivide the population of registered voters into the following twelve strata,

1	Younger women living in cities	7	Younger men living in cities
2	Younger women living in suburbs	8	Younger men living in suburbs
3	Younger women living in rural areas	9	Younger men living in rural areas
4	Older women living in cities	10	Older men living in cities
5	Older women living in suburbs	11	Older men living in suburbs
6	Older women living in rural areas	12	Older men living in rural areas

and sample separately from each stratum. This sampling design would allow the researchers to compare results for men and women, for young and old, for voters in cities versus suburbs versus rural areas, and for different combinations of the given stratification variables.

Cluster random samples. In a *cluster random sample*, the study population is first divided into a large number of groups or *clusters*. A simple random sample of H clusters is chosen. All individuals in the chosen clusters are sampled. The total sample size is

$$n = c_1 + c_2 + \cdots + c_H,$$

where c_i is the total number of individuals in the i^{th} chosen cluster.

Examples of clusters include

1. all students living on one floor in a given dormitory,
2. all voters living on one city block in a large urban area,
3. all freshmen taking statistics at a small liberal arts college.

Cluster random samples are cost effective (that is, information from a cluster random sample of size n would cost less to gather than information from a simple random sample of size n), but are less informative than simple random samples of the same size.

In some cases, researchers choose to use cluster sampling techniques for reasons of safety. For example, in an article co-written by public health researchers at the Johns Hopkins Bloomberg School of Public Health and the Al Mustansiriya University School of Medicine in Baghdad, researchers reported on their use of cluster sampling techniques to determine the excess mortality rates in Iraq attributed to the invasion of Iraq, for the period from March 2003 to September 2004 (*Lancet* (2006), 368:1421-1428). Researchers worked with 50 clusters of 40 households each, where a household was defined as a unit that ate together, and had a separate entrance from the street or a separate apartment.

> "By confining the survey to a cluster of houses close to one another it was felt the benign purpose of the survey would spread quickly by word of mouth among households, thus lessening risk to interviewers." (page 1422)

Footnotes. Recall that a variable is a characteristic of an individual, that different individuals have different values of a given variable, and that a statistic is a summary of the values from a sample of individuals. Different samples produce different summary values. That is, different samples produce different statistics.

Statistics based on information from probability samples are reliable estimates of unknown population parameters. Although the variability of estimates depends on the particular sampling scheme (simple, stratified or cluster), valid inferences from sample to population can be made in all cases. By contrast, statistics based on information from nonprobability samples cannot be used to make valid inferences about unknown population parameters.

Although not ideal, nonprobability samples (also called *convenience samples*) can still provide valuable information. In medical research, for example, volunteers are often used in comparative experiments of new versus standard treatments for a particular disease. If designed properly, the results of such studies can be used to help determine if the proposed new treatment is better than the one in current use. The results, however, cannot be generalized to the population of all individuals with the disease.

2.5 Focus on Experiments

Researchers are often interested in determining (if possible) a causal link between treatment and response. Experiments are preferred to observational studies since, if conducted properly, their results can provide stronger cause-and-effect evidence. In order to make valid cause-and-effect inferences we need to focus on the methods used to design experiments.

2.5.1 Lurking Variables and Confounding

A *lurking variable* is one that influences the relationships among the variables in a study but is not included in the variables studied. The following example illustrates this concept.

Example 2.3 (Smoking and Health) (Agresti & Franklin, 2007, pages 132-3) As part of a study examining the possible harmful effects of cigarette smoking, researchers in the United Kingdom surveyed 1314 women about their cigarette use. The survey was conducted during the period from 1972 to 1974. Twenty years later, the researchers determined the survival status of each woman. The following table summarizes their results, where the rows correspond to whether or not the woman was a smoker at the time of the initial interview and the columns correspond to whether or not the woman was alive twenty years later.

	Alive	Dead	Total
Smoker	443 (33.7)	139 (10.6)	582 (44.3)
Nonsmoker	502 (38.2)	230 (17.5)	732 (55.7)
Total	945 (71.9)	369 (28.1)	1314 (100.0)

The table contains both frequencies and percents (in parentheses). For example, 582 women (44.3% of 1314) reported that they were smokers at the time of the initial interview, 945 women (71.9% of 1314) were alive twenty years after the initial interval, and 443 women (33.7% of 1314) were smokers who survived the twenty years.

If we consider the smokers and nonsmokers separately, a rather curious result is observed.

	Alive	Dead	Total
Smoker	443 (76.1)	139 (23.9)	582 (100.0)

	Alive	Dead	Total
Nonsmoker	502 (68.6)	230 (31.4)	732 (100.0)

Namely, that 76.1% (443/582) of smokers were alive twenty years later compared to only 68.6% (502/732) of nonsmokers.

The summaries above include the main explanatory variable (initial smoking status) and response variable (twenty year survival status) only, and seem to suggest that smoking is beneficial. But other important variables that could contribute to survival status are missing. One such variable is age at the time of the initial interview, which will be considered further in the continuation of this example below.

Confounding. Two variables are said to be *confounded* if their effects on a response variable cannot be distinguished from each other. Note that confounded variables can be explanatory variables or lurking variables.

Example 2.4 (Smoking and Health, continued) The summaries of the smoking and health study above leave us wondering if there are other important variables lurking in the background. In fact, age at time of initial interview can be used to help explain survival status twenty years later and is confounded with initial smoking status.

The first table below considers the initial ages of the smokers, grouped as follows: between 18 and 34 years old, between 35 and 54, between 55 and 64, and 65 or greater. The second table considers the ages of the nonsmokers.

	18-34	35-54	55-64	65+	Total
Smoker	179 (30.8)	239 (41.1)	115 (19.8)	49 (8.4)	582 (100.0)

	18-34	35-54	55-64	65+	Total
Nonsmoker	219 (29.9)	199 (27.2)	121 (16.5)	193 (26.4)	732 (100.0)

The women classified as nonsmokers tend to be older than those classified as smokers. Notice, in particular, that 26.4% of those classified as nonsmokers are 65 years or older compared to only 8.4% of those classified as smokers. Since age and survival status are related, the curious result reported earlier could be due to the effects of age and not smoking status. The effects of age and smoking cannot be separated.

The following table gives the percentage and proportion (in parentheses) of women alive after twenty years in each of 8 groups, where groups are defined using each combination of smoking status and age.

	18-34	35-54	55-64	65+
Smoker	97.2% (174/179)	82.8% (198/239)	55.7% (64/115)	14.3% (7/49)
Nonsmoker	97.3% (213/219)	90.5% (180/199)	66.9% (81/121)	14.5% (28/193)

Within each age group, the nonsmokers were more likely to survive. For example, among women between 35 and 54 when initially interviewed, 90.5% (180/199) of nonsmokers survived compared to only 82.8% (198/239) of smokers.

Footnotes. The extended example above reports on an observational study of the health effects of smoking. If the researchers had observed initial smoking status and twenty year survival status only, then age would be a lurking variable (an important but missing variable). The effects of age and smoking status were confounded. Once age was taken into account, a clearer picture of the relationship between smoking and survival emerged.

In observational studies, researchers can only hope that they have gathered information on all potential confounders (that is, that there are no other important variables lurking in the background). In experiments using randomization (where subjects are randomly assigned to treatments), researchers can be surer of their results even if they have not gathered information on all potential confounders.

2.5.2 Randomized Comparative Experiments

Randomized comparative experiments (also called *randomized controlled experiments*, *intervention trials* or *clinical trials*) have two important features. First, they control the effects of lurking variables on the response by using the method of comparison. Second, they reduce potential bias in treatment assignment by randomly assigning subjects to treatments.

Simple scenario. Consider, for example, the following simple scenario. A medical researcher is interested in determining if a proposed new treatment or intervention is effective. In general, it would be hard to judge the effectiveness of the proposed treatment or intervention properly without comparing it to something else.

Comparisons are made between the *treatment group* (the individuals chosen to receive the proposed new treatment or intervention) and the *control group* (the individuals chosen to receive a standard treatment or no treatment at all).

Once a sample of n individuals is identified, the next step is to randomly assign the individuals to either the treatment group or the control group. Let n_1 be the number to be assigned to the treatment group and $n_2 = n - n_1$ the number to be assigned to the control group. Computer-generated random numbers are often used to select the individuals in each group. Specifically,

1. Each individual in the sample is assigned a whole number between 1 and n.
2. The computer is used to construct a random subset of size n_1 from the set $\{1, 2, \ldots, n\}$.
3. Individuals with numbers in the chosen subset are assigned to the treatment group. The remaining individuals are assigned to the control group.

The randomization of individuals to treatment or control produces groups that should be similar in all respects before the treatments are given. The use of a control group ensures that influences other than the experimental treatments operate equally on both groups (thus, controlling the effects of any lurking variables). Any observed differences in responses can then be attributed to the effects of the treatments.

Blinding and double-blind experiments. If a subject in a randomized comparative experiment does not know his/her treatment assignment, then the subject is said to be *blind* to the assignment. If the people who work with the subjects (doctors and nurses, for example) do not know the subjects' treatment assignments, then they are said to be *blind* to the assignments.

In a *double-blind experiment*, neither the subjects nor the people who work with the subjects know the treatment assignments. Whenever possible, experiments should be conducted double-blind.

The ideas above are illustrated in the following famous example.

Example 2.5 (The Salk Vaccine Field Trial) (Freedman et al, 1991, pages 3-7)

> "The first polio epidemic hit the United States in 1916, and during the next forty years polio claimed many hundreds of thousands of victims, especially children. By the 1950s, several vaccines against this disease had been discovered. The one developed by Jonas Salk seemed the most promising. In laboratory trials, it had proved safe and had caused the production of antibodies against polio. A large-scale field trial was needed to see whether the vaccine would protect children against polio outside the laboratory.
>
> "In 1954, the Public Health Service decided to organize this kind of experiment. The subjects were children in the most vulnerable age groups — grades 1, 2 and 3. The field trial was carried out in selected school districts throughout the country, where the risk of polio was believed to be the worst. Two million children were involved, and half a million were vaccinated. A million were deliberately left unvaccinated, and half a million refused vaccination.
>
> "This illustrates the method of comparison. Only the subjects in the treatment group were vaccinated; the other subjects did not get the treatment and were used as controls. The responses of the two groups could then be compared to see if the treatment made any difference. In the Salk vaccine field trial, the treatment group and control group were of different sizes, but that did not matter. The investigators compared the rates at which children got polio in the two groups — cases per hundred thousand. Looking at rates instead of absolute numbers adjusts for the difference in the sizes of the two groups.
>
> "A troublesome question of medical ethics: Shouldn't all the children have been given the vaccine? One answer is that with new drugs, even after extensive laboratory testing, it may be unclear whether the benefits outweigh the risks. A field trial is needed to find out what the treatment does when used in the real world. Of course, giving the vaccine to a large number of children might seem to provide decisive evidence, even without controls. For instance, if the incidence of polio in 1954 had dropped sharply from 1953, that would seem to be proof of the effectiveness of the Salk vaccine. But it really would not be, because polio was an epidemic disease whose incidence varied a lot from year to year. In 1952, there were about 60,000 cases; in 1953, there were only half as many. Without controls, low incidence in 1954 could have meant one of two things: either the vaccine was effective, or there was no epidemic that year.
>
> "The only way to find out whether the vaccine worked was to leave some children unvaccinated. Of course, children could be vaccinated only with their parents' permission. So one possible design was this: The children whose parents consented would form the treatment group and get the vaccine. The other children would

form the control group. But it was known that higher-income parents would consent to treatment more often than lower-income parents. And this would have created a bias against the vaccine, because children of higher-income parents are more vulnerable to polio.

“This seems paradoxical at first, but polio is a disease of hygiene. Children who live in less hygienic surroundings tend to contract mild cases of polio early in childhood, while still protected by antibodies from their mothers. After being infected, they generate their own antibodies which protect them against more severe infection later. Children who live in more hygienic surroundings do not develop the antibodies.

“The statistical lesson is that to avoid bias, the treatment group and control group should be as similar as possible — except for the treatment. Then, any difference in response between the two groups is due to the treatment, rather than something else. If the two groups differ with respect to some factor other than the treatment, the effects of this other factor might be *confounded* (mixed up) with the effects of the treatment. Separating these effects can be difficult or impossible. Confounding is a major source of bias.

“For the Salk vaccine field trial, several designs were proposed. The National Foundation for Infantile Paralysis (NFIP) wanted to vaccinate all grade 2 children whose parents would consent, leaving the children in grades 1 and 3 as controls. And this NFIP design was accepted by many school districts. However, polio is a contagious disease, spreading through contact. So the incidence could have been higher in grade 2 than in grades 1 or 3. This would have prejudiced the study against the vaccine. Or the incidence could have been lower in grade 2, prejudicing the study in favor of the vaccine. Furthermore, children in the treatment group, where parental consent was needed, were bound to have different family backgrounds from those in the control group, where parental consent was not required. With the NFIP design, the treatment group would include too many children from the higher-income families, making this group more vulnerable to polio than the control group. Here was a definite bias — against the vaccine.

“Many school districts saw these flaws in the NFIP design, and used a different design. The control group had to be chosen from the same population as the treatment group: children whose parents consented to vaccination. Otherwise, the effect of family background would have been confounded with the effect of the vaccine. The next issue was how to assign the children to treatment or control. Human judgment seems to be needed, to make the control group like the treatment group with respect to the relevant variables — family income, or the children’s general health, personality, and social habits.

“Experience shows, however, that human judgments often result in substantial bias. It is better to use a carefully designed chance procedure. For the Salk trial, the procedure was [roughly] equivalent to tossing a coin for each child, with a 50-50 chance of assignment to the treatment or the control group. Such a procedure is objective and impartial. And the laws of chance guarantee that with enough subjects, the treatment group and the control group will resemble each other very closely with respect to all the important variables, whether or not these variables have been identified. . . .

“Another basic precaution in the Salk trial was the use of a *placebo*. Children in the control group were given an injection of salt dissolved in water. During

Table 2.2: *The results of the Salk vaccine trial of 1954. Size of groups and rate of polio cases per 100,000 in each group. The numbers are rounded.*

The randomized controlled double-blind experiment		
	Size	*Rate*
Treatment	200,000	28
Control	200,000	71
No consent	350,000	46

The NFIP study		
	Size	*Rate*
Grade 2 (vaccine)	225,000	25
Grades 1 and 3 (control)	725,000	54
Grade 2 (no consent)	125,000	44

(*Source*: Thomas Francis, Jr., *American Journal of Public Health* vol 45 (1955) pp 1-63)

the experiment the subject did not know whether they were in treatment or in control. So their response was to the vaccine, not the idea of treatment. It may seem unlikely that subjects could be protected from polio just by the strength of an idea. However, hospital patients suffering from severe post-operative pain have been given a "pain killer" which was made of a completely neutral substance: about one-third of the patients experienced prompt relief. [This phenomenon is known as the *placebo effect*.]

"Still another precaution: diagnosticians had to decide whether the children contracted polio during the experiment. Many forms of polio are hard to diagnose, and in borderline cases the diagnosticians could have been affected by knowing whether the child was vaccinated. So the doctors were not told which group the child belonged to. This was *double-blinding*: the subjects did not know whether they got the treatment or the placebo, and neither did those who evaluated the responses. This part of the Salk trial was a randomized controlled double-blind experiment, which is about the best design there is.

"How did it turn out? [Table 2.2 (page 15)] shows the rate of polio cases (per hundred thousand subjects) in the randomized controlled experiment, for the treatment group and the control group. The rate is much lower than for the treatment group, decisive proof of the effectiveness of the Salk vaccine.

"[Table 2.2] also shows that the NFIP study was biased against the vaccine. In the randomized controlled experiment, the vaccine cut the polio rate from 71 to 28 per hundred thousand; the apparent reduction in the NFIP study, from 54 to 25 per hundred thousand, is quite a bit less. The main source of the bias was confounding. The NFIP treatment group included only children whose parents consented to vaccination. The NFIP control group also included children whose parents would not have consented. The control group was not comparable to the treatment group.

"The randomized controlled double-blind design reduces bias to a minimum, and that is the main reason for using it whenever possible. Furthermore, this design has an important technical advantage. To see why, let us play devil's advocate for a moment and assume that the Salk vaccine really had no effect. Then, the

> difference between the polio rates for the treatment and control groups is just due to chance. How likely is that?
>
> "With the NFIP design, the results are affected by many factors that (from the point of view of the investigators) are random: which families volunteer, which children are in grade 2, and so on. However, the investigators do not have enough information to estimate the chances for these outcomes. So they cannot figure the odds against the difference in polio rates being due to these accidental factors. With a randomized controlled experiment, on the other hand, chance enters in a planned and simple way — when the assignment is made to treatment or control.
>
> "To spell this out, the devil's-advocate hypothesis says that the vaccine has no effect. On this hypothesis, a few children are fated to contract polio; assignment to treatment or control has nothing to do with it. Each child has a 50-50 chance to be in treatment or control, just depending on the toss of a coin. So each polio case has a 50-50 chance to turn up in the treatment group or the control group.
>
> "Therefore, the number of polio cases in the two groups must be about the same; any difference is due to the chance variability in coin tossing. Statisticians understand this kind of variability. They can figure the odds against it making a difference as large as the observed one. [In this case,] the odds are astronomical – a billion to one against."

Footnotes. Randomized comparative experiments (also called *randomized controlled experiments* or *clinical trials*) give the best information about cause and effect, but may not always be feasible. For example, the cost of running a large-scale field trial like the one described in the extended example above may be prohibitive. Or, it may be unethical to assign individuals to treatments (as it would be if one of the "treatments" involved using cigarettes for a fixed period of time).

When randomized comparative experiments are not possible, researchers must rely on the results of good observational studies. In a good observational study, the researchers will use the method of comparison with a control group and will attempt to measure and adjust for all confounding variables.

In the Salk vaccine field trial, $n = 400,000$ children participated in the randomized controlled experiment, with $n_1 = n_2 = 200,000$ children randomly assigned to each group. The informal description of the random assignment process (by the flip of a coin) is roughly equivalent to the general description for choosing random subsets given on page 12.

Note the similarity between using the computer to choose a simple random sample from a population (page 8) and using the computer to choose the individuals assigned to a treatment group in the simple scenario (page 12). In each case, a probability method (often called a *chance method*) is used in the design stage of a study. If chance is used in the design of a study, then the methods of probability and statistics can be used to make inferences.

2.5.3 Beyond the Simple Scenario

There are several ways to generalize the ideas of the last section. For example, the researcher might be interested in comparing more than two groups by using a completely randomized design. Or, the researcher might be interested in explicit control of a potential confounder by using a randomized block design.

Completely randomized design. Let k be the number of treatments under study. A *completely randomized design* is one in which all individuals in the sample are allocated at random among the k groups. Note that the simple scenario of the last section is a completely randomized design with $k = 2$ groups.

Once a sample of n individuals is identified, the random allocation can be accomplished using a simple generalization of the method outlined on page 12 for random allocation to two groups. For example, if $k = 3$ treatments are under study and the researcher wishes to assign n_1 individuals to the first treatment group, n_2 individuals to the second treatment group, and $n_3 = n - n_1 - n_2$ individuals to the third treatment group, then the procedure would be as follows:

1. Each individual in the sample is assigned a whole number between 1 and n.
2. The computer is used to randomly select a subset of size n_1 from the set $\{1, 2, \ldots, n\}$ and the individuals with numbers in the chosen subset are assigned to the first treatment group.
3. The computer is then used to randomly select a subset of size n_2 from the remaining set of $n_2 + n_3$ numbers. Individuals in the chosen subset are assigned to the second treatment group. The remaining individuals are assigned to the third treatment group.

The following example uses $k = 4$ treatment groups.

Example 2.6 (Trying to Quit) (Agresti & Franklin, 2007, pages 174 and 186) As part of a study to evaluate the use of an antidepressant known as Zyban and/or the use of a nicotine patch as treatments for quitting smoking, 893 smokers were randomly assigned to four treatment groups:

1	Zyban pill plus nicotine patch ($n_1 = 245$)	3	Placebo pill plus nicotine patch ($n_3 = 244$)
2	Zyban pill plus placebo patch ($n_2 = 244$)	4	Placebo pill plus placebo patch ($n_4 = 160$)

Each individual was given a one-year supply of pills and a one-year supply of patches. The placebo pills and placebo patches contained no active ingredients. The response of interest was whether the participant abstained from smoking for one year. The study was double-blind. That is, neither the participants nor the individuals assigned to measure the responses knew which pills or patches contained active ingredients.

The following table contains one-year abstinence percentages and proportions (in parentheses) for each treatment group:

	Zyban pill/ nicotine patch	Zyban pill/ placebo patch	Placebo pill/ nicotine patch	Placebo pill/ placebo patch
One-Year Abstinence	35.5% (87/245)	30.3% (74/244)	16.4% (40/244)	15.6% (25/160)

The Zyban pill plus nicotine patch treatment was the most effective, although the results for the group taking Zyban alone (that is, Zyban pill plus placebo patch) are a close second. Our study of statistical methods later in the course will help us sort out whether the observed differences in effectiveness percentages are likely to be real (that is, are likely to indicate a real difference in effectiveness) or due to chance alone.

The results for the last two groups are also quite close. It is interesting to note that 15.6% of individuals who took placebo only abstained from cigarette smoking for one year. This is an example of what is known as the *placebo effect*.

Randomized block design. A *block* is a group of individuals that are similar in ways that are expected to affect the response to treatments. In a *randomized block design*, the random assignment of individuals to treatments is carried out separately within each block.

Common blocking factors are gender, age and race. The following block design example uses a study scenario similar to the one above.

Example 2.7 (Effects of Partner Smoking) (Agresti & Franklin, 2007, page 183) Smokers whose partners also smoke have, in general, a more difficult time quitting than smokers whose partners do not smoke. Thus, a reasonable blocking factor would be whether or not the smoker lives with another smoker.

Consider, for example, a study comparing Zyban pill to placebo pill and assume that a sample of n individuals have been identified. Further, assume that the response of interest is whether the smoker abstains from smoking for one year.

Block 1 would consist of those individuals whose partner does not smoke and Block 2 would consist of those individuals whose partner does smoke. The individuals in Block 1 would be randomly assigned to either treatment or control. Similarly, individuals in Block 2 would be randomly assigned to either treatment or control.

At the end of the study, researchers would be able to determine (using statistical methods) if Zyban was an effective treatment and if the level of effectiveness was different for individuals in the two blocks.

Footnotes. The first example considers a study whose design is completely randomized with $k = 4$ treatments. This study is also an example of a *completely randomized two-factor design*, where each factor has two levels. In the first factor, the treatment is the Zyban pill and the control is the placebo pill. In the second factor, the treatment is the nicotine patch and the control is the placebo patch.

In the second example, the randomization is not complete. Instead, it is restricted to operate within groups of individuals that the researcher believes will respond similarly to the treatments. By restricting the randomization to within blocks, the researcher will be able to draw separate conclusions about each block.

The idea of blocking in designing experiments is similar to the idea of stratifying in sampling from populations (see page 8). In each case, the researcher is attempting to explicitly control for variables that may affect the results.

An important special case of the randomized block design is the matched pairs design. In a *matched pairs design*, researchers are interested in comparing two treatments and each block consists of two individuals who are as closely matched as possible. That is, who are as similar as possible on potential confounding variables. Within each matched pair, one individual is randomly selected to receive the first treatment while the other individual receives the second treatment. Matched pairs designs are often used in medical studies of chronic diseases.

2.5.4 Principles of Experimental Design

Each of the designs discussed in the last two sections satisfies the following basic principles of experimental design (Moore & Notz, 2006, page 83):

1. ***Control*** the effects of lurking variables on the response, most simply by comparing two or more treatments.
2. ***Randomize*** – use impersonal chance to assign subjects to treatments.
3. ***Use enough subjects*** in each group to reduce chance variation in the results.

These principles were formulated by Sir Ronald A. Fisher in the 1920's.

Ronald A. Fisher (1890–1962). The British statistician Ronald Ayler Fisher is one of the most important statisticians of the twentieth century. His early training was in mathematics, astronomy and biology.

R.A. Fisher was a man who liked to get his hands dirty – literally. One of his interests was farming. While working as a statistician at the Rothamsted Agricultural Experiment Station, where his job was to study the effects of nutrition and soil types on plant fertility, he formulated the practical principles of experimentation outlined above and worked out the details of many different design strategies. The centerpiece of his work — something that hadn't been used before — was the principle of randomization; it remains one of the most important ideas in statistics.

During his career, Fisher made fundamental contributions to both the design and the analysis of experiments, to genetics (another one of his passions), to estimation theory, where he was the first person to formulate principles for comparing alternative methods to estimate an unknown parameter using sample data, and to many other subjects. He was a skilled theoretician, who believed that good theoretical work should be tied to explicit real world applications. He had very little patience for theoreticians who worked on problems with only marginal (or no) practical applications. He was not shy about expressing his (often negative) views about the work of other statisticians. Although he could be a difficult man to work with, he was always kind and generous in his dealings with students and close colleagues.

2.6 Brief Summary and Additional Examples

This chapter introduces basic definitions (e.g. population, sample, parameter, statistic) and basic types of studies (e.g. observational studies, experiments), and focuses on important issues in sampling and experimental design. This section reiterates a few key points and gives additional examples.

Sampling revisited. If a sample is chosen from a population using a probability method (also called a chance method), then the results of a study can be generalized to the population from which the sample was drawn. That is, the results can be used to make inferences from the sample to the population. Otherwise, the results cannot be generalized.

Sample summaries vary from sample to sample. For example, in the nursing home example beginning on page 3, the population mean rate was 43.8941 (per 1000 population at least 65 years old). The following table gives sample mean rates for several different samples of 5 states each:

Sample of 5 States	*Mean Rate for States in Sample*
Delaware, Hawaii, Michigan, New York, Ohio	35.84
DC, Maryland, Minnesota, New Hampshire, Rhode Island	51.92
Maryland, Montana, Oklahoma, South Dakota, Utah	50.38
Arkansas, Mississippi, New Mexico, New York, Tennessee	40.82
Iowa, Kentucky, Montana, New Hampshire, Texas	52.58
DC, Nebraska, Ohio, Oklahoma, Tennessee	51.20
Iowa, Maryland, Oklahoma, Rhode Island, Washington	54.46

Quantifying the variability of sample statistics is an important topic we will study later.

Note that if the study population and the sampling and design methods are the same in two different studies and yet the studies give conflicting results, then the variability of sample statistics can be used to help explain the differences.

Observation and experimentation revisited. Good studies (whether observational or experimental) should include a control group and should try to adjust for potential confounding variables. When reviewing the results of studies, you should pay close attention to how the data were produced.

Example 2.8 (Ultrasound and Low Birthweight) (Freedman et al, 1991, page 15)

> "Human babies can now be examined in the womb using ultrasound. Several experiments on lab animals have shown that ultrasound examinations can cause low birthweight. If this is true for humans, there are grounds for concern. Investigators ran an observational study to find out, at the Johns Hopkins Hospital in Baltimore [with results reported in the *Journal of Obstetrics and Gynecology* (1988) 71: 513-7].
>
> "Of course, babies exposed to ultrasound differed from unexposed babies in many ways besides exposure; this was an observational study. The investigators found a number of confounding variables, and adjusted for them. Even so, there was an association. Babies exposed to ultrasound in the womb had lower birthweight, on average, than babies who were not exposed. Is this evidence that ultrasound causes lower birthweight?
>
> "*Discussion.* Obstetricians suggest ultrasound examinations when something seems to be wrong. The investigators concluded that the ultrasound exams and low birthweights had a common cause – problem pregnancies. Later, a randomized controlled experiment was done to get more definite evidence [with results reported in *Lancet* (September 10, 1988), pages 585-8]. If anything, ultrasound was protective."

Example 2.9 (Use of Cell Phones) (Agresti & Franklin, 2007, pages 147, 149, 180)

> “Cell phones have become the must-have communication gadget of the new millennium. There’s no doubt about it: The use of cell phones has become a routine part of our lives. But how safe are they? Cell phones emit electromagnetic radiation, produced in the form of nonionizing radio-frequency energy. A cellular phone’s antenna is the main source of this energy. The closer the antenna is to the user’s head, the greater the exposure to cell phone radiation.
>
> “With the increase in the popularity of cell phones has come a growing concern that heavy use of cell phones may increase a persons’ risk of getting cancer. Several studies have explored whether there’s an association between cell phone use and the occurrence of cancer. For instance,
>
> “**Study 1:** A German study [reported in *Epidemiology* (2001) 12(1):7-12] compared 118 patients with a rare form of eye cancer called uveal melanoma to 475 healthy patients who did not have the eye cancer. The patient’s cell phone use was measured using a questionnaire. The eye cancer patients used cell phones more often, on average.
>
> “**Study 2:** A U.S. study [reported in the *Journal of the American Medical Association* (2000) 284:3001-7] compared 469 patients with brain cancer to 422 patients who did not have brain cancer. The patients’ cell-phone use was measured using a questionnaire. The two groups’ use of cell phones was similar.
>
> “**Study 3:** An Australian study [reported in *Environmental Health Prospect* (1997) 105:1565-8] conducted an experiment with 200 transgenic mice, specially bred to be susceptible to cancers of the immune system. One hundred mice were exposed for two half-hour periods a day to the same kind of microwaves with roughly the same power as the kind transmitted from a cell phone. The other 100 mice were not exposed. After 18 months, the brain tumor rate for the mice exposed to cell phone radiation was twice as high as the brain tumor rate for the unexposed mice.
>
> “Studies 1 and 3 found an association between cell phone use and cancer. Study 2 did not. . . . [Note that] the Australian study was an experiment, using mice. The German and the U.S. studies both observed the amount of cell phone use for cancer patients and for non-cancer patients using a questionnaire.
>
> “. . . One reason that results of different medical studies sometimes disagree is that they are not the same *type* of study. An experimental study with mice is not directly comparable to an observational study with humans.
>
> “The U.S. and German studies . . . were both case-control studies [that is, retrospective observational studies in which subjects who have a response outcome of interest (the cases) and subjects who have the other response outcome (the controls) are compared on an explanatory variable]. In the U.S. study, the cases were brain cancer patients. The controls were hospitalized for benign conditions but were matched with the cases on age, gender, race, and month of admission. In the German study, the cases had a type of eye cancer. In forming a sample of controls, the German study did not attempt to match subjects with the cases.
>
> “[Since matching is an attempt to achieve the kind of balance that randomization would provide in an experimental setting, the] lack of matching in the German

> study may be one reason that the results from the two studies differed, the U.S. study not finding an association and the German study finding one. For example, in the German study suppose the eye cancer patients tended to be older than the controls and suppose that older people tend to be heavier users of cell phones. Then, age could be responsible for the observed association between cell-phone use and eye cancer."

Ethical considerations in randomization. The importance of the randomized comparative experiment in medical studies is underscored by the following quote from the *New England Journal of Medicine*:

> "Randomized, double-blind, placebo-controlled trials are the gold standard for evaluating new interventions and are routinely used to assess new medical therapies." (Source: Moore & Notz, 2006, page 115)

To date, there have been hundreds of thousands of such studies (Agresti & Franklin, 2007, page 170). Yet, people continue to question the ethics of randomization in experiments whose subjects are humans.

Random allocation is ethical when researchers are completely ignorant about which of several treatments is best. The results of observational studies (which might include a medical practitioner's observations over many years of clinical practice) might suggest that a particular treatment is best, but might not have accounted for all potential confounding variables. Randomized experiments give surer cause-and-effect results since the randomization will roughly balance the groups on any lurking variables.

Random samples versus randomization. Keep in mind that random samples and randomization are two different things:

1. In random samples, a chance method is used to determine the individuals in the sample.
2. In randomization, a chance method is used to determine the assignment of treatments to the individuals in a given sample.

The three most common types of random samples are simple random samples, stratified random samples and cluster random samples. Most introductory courses focus on methods for analyzing information from simple random samples. Different formulas are needed to analyze information from stratified or cluster random samples. (See the end of Chapter 6 for an example comparing the variability of sample means from simple random samples and from cluster random samples.)

Completely randomized designs and randomized block designs are two types of designs involving randomization. We will study methods for analyzing information from completely randomized designs with two treatments and from randomized block designs where each block has two individuals in Chapters 7 and 8.

Another interesting randomization scheme is the *group-randomized study design*, where entire groups of individuals are assigned to different treatments.

For example,

(1) In order to evaluate the impact of an intervention designed to change the instructional program and the school environment to increase support for physical activity among high school girls, researchers may randomly choose some schools to implement the proposed changes and others not to implement the changes.

(2) In order to evaluate the impact of a breast cancer screening program, researchers may randomly choose some community-based health centers to set up screening programs and others not to set up the programs. This may be necessary since widespread publicity is needed to encourage women to come for screening, or because members of the screening group might pass on information to neighbors who have been allocated to the "no screening" group, leading them to demand screening.

In group-randomized trials, randomization is done at the group level, information is collected at both the individual and group levels, and researchers are interested in drawing conclusions about differences between treatments at both the individual and group levels. Specialized formulas are needed to analyze information from these types of studies.

3 Tables, Graphs and Numerical Summaries

This chapter introduces *descriptive statistics*, that is, methods for organizing, summarizing and presenting the results of studies. References for this chapter include the texts by Agresti & Franklin (2007, Chapters 2 and 3), Baldi & Moore (2009, Chapters 1 through 6), Freedman et al (1991, Parts II and III), Moore & McCabe (1999, Chapters 1 and 2), Moore & Notz (2006, Part II) and Pagano & Gauvreau (2000, Chapters 2, 3 and 17).

3.1 Types of Variables

Recall that a *variable* is a characteristic of an individual. There are two broad types of variables — called categorical variables and quantitative variables — and several subtypes.

3.1.1 Categorical Variables

A *categorical variable* places an individual into one of several categories or classes. If there are exactly two classes, then the variable is said to be *dichotomous* (or *binary*).

The categorical variable is said to be *nominal* if the categories are unordered; it is said to be *ordinal* if there is a natural ordering to the classes.

Here are some examples:

(1) Suppose there are three political parties in a town (say Democrat, Republican and Independent) and that each registered voter has expressed an affiliation with one of the parties. Then party affiliation is a categorical variable with 3 classes. Since the classes are unordered, party affiliation is an example of a nominal variable.

(2) Researchers often use a 5-point scale to judge patient performance:

0. Patient fully active.
1. Patient is restricted in physically strenuous activity only.
2. Patient can handle self-care activities but not normal work activities.
3. Patient is capable of only limited self-care activities.
4. Patient is completely disabled.

Using this scale, patient performance status is a categorical variable with 5 classes. Since there is a natural ordering to the classes, patient performance status is an example of an ordinal variable.

(3) The response variable of interest in the Salk vaccine field trial example (page 13) was whether or not the child developed polio. Since the response variable is a categorical variable with 2 classes, it is an example of a dichotomous (or binary) variable.

3.1.2 Quantitative Variables

A *quantitative variable* is one for which arithmetic operations (such as adding or averaging) make sense. The quantitative variable is said to be *discrete* if its values form a finite set of numbers or a sequence of numbers such as $0, 1, 2, \ldots$. The quantitative variable is said to be *continuous* if its values form an interval on the real line.

Here are some examples:

(1) In a simple chance experiment, you decide to toss a fair coin 20 times and record the number of times a head appears on the top face. The recorded number of times can be any integer between 0 and 20. That is, the set of possible values is the finite set $\{0, 1, \ldots, 20\}$.

(2) In a simple chance experiment, you decide to toss a fair coin until a head appears on the top face and record the total number of tosses. The recorded total number is a discrete variable whose value can be any positive integer. That is, the set of possible values is the countably infinite set $\{1, 2, 3, \ldots\}$.

(3) Suppose that in a retrospective observational study of the relationship between smoking and lung cancer, the explanatory variable of interest is the average daily cigarette use in the ten-year period prior to onset of disease. Since average daily cigarette use can be any nonnegative real number (that is, any number greater than or equal to 0), the explanatory variable is a continuous variable whose values lie in the interval $[0, \infty)$.

Footnote. The distinctions among the types of variables introduced above are important. In general, the statistical methods we use (both descriptive and inferential methods) will depend on the types of variables we measure.

3.2 Data Tables, Distributions and Graphs

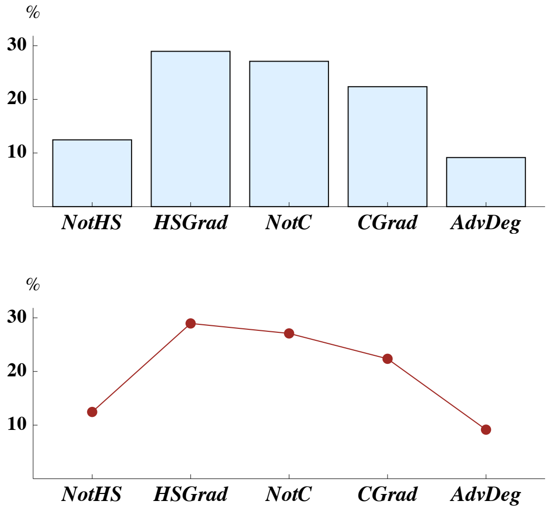
One of the simplest ways to summarize a set of observations is by using a table of frequencies and/or relative frequencies. A table is an appropriate summary tool for both categorical and quantitative data.

3.2.1 Frequency and Relative Frequency Distributions

The *distribution* of a variable tells us the values of the variable and how often each value is assumed. In a *frequency distribution*, raw counts are reported; in a *relative frequency distribution*, proportions or percents are recorded.

Example 3.1 (Educational Level of Adults) (Moore & Notz, 2006, page 183) The following table summarizes information from the 2003 Current Population Survey on the highest educational level of U.S. adults between 30 and 34 years old. Five categories are used: not a high school graduate, high school graduate, some college but not a college graduate, college graduate, and having earned an advanced degree.

Figure 3.1: *Bar chart (upper plot) and polygon plot (lower plot) of the educational levels data. Relative frequencies are reported as percents of individuals in each category.*



	NotHS	HSGrad	NotC	CGrad	AdvDeg	Total
Thousands of Persons	2554 (12.45)	5942 (28.95)	5559 (27.09)	4589 (22.36)	1878 (9.15)	20522 (100.00)

The table shows the frequency distribution in thousands and the relative frequency distribution as percents (in parentheses) for the more than 20.5 million adults in the study.

3.2.2 Bar Charts and Polygon Plots

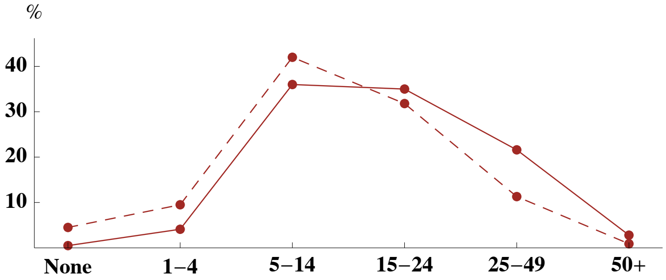
Bar charts and polygon plots are often used to plot categorical data.

In a *bar chart*, each category corresponds to a rectangle (a *bar*) with height equal to the frequency or relative frequency of the category. The bars are arranged so that all bars are of equal width and so that the bars do not touch.

In a *polygon plot*, each category corresponds to an ordered pair where the first coordinate is the center of the interval used to define the bar in a bar chart and the second coordinate is either the frequency or relative frequency of the category. Successive ordered pairs are connected by line segments.

Example 3.2 (Educational Level of Adults, continued) The upper part of Figure 3.1 (page 27) is a bar chart of relative frequencies for the educational levels data summarized in the table above; the lower part is a polygon plot of relative frequencies superimposed on the bar chart. In each case, the relative frequency distribution is reported as proportions.

Figure 3.2: *Polygon plots of smoking distributions for the lung cancer patients (solid line segments) and control patients (dashed line segments). Relative frequency distributions are reported as percents of individuals in each category.*



In applications, we would use one type of plot only (either the bar chart or the polygon plot). The polygon plot of relative frequencies is useful when we want to compare two or more distributions using the same set of axes.

Example 3.3 (Smoking and Lung Cancer) (Agresti, 1991, page 31) In a famous retrospective observational study of the relationship between smoking and lung cancer, Doll and Hill (*British Journal of Medicine* (1952) 2:1271-1286) collected information on the smoking habits of 1357 lung cancer patients among patients in hospitals in several English cities and on the smoking habits of 1357 patients from the same hospitals who had different respiratory diseases. The patients with other respiratory diseases were matched to the lung cancer patients on possible confounding variables.

For each individual, the researchers recorded average smoking use in 5 categories: no cigarette use on average, between 1 and 4 cigarettes per day on average, between 5 and 14 cigarettes per day on average, between 15 and 24 cigarettes per day on average, between 25 and 49 cigarettes per day on average, and 50 or more cigarettes per day on average

The following tables present the frequency distributions and relative frequency distributions as percents (in parentheses) for the lung cancer patients (the LC Patients) and for the patients with other respiratory diseases (the Control Patients).

	None	1-4	5-14	15-24	25-49	50+	Total
LC Patients	7 (0.5)	55 (4.1)	489 (36.0)	475 (35.0)	293 (21.6)	38 (2.8)	1357 (100.0)

	None	1-4	5-14	15-24	25-49	50+	Total
Control Patients	61 (4.5)	129 (9.5)	570 (42.0)	431 (31.8)	154 (11.3)	12 (0.9)	1357 (100.0)

Figure 3.2 (page 28) compares the relative frequency distributions as percentages for the two groups, where the distribution for lung cancer patients is plotted using solid line segments and the distribution for control patients is plotted using dashed line segments.

Notice that the proportions for lung cancer patients are generally shifted to higher levels of smoking than for control patients, suggesting a link between smoking and lung cancer.

Footnotes. The retrospective observational study reported in the example above is an example of a *case-control study*, where a fixed number of individuals with the response of interest (the cases) are compared to a fixed number of individuals who have the other response (the controls) on an explanatory variable. In good case-control studies, the researchers will attempt to match the cases and controls on possible confounding variables. But even with these precautions, a case-control study cannot be used to determine cause and effect definitively.

Smoking has been implicated in heart disease, lung cancer, and many other diseases. When considering the case against smoking, statisticians like Joseph Berkson and R.A. Fisher (page 19) disagreed with the approach taken by Doll and Hill, arguing that many important confounding variables were not controlled in their design. It was only after many studies were conducted — all with results similar to the ones reported above — that the medical community and the public at large accepted the causal link between smoking and various diseases (Freedman et al, 1991, page 12).

3.2.3 Modal Class

The *modal class* (or *mode*) of a categorical variable is the category with the highest frequency. The mode is the most common outcome. For example, in the educational level of adults example (page 26), the modal class is the class of high school graduates.

If a distribution has a single peak, it is said to be *unimodal.* If it has two distinct peaks, it is said to be *bimodal.*

Example 3.4 (Rating Sweetness) For example, suppose that 45 Australian and 48 Japanese consumers were asked to rate a particular brand of chocolate on a ten-point scale, where a score of 10 indicates the consumer liked the sweetness, while a score of 1 indicates a consumer did not like the sweetness at all.

The first table below gives the frequency distribution and the relative frequency distribution as percents (in parentheses) for the responses of the Australian consumers. The second table gives the same information for the responses of the Japanese consumers.

Australian Group ($n = 45$)

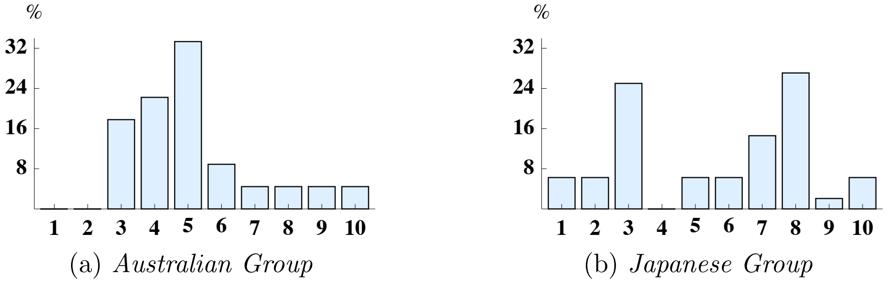
1	2	3	4	5	6	7	8	9	10	Total
0 (0.00)	0 (0.00)	8 (17.78)	10 (22.22)	15 (33.33)	4 (8.89)	2 (4.44)	2 (4.44)	2 (4.44)	2 (4.44)	45 (100.00)

Japanese Group ($n = 48$)

1	2	3	4	5	6	7	8	9	10	Total
3 (6.25)	3 (6.25)	12 (25.00)	0 (0.00)	3 (6.25)	3 (6.25)	7 (14.58)	13 (27.08)	1 (2.08)	3 (6.25)	48 (100.00)

Figure 3.3(a) (page 30) is a bar chart of the relative frequency distribution of sweetness ratings for the Australian sample. The distribution is unimodal: the most frequent sweetness category is 5. Figure 3.3(b) is a bar chart of the relative frequency distribution of sweetness ratings

Figure 3.3: *Comparison of relative frequency distributions of sweetness ratings. The distribution for the Australian group is unimodal, with modal class 5. The distribution for the Japanese group is bimodal, with separate modes at 3 and 8.*



(a) *Australian Group*

(b) *Japanese Group*

for the Japanese sample. In this case, the distribution is bimodal: the two most frequent sweetness categories are 3 and 8.

Footnote. The sweetness-rating scale used above is an example of an ordinal categorical variable (since the classes have a natural ordering) and of a discrete quantitative variable. It is interesting to note that average sweetness ratings for the two groups are close: 4.5 for the Australians versus 4.8 for the Japanese. But the shapes of the distributions are quite different. The Japanese consumers had stronger opinions either for or against the sweetness; the typical values (in terms of frequency) were far from the average for the group.

3.3 Stem Plots and Histograms

Stem plots and histograms are common graphical methods for quantitative data.

3.3.1 Stem Plots

A *stem plot* (or *stem-and-leaf plot*) gives a picture of the shape of a distribution while retaining all of the original observations. To construct a stem plot,

1. Separate each observation into a stem and a leaf. The *stem* of the observation consists of all but its last digit. The *leaf* of the observation is its last digit.
2. Write the stems in a column from smallest to largest. Draw a vertical line to the right of the column. Include enough stems so that the column is a representation of the number line.
3. Write each leaf in the row to the right of its stem, in increasing order.

Example 3.5 (Birthweights) Low birthweight is associated with many childhood disorders. In order to study the possible relationship between low birthweight and prenatal care, researchers examined the records of 14 newborns whose mothers visited their doctors 5 or fewer times during pregnancy (the 5 or Fewer Group) and 14 newborns whose mothers visited

Figure 3.4: *Stem plot (left) and back-to-back stem plot (right) of birthweights.*

5 or Fewer Group

4	9
5	2
6	
7	
8	2
9	36
10	18
11	04446
12	0
13	4

5 or Fewer Group (left) versus
6 or More Group (right)

9	4	
2	5	
	6	
	7	
2	8	7
63	9	378
81	10	68
64440	11	03699
0	12	9
4	13	1
	14	
	15	3

their doctors 6 or more times during pregnancy (the 6 or More Group). The following tables give the birthweights in ounces for each child in each group.

5 or Fewer Group ($n = 14$):

49	52	82	93	96	101	108	110	114	114	114	116	120	134

6 or More Group ($n = 14$):

87	93	97	98	106	108	110	113	116	119	119	129	131	153

The left part of Figure 3.4 (page 31) is a stem plot of birthweights for children in the 5 or Fewer group. The column of stems includes all integers between 4 (representing 40 ounces) and 13 (representing 130 ounces). The number of leaves in each row corresponds to the number of birthweights in the given 10-gram interval. For example, stem 10 has 2 leaves corresponding to the observed birthweights of 101 and 108.

Back-to-back stem plots can be used to compare two distributions. To construct a back-to-back stem plot,

1. Separate all observations into stems and leaves, as above.
2. Create a central column of stems serving as a representation of the number line long enough to plot the observations in both groups. Draw vertical lines to the left and right of the column.
3. Write each leaf for the first group in the row to the *left* of the stem, in increasing order as you move to the left. Write each leaf for the second group in the row to the *right* of the stem, in increasing order as you move to the right.

Example 3.6 (Birthweights, continued) The right part of Figure 3.4 (page 31) is a back-to-back stem plot of the birthweights data above. The stems include all integers between 4 (representing 40 ounces) and 15 (representing 150 ounces). This simple plot shows that most observations lie between 80 and 140, that the two observations below 80 are in the 5 or Fewer group and that the one observation above 140 is in the 6 or More group.

Although no firm conclusions can be drawn from this small study, the results do suggest that the problem be studied further. (In future studies, researchers should use larger samples and should include possible confounding variables.)

3.3.2 Histograms

Stem plots provide a quick way to visualize small data sets. By contrast, histograms can be used with both small and large data sets.

The key to understanding histograms is to note that histograms use *area* to represent numbers. To construct a histogram of quantitative data,

1. Subdivide the range of observed values into a convenient number of non-overlapping subintervals (called *class intervals*) so that each observation lies in one and only one class interval.
2. Compute the number of observations in each class interval.
3. For each class interval, draw a rectangle whose base is the given interval and whose height equals the proportion of observations in the interval divided by the length of the interval:

   $$\text{Height} = (\text{Proportion in Class Interval})/(\text{Length of Class Interval}).$$

   The height of the rectangle is called the *density* of the class interval.

Note that the area of each rectangle is the proportion of observations in the given class interval and the sum of the areas is exactly one.

Example 3.7 (Birthweights, continued) Figure 3.5(a) (page 33) is a histogram of the 14 birthweights in the 5 or Fewer group and Figure 3.5(b) is a histogram of the 14 birthweights in the 6 or More group.

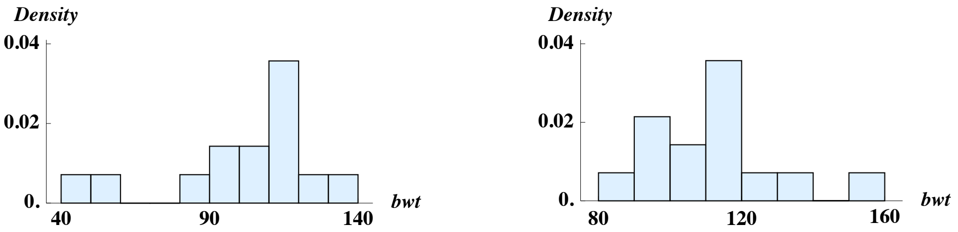
In part (a), the class intervals are $[40, 50)$, $[50, 60)$, …, $[130, 140)$. Since each subinterval is 10 units wide, the height of a given rectangle is the proportion of newborns whose birthweight falls in the interval divided by 10.

In part (b), the class intervals are $[80, 90)$, $[90, 100)$, …, $[150, 160)$.

Footnotes. The histograms constructed in this section are often called *density histograms*. We will see later how density histograms can be compared to "idealized distributions" for continuous data.

Although class intervals can be of different lengths, histograms using equal-length subintervals are easier to interpret. When the class intervals are of equal length, histograms are comparable to bar charts. In fact, researchers often use the terms "histogram" and "bar chart" interchangeably when they use equal-length subintervals.

Figure 3.5: *Histograms of birthweights for the two groups.*



(a) *5 or Fewer Group*

(b) *6 or More Group*

3.4 Quantiles and Additional Plots

This section discusses quantile summaries of quantitative data and introduces two additional plots for quantitative data, the one-way scatter plot and the box plot.

3.4.1 Computing Sample Quantiles

Let p be a proportion between 0 and 1 and let x_p be the number satisfying the following:

1. $100p\%$ of the population distribution lies to the left of x_p and
2. $100(1-p)\%$ of the population distribution lies to the right of x_p.

Then x_p is said to be the p^{th} *quantile* (or $100p^{\text{th}}$ *percentile*) of the distribution. Important special cases include the following:

1. The 50th percentile is known as the population *median.*
2. The 25th, 50th and 75th percentiles are known as the population *quartiles.*
3. The 10th, 20th, . . ., 90th percentiles are known as the population *deciles.*

This section considers estimating population quantiles from sample data using an interpolation method.

Order statistics. Starting with a sample of size n, the first step is to list the observations in increasing order, where the k^{th} observation in order is denoted by $x_{(k)}$,

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

$x_{(k)}$ is often called the k^{th} *order statistic.*

Sample p^{th} quantile. The general rule for finding the sample p^{th} quantile is

$$\text{Sample } p^{\text{th}} \text{ quantile} = (1-w)\; x_{(k)} \; + \; w\; x_{(k+1)} = x_{(k)} + w\Big(x_{(k+1)} - x_{(k)}\Big),$$

where k is the integer part and w is the fractional part of $(n+1)p$:

$$(n+1)p = \text{Integer Part} + \text{Fractional Part} = k + w.$$

This rule is valid as long as

$$\frac{1}{(n+1)} \leq p \leq \frac{n}{(n+1)}.$$

Otherwise, there is not enough information to estimate the population quantile.

Note that when $p = 1/2$, the procedure above reduces to the following simple formula:

$$\text{Sample Median} = \begin{cases} x_{((n+1)/2)} & \text{when } n \text{ is odd} \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right) & \text{when } n \text{ is even} \end{cases}$$

That is, the sample median is the middle number (in order) when n is odd and is the average of the two middle numbers when n is even.

Example 3.8 (Birthweights, continued) The left table below shows estimated quartiles for birthweights of children whose mother visited her doctor 5 or fewer times during pregnancy. The right table shows similar information for birthweights of children whose mother visited her doctor 6 or more times during pregnancy.

5 or Fewer Group

				Sample
p	$(n+1)p$	k	w	p^{th} Quantile
0.25	3.75	3	0.75	90.25
0.50	7.50	7	0.50	109.00
0.75	11.25	11	0.25	114.50

6 or More Group

				Sample
p	$(n+1)p$	k	w	p^{th} Quantile
0.25	3.75	3	0.75	97.75
0.50	7.50	7	0.50	111.50
0.75	11.25	11	0.25	121.50

Each table includes information that was used to compute the sample quantiles. Specifically, the second column shows $(n+1)p$, the third column shows k (the integer part of $(n+1)p$) and the fourth column shows w (the fractional part of $(n+1)p$). In each case, the sample quantile was computed using the formula on the previous page. Since both sample sizes are equal, the first four columns are the same in both tables.

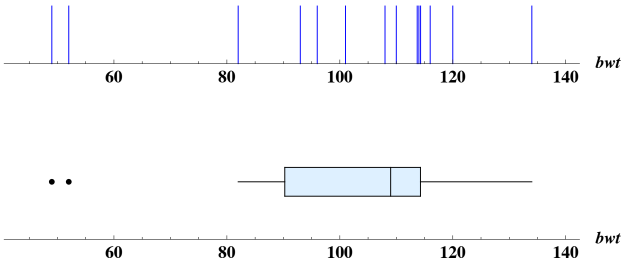
Notice that the sample quartiles for the first group are smaller than the corresponding sample quartiles for the second group.

3.4.2 One-Way Scatter Plots and Box Plots

One-way scatter plots and box plots are similar in that they each require a single real axis. In practice, box plots are used more often since plots for several distributions can be displayed in parallel for easy comparison.

One-way scatter plot. A one-way scatter plot uses a horizontal axis with values ranging from the minimum to the maximum observed value. Each observation corresponds to a

Figure 3.6: *The upper plot is a one-way scatter plot of the 14 birthweights in the 5 or Fewer Group. The lower plot is a box plot of the same data. There are two outliers.*



position on the line, represented using a short line segment. (Observations are visualized using position only.)

For example, the upper part of Figure 3.6 (page 35) is a one-way scatter plot of birthweights of the 14 children in the 5 or Fewer Group. Since the number 114 appears three times in the data set, three segments were drawn at positions at or near 114 on the numberline. (The data were *jiggled* so that the number of segments corresponds to the sample size.)

Box plot. A *box plot* (or *box-and-whisker plot*) is a visual display of quantitative data that uses sample quartiles (sample 25th, 50th and 75th percentiles), as described below.

To construct a box plot:

1. A *box* is drawn from the sample 25th percentile (q_1) to the sample 75th percentile (q_3).
2. A *bar* is drawn through the box at the sample median (q_2).
3. A *whisker* is drawn from q_3 to the largest observation that is less than or equal to $q_3 + 1.50(q_3 - q_1)$. Another whisker is drawn from q_1 to the smallest observation that is greater than or equal to $q_1 - 1.50(q_3 - q_1)$.
4. Observations outside the interval

   $$[q_1 - 1.50(q_3 - q_1), q_3 + 1.50(q_3 - q_1)]$$

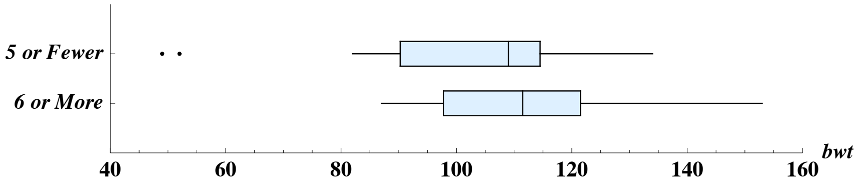
   are drawn as separate points. These observations are called the *outliers*.

The interval $[q_1 - 1.50(q_3 - q_1), q_3 + 1.50(q_3 - q_1)]$ is often called the *whisker interval*. Observations in this interval are either covered by the box or by one of the whiskers.

Example 3.9 (Birthweights, continued) The lower part of Figure 3.6 (page 35) is a box plot of the 14 birthweights in the 5 or Fewer Group. For these data, the sample quartiles are

$$q_1 = 90.25,\ q_2 = 109.00 \text{ and } q_3 = 114.50,$$

Figure 3.7: *Parallel box plots of birthweights for the two groups.*



and the whisker interval is $[q_1 - 1.50(q_3 - q_1), q_3 + 1.50(q_3 - q_1)] = [53.875, 150.875]$.

Eight observations lie within the central box. Twelve observations lie within the whisker interval. The smallest and largest of the 12 observations in the whisker interval are the endpoints of the whiskers. The remaining two observations (49 and 52) are the outliers and are represented as separate points.

Two or more distributions can be compared graphically using parallel box plots.

Example 3.10 (Birthweights, continued) Figure 3.7 (page 36) compares birthweight distributions for the 14 children whose mothers visited their doctors 5 or fewer times during pregnancy and for the 14 children whose mothers visited their doctors 6 or more times during pregnancy using parallel box plots. Notice, in particular, that the central part of the "6 or More" distribution is an approximate shift to higher values of the central part of the "5 or Fewer" distribution.

Example 3.11 (Cholesterol Reduction and Compliance) The Lipid Research Clinics Coronary Primary Prevention Trial (LRC-CPPT) was a placebo-controlled double-blind randomized comparative experiment concerning the efficacy of the drug cholestyramine for lowering cholesterol level and thereby reducing coronary artery disease (Efron & Feldman, *Journal of the American Statistical Association* (1991) 86:9-26).

The sample consisted of 335 men between 35 and 59 years old with an initial serum cholesterol level of 265 milligrams per deciliter (mg/dL) or more. Of these men, 164 were randomly assigned to the treatment group and the remaining 171 were assigned to the control group. Each man was supposed to take 6 packets of the "medication" per day for a fixed period of time. At the end of the study, the researchers recorded each subject's cholesterol reduction,

$$\text{CR} = \text{Initial Cholesterol Level} - \text{Final Cholesterol Level},$$

and compliance with the treatment protocol as a percent,

$$\text{Compliance} = \text{Percent of Packets Actually Used}.$$

Compliance with treatment protocol is a variable that cannot be controlled in a randomized experiment. It may be that better compliers are better patients to begin with. Further, although compliance can be adjusted for at the end of the study by, for example, comparing results for the subgroup of men in the treatment group with "good compliance" to the subgroup of men in the control group with "good compliance," compliance has different meanings in the treatment and control groups. In the treatment group, compliance determines the amount of active drug taken by a subject and also indicates something about the patient's psychological status. In the control group, where each subject receives no active drug, only the psychological component of compliance applies.

For the reasons mentioned above, and for other reasons not stated here, the appropriate analysis of experimental studies where compliance is an issue is an important and difficult statistical problem.

This example considers the 164 men who received the proposed cholesterol-reducing drug and uses box plots to explore how compliance and cholesterol reduction are related.

The tables below give cholesterol reduction for the 164 men who received the active drug, where the men have been grouped by their compliance to treatment into one of four classes: 0 to 24 percent compliance, 25 to 49 percent compliance, 50 to 74 percent compliance, and 75 to 100 percent compliance.

0 to 24 Percent Compliance Group ($n = 37$)*:*

−23.00	−21.00	−16.50	−13.00	−10.50	−10.25	−10.25	−7.25	−6.25	−5.75	−5.25
−0.50	−0.50	0.25	2.50	3.00	3.25	3.50	4.50	5.50	5.75	7.50
8.25	8.75	10.75	11.25	11.50	17.25	19.50	19.75	21.00	21.25	24.00
29.25	29.75	36.25	39.00							

25 to 49 Percent Compliance Group ($n = 26$)*:*

−19.00	−15.75	−1.50	−1.50	1.00	3.25	4.25	4.25	5.50	6.00	7.75
8.50	10.50	15.75	18.75	20.25	23.50	25.50	27.75	30.75	32.50	33.00
33.25	33.50	36.25	56.75							

50 to 74 Percent Compliance Group ($n = 25$)*:*

−20.75	−14.50	−4.25	0.00	13.75	14.75	14.75	18.00	22.50	27.75	30.25
39.50	41.25	42.00	42.75	43.00	44.50	46.25	47.25	48.75	54.50	59.50
62.75	63.00	64.25								

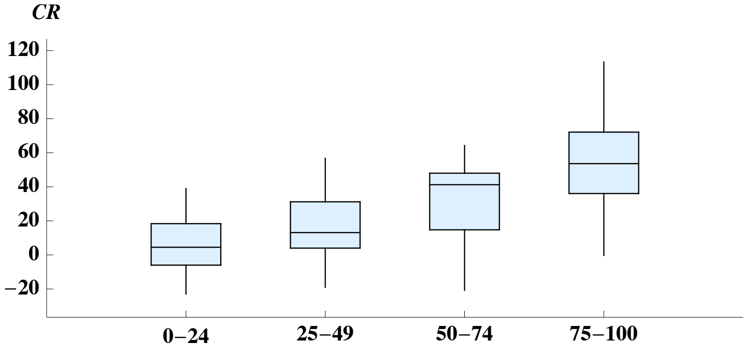
75 to 100 Percent Compliance Group ($n = 76$)*:*

−0.25	1.00	1.00	2.75	3.25	6.00	9.75	11.50	14.00	18.25	21.00
26.75	26.75	28.75	29.25	29.50	30.25	32.50	36.00	36.25	37.75	39.00
39.00	39.50	40.25	41.00	41.25	41.75	42.25	44.75	46.50	46.75	47.50
48.50	48.75	51.00	51.50	53.50	53.75	54.00	54.75	54.75	56.00	56.75
59.75	60.00	61.25	61.75	62.50	66.50	68.00	68.00	69.00	69.50	70.00
70.75	71.00	72.50	73.00	73.75	75.75	76.00	77.00	78.75	79.00	80.00
80.00	82.75	85.00	86.00	86.75	92.00	94.25	101.50	104.75	113.25	

Numbers are listed in increasing order within each group. Values range from a minimum of −23.00 (an increase of 23 over the study period) to a maximum of 113.25 (a decrease of 113.25 over the study period) and exhibit quite a bit of variability.

Figure 3.8 (page 38) compares cholesterol reduction in the four compliance groups defined above using parallel box plots. The plot suggests that compliance and cholesterol-reduction are associated (although the exact nature of the association is still in question). Notice, in particular, that median values increase with increasing compliance. The median values for the four compliance groups are 4.5, 13.125, 41.25 and 53.625, respectively.

Figure 3.8: *Parallel box plots of cholesterol reduction in 4 percent-compliance groups.*



3.5 Measures of Center and Spread

This section considers two commonly used measures of the center of a sample of observations (the sample median and the sample mean) and two commonly used measures of the spread of the sample (the sample interquartile range and the sample standard deviation).

3.5.1 Median and Interquartile Range

Natural measures of the center and spread of a sample distribution can be based on the sample quartiles, q_1, q_2 and q_3. The *sample median* (or *median*) is defined as follows:

$$\text{Median} = q_2.$$

The *sample interquartile range* (or *interquartile range*) is defined as follows:

$$\text{IQR} = q_3 - q_1.$$

That is, the median is the sample 50^{th} percentile and the interquartile range is the length of the interval from the sample 25^{th} to the sample 75^{th} percentiles.

Example 3.12 (Birthweights, continued) Continuing with the birthweights and prenatal care example, the computations on page 34 imply that

- The median and IQR of the 5 or Fewer Group are 109 ounces and 24.25 ounces, respectively; and
- The median and IQR of the 6 or More Group are 111.5 ounces and 23.75 ounces, respectively.

3.5.2 Mean and Standard Deviation

Let x_1 be the value of the variable for the first individual, x_2 the value of the variable for the second individual, and so forth. The *sample mean* (or *mean*) of the n observations is defined as follows:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n\right)$$

(read "x-bar"). The sample mean is the simple average of the n observations.

The middle term in the definition of $\overline{x}$ uses Σ-notation ("Sigma-notation") to represent the sum of the n observations. Σ-notation is a convenient way to represent sums. In this case, the Σ-notation specifies that we should sum all x_i's, starting with $i = 1$ and ending with $i = n$.

The *sample variance* (or *variance*) of the observations is the average of the squares of the differences $(x_i - \overline{x})$, where $n - 1$ is used in the average formula instead of n,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\Big((x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2\Big).$$

The *sample standard deviation* (or *standard deviation*) of the observations is the square root of the sample variance, $s = \sqrt{s^2}$.

The standard deviation is a measure of the average distance between each observation and the overall mean. The fact that squares and square roots are used in the formulas should remind you of finding distances between points in the plane or in three-space. The fact that $(n-1)$ is used instead of n when computing the average reflects the fact that we are working with information from a sample and not from an entire population.

Example 3.13 (Birthweights, continued) Let the x_i's represent the observations in the 5 or Fewer Group, and the y_i's represent the observations in the 6 or More Group.

Using 2 decimal places of accuracy,

1. the mean of observations in the 5 or Fewer Group is

$$\overline{x} = \frac{1}{14}\sum_{i=1}^{14} x_i = \frac{1}{14}\left(49 + 52 + \cdots + 134\right) = \frac{1403}{14} = 100.21 \text{ ounces and}$$

2. the mean of observations in the 6 or More Group is

$$\overline{y} = \frac{1}{14}\sum_{i=1}^{14} y_i = \frac{1}{14}\left(87 + 93 + \cdots + 153\right) = \frac{1579}{14} = 112.79 \text{ ounces.}$$

The computations needed to find the standard deviations are as follows:

5 or Fewer Group

x_i	$(x_i - \overline{x})$	$(x_i - \overline{x})^2$
49	−51.21	2622.4641
52	−48.21	2324.2041
82	−18.21	331.6041
93	−7.21	51.9841
96	−4.21	17.7241
101	0.79	0.6241
108	7.79	60.6841
110	9.79	95.8441
114	13.79	190.1641
114	13.79	190.1641
114	13.79	190.1641
116	15.79	249.3241
120	19.79	391.6441
134	33.79	1141.7641
1403	0.06	7858.3574

6 or More Group

y_i	$(y_i - \overline{y})$	$(y_i - \overline{y})^2$
87	−25.79	665.1241
93	−19.79	391.6441
97	−15.79	249.3241
98	−14.79	218.7441
106	−6.79	46.1041
108	−4.79	22.9441
110	−2.79	7.7841
113	0.21	0.0441
116	3.21	10.3041
119	6.21	38.5641
119	6.21	38.5641
129	16.21	262.7641
131	18.21	331.6041
153	40.21	1616.8441
1579	−0.06	3900.3574

Note that the sums of the differences $(x_i - \overline{x})$ would be exactly 0 if more decimal places of accuracy were used to compute $\overline{x}$. Similarly, the sum of the differences $(y_i - \overline{y})$ would be exactly 0 if more decimal places of accuracy were used to compute $\overline{y}$.

Using 2 decimal places of accuracy,

1. the standard deviation of observations in the 5 or Fewer Group is

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{7858.3574}{13}} = \sqrt{604.489} = 24.59 \text{ ounces and}$$

2. the standard deviation of observations in the 6 or More Group is

$$s_y = \sqrt{s_y^2} = \sqrt{\frac{3900.3574}{13}} = \sqrt{300.027} = 17.32 \text{ ounces.}$$

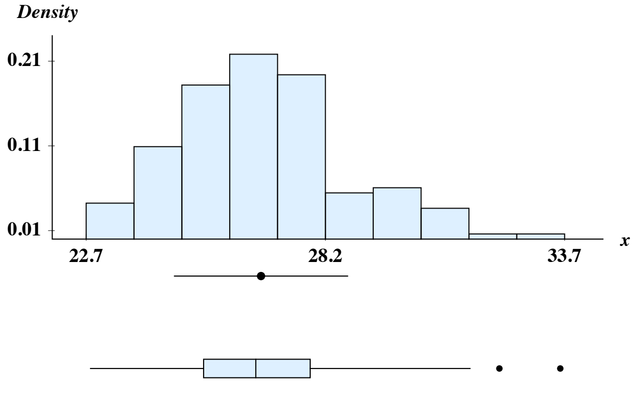
Example 3.14 (Waist Measurements) (Heinz et al, *Journal of Statistics Education* Volume 11, Number 2 (2003), `www.amstat.org/publications/jse/v11n2`) As part of a study investigating the correspondence between body build, weight, and girths in a group of physically active men and women, researchers gathered information on more than 500 adults. This example considers waist measurements (in inches) for 150 women under 30 who participated in the study.

The following table gives the waist measurements for these women in increasing order.

Waist Measurement Data ($n = 150$)

22.8	22.8	23.1	23.1	23.2	23.4	23.7	23.8	23.8	23.9	23.9	24.0	24.0	24.1
24.2	24.2	24.3	24.3	24.5	24.6	24.7	24.7	24.7	24.8	24.8	24.9	24.9	25.0
25.0	25.1	25.1	25.2	25.2	25.2	25.3	25.4	25.4	25.4	25.5	25.6	25.6	25.6
25.6	25.6	25.6	25.6	25.7	25.7	25.8	25.8	25.8	25.9	25.9	25.9	25.9	26.0
26.0	26.0	26.0	26.0	26.0	26.0	26.1	26.1	26.2	26.2	26.2	26.3	26.3	26.4
26.4	26.5	26.5	26.5	26.6	26.6	26.6	26.6	26.7	26.7	26.7	26.8	26.8	26.8
26.8	26.9	26.9	26.9	27.0	27.0	27.0	27.1	27.2	27.2	27.2	27.2	27.4	27.4
27.4	27.4	27.5	27.5	27.5	27.6	27.6	27.6	27.6	27.6	27.7	27.7	27.7	27.8
27.8	28.0	28.0	28.0	28.0	28.0	28.0	28.1	28.1	28.1	28.1	28.2	28.4	28.5
28.5	28.6	28.8	28.9	29.0	29.2	29.3	29.4	29.5	29.5	29.5	29.7	29.7	29.8
29.8	29.9	30.7	30.7	30.8	31.0	31.3	31.5	32.2	33.6				

Figure 3.9: *The upper plot is a histogram of the waist measurement data. A line segment from $\overline{x} - s$ to $\overline{x} + s$ is drawn under the horizontal axis with a point marking the mean, $\overline{x}$. The lower plot is a box plot of the same sample. Note that there are two outliers.*



Using 2-decimal place accuracy, the sample mean is $\overline{x} = 26.72$ inches and the sample standard deviation is $s = 1.99$ inches.

The upper part of Figure 3.9 (page 41) is a histogram of the waist measurement data using 10 class intervals:

$$[22.7, 23.8),\ [23.8, 24.9),\ \ldots,\ [32.6, 33.7).$$

A line segment from $\overline{x} - s = 24.73$ to $\overline{x} + s = 28.71$ is drawn under the horizontal axis of the histogram with a point marking the position of the mean $\overline{x} = 26.72$.

A box plot of the waist measurement data is displayed in the lower part of Figure 3.9 (page 41).

The following tables summarize the computations needed to construct the plot.

				Sample
p	$(n+1)p$	k	w	p^{th} Quantile
0.25	37.75	37	0.75	25.40
0.50	75.50	75	0.50	26.60
0.75	113.25	113	0.25	27.85

Median = 26.6 inches
IQR = 2.45 inches

Whisker Interval: [21.725, 31.525]
Outliers: 32.2, 33.6

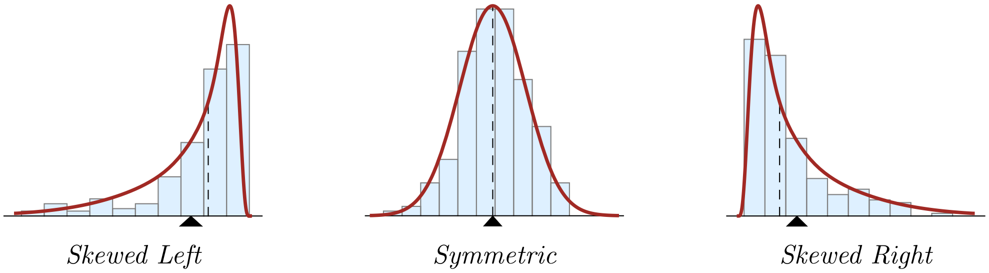
Notice, in particular, that two observations lie outside the whisker interval. These observations (the outliers) are displayed as separate points in the box plot.

Footnotes. A plot of the interval $[\overline{x} - s, \overline{x} + s]$ with a point at $\overline{x}$ (the mean) is similar to a box plot of the $[q_1, q_3]$ interval with a bar at q_2 (the median).

But, the box plot gives more information. Specifically,

1. The central box is constructed to enclose 50% of the sample distribution, with 25% in the subbox for the $[q_1, q_2]$ interval and 25% in the subbox for the $[q_2, q_3]$ interval;

Figure 3.10: *Histograms for observations from 3 types of population distributions with idealized curves superimposed. Dashed lines are drawn at population medians. Triangles are drawn at population means.*



2. The whiskers extend the box to incorporate observations that are “reasonably close” to the central part of the distribution; and
3. The plot of outliers shows the unusual observations.

Be sure you know how to use the formulas for finding the sample median, interquartile range, mean and standard deviation. You should be able to apply the formulas in problems where the number of observations is small.

3.5.3 Comparison of Measures

The mean and standard deviation are the preferred measures of center and spread when the observed distribution is approximately symmetric with no outliers or with few outliers. The median and interquartile range are the preferred measures when distributions are highly skewed with extreme outliers.

For example, consider the three histograms for observations from three types of distributions (skewed left, symmetric and skewed right) shown in Figure 3.10 (page 42). In each case, a density curve for an idealized population distribution has been superimposed, a vertical dashed line has been drawn at the location of the population median and a triangle has been drawn at the location of the population mean.

In symmetric population distributions, the mean and median are equal. Their common value represents both the average value and the point at which 50% of the distribution lies to the left and 50% of the distribution lies to the right.

If a population distribution is skewed left, the mean will be to the left of the median. If a population distribution is skewed right, the mean will be to the right of the median. In both cases, most observations will be closer to the median than to the mean.

For population distributions that are highly skewed (to the left or right), the locations of the mean and median could be quite far apart. Since most observations are closer to the median than to the mean, the median is a better measure of center in skewed distributions.

A similar argument can be made for preferring the interquartile range to the standard deviation in highly skewed distributions. In particular, if the distribution is highly skewed, then some observations will be very far from the sample mean and will contribute a large amount to the sample standard deviation.

3.6 Contingency Tables and Association

Contingency tables allow us to study associations among categorical variables.

3.6.1 Two-Way Tables, Joint and Marginal Distributions

A *joint distribution* is the distribution of two or more variables. In *two-way tables*, joint frequency distributions and/or joint relative frequency distributions of two categorical variables are reported.

Example 3.15 (CHD and Anger) (Moore & Notz, 2006, page 496) As part of a study on coronary heart disease (CHD), researchers gathered information on 8474 people with normal blood pressure. Each subject was rated using the Spielberger Trait Anger Scale, a method of determining how prone an individual is to sudden anger.

The following two-way table summarizes the results of the study. The rows in the table correspond to levels of CHD (present or absent). The columns correspond to levels of the anger score (low, moderate or high).

	Low	Moderate	High	Total
CHD Present	53 (0.6)	110 (1.3)	27 (0.3)	190 (2.2)
CHD Absent	3057 (36.1)	4621 (54.5)	606 (7.2)	8284 (97.8)
Total	3110 (36.7)	4731 (55.8)	633 (7.5)	8474 (100.0)

The table contains both frequencies and relative frequencies as percents (in parentheses). For example, 190 (2.2% of 8474) subjects had CHD, 3110 (36.7% of 8474) subjects scored low on the anger scale, and 53 (0.6% of 8474) had CHD and scored low on the anger scale.

Marginal distributions. The *margins* of two-way tables can be used to report the distributions of the row and column variables. For this reason, the distributions of the row and column variables are often called the *marginal distributions*.

Example 3.16 (CHD and Anger, continued) The bottom row of the table above gives the marginal distribution of the anger variable.

The right column of the table gives the marginal distribution of the CHD variable. Since CHD is dichotomous, the distribution could easily be reported using a single percent. That is, it is sufficient to report that 2.2% (190/8474) of the subjects had CHD.

3.6.2 Conditional Distributions and Association

Conditional distributions are used to study the association between row and column variables in a two-way table. Either

- the distributions of the column variable for each value of the row variable or
- the distributions of the row variable for each value of the column variable

are used. The distribution of one variable given a fixed value of another variable is called a *conditional distribution.*

Example 3.17 (CHD and Anger, continued) The left table below shows the conditional distribution of anger scores for subjects with CHD. The right table gives the same information for subjects without CHD.

Subjects with CHD

Low	Medium	High	Total
53 (28%)	110 (58%)	27 (14%)	190 (100%)

Subjects without CHD

Low	Medium	High	Total
3057 (37%)	4621 (56%)	606 (7%)	8284 (100%)

Notice, in particular, that 14% (27/190) of subjects with CHD were rated high for sudden anger compared to only 7% (606/8284) of subjects without CHD.

Since CHD is dichotomous, the three conditional distributions of CHD given anger score can be reported using three percents:

	Low	Medium	High
CHD Present	1.7% (53/3110)	2.3% (110/4731)	4.3% (27/633)

Notice that the percent of subjects with CHD increases with increasing anger score.

Footnotes. In the extended example above, CHD is a response variable and anger is an explanatory variable. The association between these variables was explored using (1) the conditional distributions of anger given level of CHD and (2) the conditional distributions of CHD given level of anger. Each method suggests that anger and CHD are associated.

3.6.3 Controlling for a Variable

An explicit way of controlling for a potential confounder to the relationship between explanatory and response variables (when all three variables are categorical) is to construct and study a three-way contingency table.

Example 3.18 (Smoking and Health, revisited) Consider again the study relating cigarette use and twenty-year survival beginning on page 10. Age at initial interview was confounded with smoking.

The following table shows the joint frequency distribution of the two-level smoking variable, the two-level survival variable and the four-level age variable used in the example.

		18-34	35-54	55-64	65+	Total
Smoker	Alive	174	198	64	7	443
	Dead	5	41	51	42	139
Nonsmoker	Alive	213	180	81	28	502
	Dead	6	19	40	165	230
Total		398	438	236	242	1314

To control for age, we consider the conditional distributions of survival given fixed values of the joint smoking-age distribution. Since survival is dichotomous, the eight conditional distributions can be reported using eight percents.

The following table shows the percent and proportion (in parentheses) of women alive after twenty years in each of the eight groups defined using the joint smoking-age distribution.

	18-34	35-54	55-64	65+
Smoker	97.2% (174/179)	82.8% (198/239)	55.7% (64/115)	14.3% (7/49)
Nonsmoker	97.3% (213/219)	90.5% (180/199)	66.9% (81/121)	14.5% (28/193)

Within each age group, the nonsmokers were more likely to survive than the smokers.

3.7 Scatter Plots and Association

Scatter plots are the most common way to display relationships between quantitative variables. This section introduces scatter plots and numerical methods for describing the association between two quantitative variables.

3.7.1 Two-Way Scatter Plots

A *two-way scatter plot* (or *scatter plot*) is a graphical method to display the relationship between two quantitative variables measured on the same individuals. For a given sample of size n, the ordered pairs

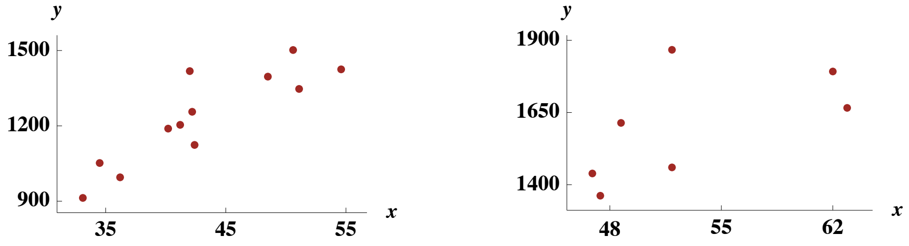
$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n)$$

are plotted on the same set of axes, where x_i is the value of the first variable and y_i is the value of the second variable for the i^{th} individual.

Example 3.19 (LBM and Metabolic Rate) (Moore & Notz, 2006, pages 282-3) As part of a study on dieting, researchers collected information on the lean body mass and resting metabolic rate for 12 female subjects.

Lean body mass (LBM) is a person's weight minus all fat. Metabolic rate is the rate at which the body consumes energy. Researchers believe that LBM is an important influence on metabolic rate.

Figure 3.11: *Scatter plots of resting metabolic rate (vertical axis) versus lean body mass (horizontal axis). Part (a) displays information for the 12 females subjects in the study. Part (b) displays information for the 7 male subjects in the study.*



(a) *Female Subjects*

(b) *Male Subjects*

The following table gives LBM measured in kilograms (the x_i's) and resting metabolic rate measured in calories burned per 24 hours (the y_i's) for the 12 women in the study. Each column of the table corresponds to one ordered pair.

x_i	33.1	34.5	36.2	40.2	41.2	42.0	42.2	42.4	48.5	50.6	51.1	54.6
y_i	913	1052	995	1189	1204	1418	1256	1124	1396	1502	1347	1425

Figure 3.11(a) (page 46) is a two-way scatter plot of resting metabolic rate (vertical axis) versus lean body mass (horizontal axis) for the 12 women. The plot suggests that as one variable increases the other generally does as well.

Footnote: Positive and negative association. Two variables are said to have a *positive association* if as one variable increases the other tends to increase. They are said to have a *negative association* if as one variable increases the other tends to decrease. The two-way scatterplot shown in Figure 3.11(a) (page 46) suggests that LBM and resting metabolic rate are positively associated.

3.7.2 Correlation

If the data pairs (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) have a roughly straight line trend, then the variables are said to have an approximate *linear relationship*. The *sample correlation* (or *correlation*) measures the direction and strength of this linear relationship.

The computation of the sample correlation, denoted by r, takes several steps:

1. Compute the mean ($\overline{x}$) and standard deviation (s_x) of the x_i's.
2. Compute the mean ($\overline{y}$) and standard deviation (s_y) of the y_i's.
3. The correlation is the average of the products of *standardized scores* for each individual,

where $(n-1)$ is used instead of n to compute the average:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{s_x}\right)\left(\frac{y_i-\overline{y}}{s_y}\right).$$

Example 3.20 (LBM and Metabolic Rate, continued) The following tables show the computations needed to find the means and standard deviations of lean body mass and resting metabolic rate for the female subjects.

Lean Body Mass

x_i	$(x_i-\overline{x})$	$(x_i-\overline{x})^2$
33.1	-9.95	99.0025
34.5	-8.55	73.1025
36.2	-6.85	46.9225
40.2	-2.85	8.1225
41.2	-1.85	3.4225
42.0	-1.05	1.1025
42.2	-0.85	0.7225
42.4	-0.65	0.4225
48.5	5.45	29.7025
50.6	7.55	5.70025
51.1	8.05	64.8025
54.6	11.55	133.4025
516.6	0.00	517.7300

Resting Metabolic Rate

y_i	$(y_i-\overline{y})$	$(y_i-\overline{y})^2$
913	-322.08	103735.5264
1052	-183.08	33518.2864
995	-240.08	57638.4064
1189	-46.08	2123.3664
1204	-31.08	965.9664
1418	182.92	33459.7264
1256	20.92	437.6464
1124	-111.08	12338.7664
1396	160.92	25895.2464
1502	266.92	71246.2864
1347	111.92	12526.0864
1425	189.92	36069.6064
14821	0.04	389954.9168

Using the left table (and 2-decimal place accuracy),

$$\overline{x} = \frac{516.6}{12} = 43.05 \text{ kg and } s_x = \sqrt{\frac{517.73}{11}} = 6.86 \text{ kg.}$$

Similarly, using the right table (and 2-decimal place accuracy),

$$\overline{y} = \frac{14821}{12} = 1235.08 \text{ cal/day and } s_y = \sqrt{\frac{389954.9168}{11}} = 188.28 \text{ cal/day.}$$

The correlation r is obtained by averaging products of standardized scores for lean body mass and resting metabolic rate. For simplicity, we use the notations

$$z_x = \frac{\text{Lean Body Mass} - 43.05}{6.86} \quad \text{and} \quad z_y = \frac{\text{Resting Metabolic Rate} - 1235.08}{188.28}$$

for the scores. The following table gives their values with 2 decimal-place accuracy.

z_x	-1.45	-1.25	-1.00	-0.42	-0.27	-0.15	-0.12	-0.09	0.79	1.10	1.17	1.68
z_y	-1.71	-0.97	-1.28	-0.24	-0.17	0.97	0.11	-0.59	0.85	1.42	0.59	1.01

Finally, using the table of standardized scores and 2 decimal-place accuracy,

$$r = \frac{1}{n-1}\sum_{i=1}^{n} z_{x_i} z_{y_i} = \frac{1}{11}\left((-1.45)(-1.71) + (-1.25)(-0.97) + \cdots + (1.68)(1.01)\right) = 0.88.$$

Interpretation. The sample correlation r takes values between -1 and 1. A positive value of r indicates a positive association between the variables and a negative value of r indicates a negative association between the variables.

Values of r near ± 1 indicate a *strong* linear association between the variables. Values of r near 0 indicate a *weak* linear association between the variables.

The value of r in the example above suggests a strong positive linear association between lean body mass and resting metabolic rate.

Example 3.21 (LBM and Metabolic Rate, continued) The researchers also collected information on the lean body mass and resting metabolic rate for 7 male subjects.

The following table gives LBM measured in kilograms (x_i's) and resting metabolic rate measured in calories burned per 24 hours (y_i's) for the 7 men in the study.

x_i	46.9	47.4	48.7	51.9	51.9	62.0	62.9
y_i	1439	1362	1614	1460	1867	1792	1666

For these data, $\overline{x} = 53.1$, $s_x = 6.69$, $\overline{y} = 1600$, $s_y = 189.24$ and $r = 0.59$.

Figure 3.11(b) (page 46) is a two-way scatter plot of resting metabolic rate (vertical axis) versus lean body mass (horizontal axis) for the 7 men. The results suggest a moderate positive linear association between lean body mass and resting metabolic rate.

The observed association was stronger for the women in the study than for the men.

Footnotes. The correlation r is the average of products of standardized scores for the two variables under study. The value of r does not depend on which variable is labeled "x" and which variable is labeled "y".

The *standardized score* (or *z-score*) for an observation is the number of standard deviations that the observation falls from the mean. Standardized scores have no units. For example, if x represents lean body mass in kilograms, then the numerator and denominator of each z-score ($x_i - \overline{x}$ and s_x, respectively) are measured in kilograms, but the ratio $z_i = (x_i - \overline{x})/s_x$ is a pure number. As the average of products of numbers with no units, the correlation r is also a pure number.

3.7.3 Correlation, Regression and Prediction

If the data pairs (x_1, y_1), (x_2, y_2), …, (x_n, y_n) have a roughly straight line trend, x is an explanatory variable and y is a response variable, then a *regression line*

$$\widehat{y} = a + bx,$$

(where "$\widehat{y}$" is read "y-hat"), can be used to predict y given x. In this setting,

1. $\widehat{y_i} = a + bx_i$ is called the i^{th} *predicted response* and
2. $e_i = y_i - \widehat{y_i}$ is called the i^{th} *prediction error* (or i^{th} *residual*).

Least squares method. The *method of least squares* is the most commonly used method to find the slope and intercept of a regression line. In this method, the slope and intercept are chosen to minimize the sum of squares of the prediction errors. That is, to minimize

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2.$$

Using this method, the formulas for the slope and intercept are

$$b = r(s_y/s_x) \text{ and } a = \overline{y} - b\overline{x}, \text{ respectively.}$$

For example, the slope of the least squares regression line for the female subjects in the lean body mass and resting metabolic rate study is

$$b = r\left(\frac{s_y}{s_x}\right) = 0.88\left(\frac{188.28}{6.86}\right) = 24.05,$$

the intercept of the line is $a = \overline{y} - b\overline{x} = 1235.08 - 24.05(43.05) = 199.64$ and the regression equation is

$$\widehat{y} = 199.64 + 24.05x \quad \text{for} \quad 33.1 \leq x \leq 54.6.$$

An individual whose lean body mass is 45 kilograms, for example, is predicted to have a resting metabolic rate of $199.64 + 24.05(45) = 1281.89$ calories per day. Note that the regression equation is valid for values of x in the observed range of lean body mass only, since we have no information beyond this range.

The interpretation of the slope of a regression line is important. In this case, the slope tells us that if lean body mass is increased by 1 kilogram, then resting metabolic rate is predicted to increase by 24.05 calories per day.

A scatter plot of (x_i, y_i) pairs, with the regression line $\widehat{y} = 199.64 + 24.05x$ superimposed, is shown in Figure 3.12(a) (page 50).

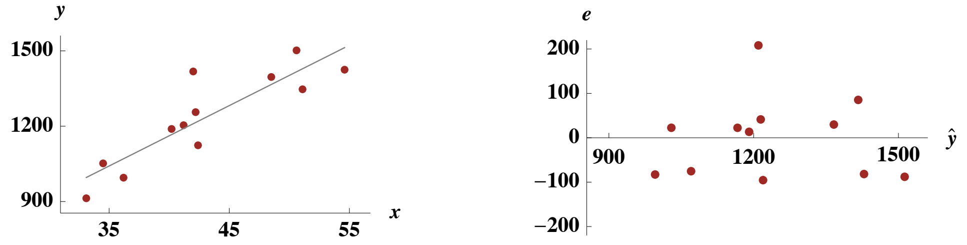
Residual plots. A *residual plot* is a two-way scatter plot of the prediction errors on the vertical axis versus either the predicted responses or a potential confounding variable on the horizontal axis.

Residuals measure the size of prediction errors. A residual of 0 says that a given (x_i, y_i) pair lies exactly on the regression line. A positive residual says that the pair lies above the line and a negative residual says that the pair lies below the line. In general, there will be some positive and some negative residuals. The sum will always be zero.

Residual plots of prediction errors versus predicted responses can be used to help assess the fit of a regression line. For example, Figure 3.12(b) (page 50) is such a plot for the female subjects in the lean body mass and resting metabolic rate study. Although the regression line fits the data reasonably well in this case, there is one unusually large residual.

Residual plots of prediction errors versus potential confounding variables can be used to help assess if the presence of that variable would change the observed association between the

Figure 3.12: *Part (a) is a scatter plot of resting metabolic rate (vertical axis) versus lean body mass (horizontal axis) for the 12 women with the regression line* $\widehat{y} = 199.64 + 24.05x$ *superimposed. Part (b) is a residual plot of errors (vertical axis) versus predicted resting metabolic rates (horizontal axis) for the same sample.*



(a) *Metabolic Rate* (y) *versus LBM* (x) (b) *Residuals* (e) *versus Predictions* $(\widehat{y})$

explanatory and response variables (that is, to assess if there is an important variable *lurking* in the background). For example, researchers often plot residuals versus the time order of the observations. A pattern here would suggest that the relationship between the explanatory and response variables has changed over time.

Footnotes. The least squares regression method uses what is often called the *five-number summary* of the paired data. That is, the values of slope and intercept can be computed once we know the five numbers $\overline{x}$, $\overline{y}$, s_x, s_y and r.

Although the correlation r is a pure number (has no units), the slope of the least squares regression line $b = r(s_y/s_x)$ is not. The slope b is the rate of change of the response variable with respect to a unit change in the explanatory variable. In the examples studied in this section, the rate is in "calories per day per kilogram".

3.8 Brief Summary and Additional Examples

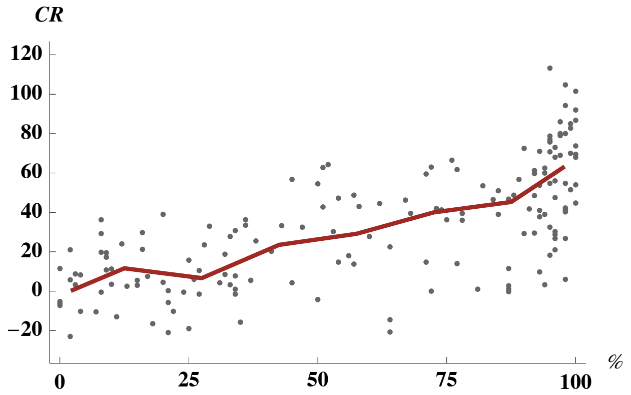
This chapter introduces different types of variables, and methods (both numerical and graphical) for summarizing and describing observed data and associations.

The emphasis is on summary and description only. To generalize from observed data, we would need to know that the data were produced properly (as studied in the last chapter) and we would need to use methods from probability theory and inferential statistics (the subject of the remaining chapters of this book).

Loss of information. Keep in mind that summary can result in the loss of valuable information. For example, if researchers in the cigarette use and twenty-year survival study example beginning on page 10, and revisited in this chapter, had summarized across age groups, then the true association would have been lost.

Alternative methods. There may be more than one way to summarize and describe a set of observations. For example, in the cholesterol reduction and compliance study example

Figure 3.13: *Scatter plot of cholesterol reduction (vertical axis) versus percent compliance (horizontal axis). An average smooth has been added to the plot.*



beginning on page 36, I created an ordinal categorical variable for compliance with 4 levels and used parallel box plots to explore the association between cholesterol reduction and compliance in the group of men who received the active drug.

An alternative graphical method is the scatter plot of cholesterol reduction versus compliance shown in Figure 3.13 (page 51). The plot does not exhibit a roughly linear trend. To summarize the relationship, I used a *scatter plot smoothing method*. Specifically, I divided the compliance interval into 8 subintervals,

$$[0,5),\ [5,20),\ [20,35),\ [35,50),\ [50,65),\ [65,80),\ [80,95),\ [95,100],$$

computed the average cholesterol reduction for men whose compliance fell in a given interval, plotted the 8 points of the form

(Midpoint of Interval, Average Cholesterol Reduction)

and connected successive points with straight line segments. The eight points are as follows:

Midpt	2.5	12.5	27.5	42.5	57.5	72.5	87.5	97.5
AvgCR	0.6	11.6	6.6	23.5	29.1	40.0	45.3	62.7

Both summary methods (using medians and parallel box plots and using a scatter plot with average smoothing) make the strong point that cholesterol reduction and compliance are related, even though the exact relationship may be hard to quantify.

Center and spread, revisited. The mean and standard deviation are the most commonly used measures of center and spread. They are good choices when distributions are approximately symmetric with no outliers, or approximately symmetric with a few outliers that lie close to the endpoints of the whiskers in a box plot. Otherwise, the median and interquartile range are better measures of center and spread.

Table 3.1: *Percentage of children immunized against diphtheria, pertussis and tetanus (DPT) and under-five mortality rate per 1000 live births in 20 countries for 1992.*

	Percentage Immunized	Mortality Rate		Percentage Immunized	Mortality Rate
Bolivia	77	118	Greece	54	9
Brazil	69	65	India	89	124
Cambodia	32	184	Italy	95	10
Canada	85	8	Japan	87	6
China	94	43	Mexico	91	33
Czech Republic	99	12	Poland	98	16
Egypt	89	55	Russian Federation	73	32
Ethiopia	13	208	Senegal	47	145
Finland	95	7	Turkey	76	87
France	95	9	United Kingdom	90	9

(*Source*: Pagano & Gauvreau, 2000, Table 17.1, page 399)

Since the sample correlation and the least squares method of finding the slope and intercept of a regression line use means and standard deviations, we should be cautious when applying these summaries to paired data.

Example 3.22 (Immunization and Mortality) Table 3.1 (page 52) lists information on the percentage of children immunized against diphtheria, pertussis and tetanus (DPT) and the under-five mortality rate per 1000 live births in 20 countries for 1992. These data are of interest since the under-five mortality rate is an important indicator of well-being for a population of children (Pagano & Gauvreau, 2000, page 398).

The following list is the percentage immunized data in increasing order:

$$13, 32, 47, 54, 69, 73, 76, 77, 85, 87, 89, 89, 90, 91, 94, 95, 95, 95, 98, 99.$$

The percentage immunized ranges from 13 (Ethiopia) to 99 (Czech Republic), with a mean of 77.4 percent and standard deviation of 23.7 percent.

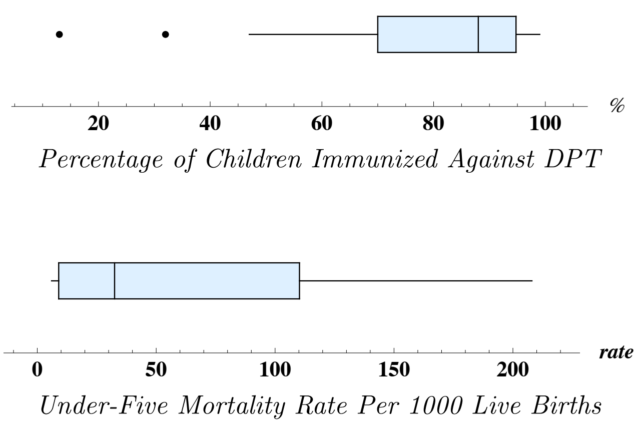
The following list is the mortality rate data in increasing order:

$$6, 7, 8, 9, 9, 9, 10, 12, 16, 32, 33, 43, 55, 65, 87, 118, 124, 145, 184, 208.$$

The mortality rate ranges from 6 deaths per 1000 live births (Japan) to 208 deaths per 1000 live births (Ethiopia), with a mean of 59 deaths per 1000 live births and a standard deviation of 63.9 deaths per 1000 live births.

Figure 3.14 (page 53) shows box plots of the percentage immunized and mortality rate data. Neither distribution is approximately symmetric. The following tables summarize the computations needed to find sample quartiles in each case.

Figure 3.14: *Box plots of the immunization and mortality study data.*



Percentage of Children Immunized Against DPT

Under-Five Mortality Rate Per 1000 Live Births

Percentage Immunized

p	$(n+1)p$	k	w	Sample p^{th} Quantile
0.25	5.25	5	0.25	70.00
0.50	10.50	10	0.50	88.00
0.75	15.75	15	0.75	94.75

Mortality Rate

p	$(n+1)p$	k	w	Sample p^{th} Quantile
0.25	5.25	5	0.25	9.00
0.50	10.50	10	0.50	32.50
0.75	15.75	15	0.75	110.25

The median percentage immunized is 88 percent. The interquartile range is 24.75 percent. The whisker interval is $[32.875, 131.875]$. There are two outliers (observations lying outside the whisker interval): 13 (Ethiopia) and 32 (Cambodia).

The median mortality rate is 32.5 deaths per 1000 live births. The interquartile range is 101.25 deaths per 1000 live births. The whisker interval is $[-142.875, 262.125]$. There are no outliers.

The correlation between percentage immunized and mortality rate is $r = -0.79$, indicating a fairly strong negative association between the variables.

Using percentage immunized as the explanatory variable (x) and mortality rate as the response variable (y), the slope of the least squares regression line is

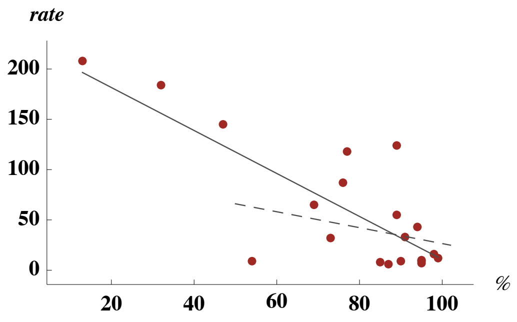
$$b = r\left(\frac{s_y}{s_x}\right) = -0.79\left(\frac{63.9}{23.7}\right) = -2.14,$$

the intercept is $a = \overline{y} - b\overline{x} = 59 + 2.14(77.4) = 224.32$, and the regression equation is

$$\widehat{y} = 224.32 - 2.14x, \quad \text{for} \quad 3 \leq x \leq 99.$$

If the percentage immunized is increased by 1 percent, then the mortality rate for a country is predicted to decrease by 2.14 deaths per 1000 live births.

Figure 3.15: *Scatter plot of mortality rate (vertical axis) versus percentage immunized (horizontal axis). The solid gray line is the least squares regression line for all 20 pairs. The dashed gray line is the least squares regression line for the rightmost 17 pairs.*



An examination of the two-way scatter plot of (x_i, y_i) pairs shown in Figure 3.15 (page 54), however, should warn us to be cautious about using the correlation and regression line summaries above. The least squares regression line is the solid gray line in the plot.

I recomputed the correlation using all countries except Ethiopia and Cambodia (whose data correspond to the uppermost two points in the upper left corner of the plot) and got $r = -0.50$. The observed negative association is now much less strong.

Since there are three apparent outliers in the plot — and to be dramatic — I recomputed the correlation using all countries except Ethiopia, Cambodia and Senegal (whose data correspond to the three points in the upper left corner of the plot) and got $r = -0.24$. I then computed the least squares regression line for the 17 countries,

$$\widehat{y} = 105.557 - 0.79x \quad \text{for} \quad 54 \leq x \leq 99,$$

and added the line to the plot as a dashed line. The relationship between the explanatory and response variables is much weaker when the three outliers are removed.

The lesson here is not that outliers should be removed before doing an analysis. In fact, outliers are often the most interesting observations. The lesson is that methods developed to be used in one situation (for example, with data that are approximately symmetric with no or few outliers) may not be appropriate summaries in other situations.

Finally, you should keep in mind that association does not imply causation (as we learned in the last chapter). Even if the negative association between percentage immunized and under-five mortality rate is real, we cannot conclude definitively that immunization causes a reduction in mortality rate.

Table 3.2: *Height in inches (x) and Weight in pounds (y) for 150 women under 30.*

x	y	x	y	x	y	x	y	x	y	x	y
58.0	109.8	62.8	111.6	63.5	139.8	64.7	122.4	66.0	159.4	67.1	149.5
59.8	101.0	62.9	110.2	63.8	120.6	64.8	132.9	66.1	130.1	67.3	119.5
59.8	130.7	62.9	117.3	63.8	121.3	65.0	118.2	66.2	108.5	67.6	123.9
60.0	102.5	63.0	103.6	63.8	118.2	65.0	123.9	66.2	117.7	68.0	134.7
60.0	104.3	63.0	107.6	63.8	130.5	65.0	125.2	66.3	120.8	68.0	136.7
61.3	118.2	63.0	110.2	63.9	110.7	65.0	159.4	66.3	143.7	68.0	153.2
61.4	120.4	63.0	110.7	64.0	115.1	65.2	132.3	66.5	121.3	68.0	155.4
61.4	142.0	63.0	110.7	64.0	128.3	65.2	161.2	66.5	133.4	68.1	137.8
61.5	132.3	63.0	119.9	64.0	120.2	65.4	117.3	66.5	137.3	68.1	131.8
61.8	105.4	63.0	120.8	64.0	135.4	65.4	153.9	66.5	138.9	68.2	152.6
61.8	138.9	63.0	130.1	64.0	146.4	65.5	124.8	66.5	152.1	68.5	122.4
62.0	107.1	63.0	132.3	64.1	127.4	65.7	124.8	66.5	128.3	68.5	162.3
62.0	119.7	63.1	114.9	64.1	127.9	65.7	126.1	66.7	139.8	69.0	137.3
62.0	120.2	63.3	123.2	64.1	131.0	65.7	131.8	66.7	148.4	69.0	147.3
62.2	122.4	63.3	152.3	64.2	136.7	65.8	137.1	66.9	123.2	69.0	140.2
62.3	102.3	63.3	119.9	64.2	158.7	65.9	130.1	66.9	130.1	69.0	140.2
62.5	108.2	63.4	118.2	64.3	119.0	66.0	116.2	66.9	160.7	69.0	144.4
62.5	122.4	63.5	113.8	64.3	123.2	66.0	120.2	67.0	120.2	69.4	147.3
62.6	107.1	63.5	120.8	64.3	124.3	66.0	128.5	67.0	123.2	69.4	148.1
62.6	122.6	63.5	121.7	64.3	131.8	66.0	134.5	67.0	130.3	69.5	158.3
62.6	104.9	63.5	126.3	64.4	114.2	66.0	135.4	67.0	140.2	69.5	162.3
62.7	114.2	63.5	132.7	64.5	129.0	66.0	138.2	67.0	160.5	70.0	132.3
62.8	100.8	63.5	132.9	64.6	118.6	66.0	138.9	67.0	162.3	70.1	155.6
62.8	104.9	63.5	134.3	64.6	157.8	66.0	142.2	67.0	142.9	70.8	147.3
62.8	108.5	63.5	149.7	64.7	131.8	66.0	151.7	67.1	142.2	71.0	133.8

(*Source:* Heinz et al (2003), `www.amstat.org/publications/jse/v11n2`)

Least squares regression line and average smooth. If the data pairs (x_i, y_i), for $i = 1, 2, \ldots, n$, have a roughly straight line trend, x is an explanatory variable and y is a response variable, then the least squares regression line can be thought of as an "average smooth" of the scatterplot.

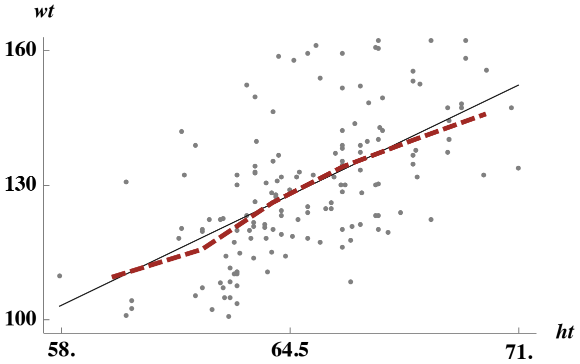
Example 3.23 (Height and Weight) For example, consider again the study investigating the correspondence between body build, weight, and girths in a group of physically active men and women (introduced in the waist measurements example beginning on page 40), and the subsample of 150 women under 30 who participated in the study.

Table 3.2 (page 55) gives the height in inches and weight in pounds for these women, where height (x) is considered to be an explanatory variable and weight (y) a response variable, and Figure 3.16 (page 56) is a scatterplot of (x, y) pairs.

Using 2 decimal-place accuracy, summary measures are as follows:

Height (inches)	Weight (pounds)	Correlation
$\overline{x} = 65.00$, $s_x = 2.44$	$\overline{y} = 129.67$, $s_y = 15.60$	$r = 0.59$

Figure 3.16: *Scatterplot of weight (vertical axis) versus height (horizontal axis). The solid line is the least squares regression line for the height-weight data. The dashed piecewise linear curve is an average smooth of the data.*



The slope of the least squares regression line is

$$b = r\left(\frac{s_y}{s_x}\right) = 0.59\left(\frac{15.60}{2.44}\right) = 3.77,$$

the intercept is $a = \overline{y} - b\overline{x} = 129.67 - 3.77(65) = -115.38$, and the regression equation is

$$\widehat{y} = -115.38 + 3.77x, \quad \text{for} \quad 58 \leq x \leq 71.$$

If height is increased by 1 inch, then weight is predicted to increase by 3.77 pounds. The least squares regression line is the solid line in Figure 3.16.

For the average smooth, I subdivided the height interval into 6 subintervals,

$$[58, 61),\ [61, 63),\ [63, 65),\ [65, 67),\ [67, 69),\ [69, 71],$$

computed the average weight for women whose heights fell in the given interval, plotted the 6 points of the form

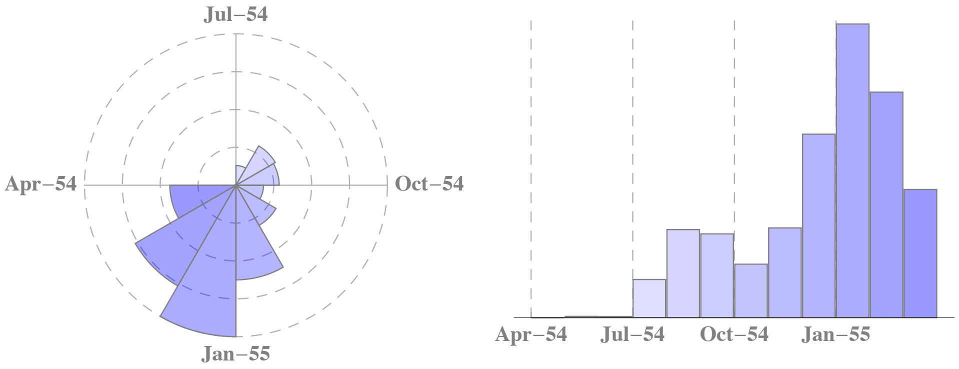
(Midpoint of Interval, Average Weight)

and connected successive points using dashed line segments in the figure. The 6 points are given in the following table

Midpt	59.5	62	64	66	68	70
AvgWT	109.7	115.7	126.1	134.1	140.1	145.7

Since the data pairs have a roughly straight line trend, the piecewise linear curve for the average smooth and the least squares regression line are close.

Figure 3.17: *Graphical representations of numbers of soldiers who died of preventable diseases during a one year period of the Crimean War. The left part of the figure is a rose diagram, and the right part is a histogram, of the information.*



(*Source:* `https://understandinguncertainty.org/node/214`)

Florence Nightingale's rose diagram. One of the most famous graphical representations of data was developed by Florence Nightingale, the "passionate statistician" who was introduced on the first page of this book.

While managing the nursing staff at a field hospital during the Crimean War, Nightingale realized that soldiers were dying more often from preventable diseases than from the wounds they had suffered in battle. After instituting basic sanitary standards in early 1855, the numbers of deaths from disease declined.

The left part of Figure 3.17 (page 57) is a "rose diagram" of the numbers of deaths from disease that occurred each month during the one year period from the beginning of April 1854 to the end of March 1855. Nightingale used this type of diagram to convince the military to institute similar reforms at all field hospitals.

Rose diagrams are not in common use today. Researchers prefer to use histograms (like the one shown in the right part of the figure) or even "pie charts" to represent information. Still, it is interesting to note that simple graphical representations can have powerful consequences.

Exploratory and confirmatory data analysis. Numerical and graphical methods used to summarize and describe data are often referred to as *exploratory data analysis* methods. By contrast, methods used to make inferences are often called *confirmatory data analysis* methods. Exploratory analysis should precede confirmatory analysis.

The importance of the exploratory phase of data analysis was underscored by the work of John W. Tukey, who also developed some of the simple techniques in this chapter.

John W. Tukey (1915–2000). John Wilder Tukey was one of the most influential statisticians of the second half of the twentieth century. He was born in New Bedford, Massachusetts, and was home schooled until completing his secondary school education. He earned undergraduate and graduate degrees in chemistry from Brown University, and graduate degrees in mathematics from Princeton University. He was a distinguished faculty member at Princeton until he retired.

Tukey was a man of many interests and talents. He was an avid consultant; he worked, for example, for the Educational Testing Service, the Xerox Corporation and Merck & Company. And he was a member of the research staff of AT&T's Bell Laboratories. He also had a penchant for coining new words. For example, he was the first person to use the word *software*. (He contrasted the words *hardware*, *software* and *brainware* in discussions of the potential impact of the then newly-developed digital computer.) In addition, he was the first person to use the word *bit* for binary digit.

Tukey developed, along with James Cooley, a computer algorithm known as the Fast Fourier Transform (FFT). The FFT is arguably the most important algorithm ever developed. The FFT is used, for example, in medical imaging devices.

Tukey made significant contributions to both theoretical and applied statistics. In the area of applied statistics, he advocated the use of exploratory methods so that researchers would be sure to understand their data before trying to apply sophisticated mathematical tools that may not be appropriate. He famously said

> "An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem."

He developed the *stem-and-leaf plot*, the *box-and-whisker plot*, and many other interesting, simple, and colorfully-named techniques for exploring data.

4 Introduction to Probability Theory

Probability theory is the study of random phenomena, and serves as the foundation for statistical inference (the subject of later chapters). This chapter is the first of two chapters introducing probability theory and its applications. References for these chapters include the texts by Agresti & Franklin (2007, Chapters 5–6), Baldi & Moore (2009, Chapters 9–12), Freedman et al (1991, Parts IV–V), Moore & McCabe (1999, Chapters 4–5), Moore & Notz (2006, Part III) and Pagano & Gauvreau (2000, Chapters 6–7).

4.1 Experiments and Sample Spaces

The term *experiment* (or *random experiment* or *chance experiment*) is used in probability theory to describe a procedure whose outcome is not known in advance with certainty. Further, experiments are assumed to be repeatable (at least in theory) and to have a well-defined set of possible outcomes. The *sample space* $\mathcal{S}$ is the set of all possible outcomes of an experiment.

Here are some examples:

(1) In a simple chance experiment, you decide to toss a coin 5 times and record h (for head) or t (for tail) each time. Each outcome is a list of 5 h's and t's and the sample space is the set of all possible lists of 5 h's and t's:

$$\begin{aligned}\mathcal{S} = \ &\{hhhhh, hhhht, hhhth, hhthh, hthhh, thhhh, hhhtt, hhtht, hthht, thhht,\\ &hhtth, hthth, thhth, htthh, ththh, tthhh, ttthh, tthth, thtth, httth, tthht,\\ &ththt, httht, thhtt, hthtt, hhttt, htttt, thttt, tthtt, tttht, tttth, ttttt\}.\end{aligned}$$

(2) In a simple chance experiment, you decide to toss a coin 5 times and record the total number of times a head appeared on the top face. Each outcome is a whole number between 0 and 5. The sample space is the set

$$\mathcal{S} = \{0, 1, 2, 3, 4, 5\}.$$

(3) A researcher is interested in determining if regular followup phone calls to the caretakers of asthmatic children is a useful intervention technique. (The phone calls would be used to check on the status of the children and to suggest changes to treatment if needed.) As a first step, the researcher decides to choose a simple random sample of 100 asthmatic children from among the 1250 asthmatic children regularly seen at a given health center. The technique described on page 8 will be used to choose the sample. Each outcome of the first step is a random subset of size $n = 100$ from the study population of size $N = 1250$. The sample space $\mathcal{S}$ for the first step is the set of all possible subsets of size 100 from the study population. Note that there are more than 8.9×10^{149} possible subsets.

(4) Continuing with the setup of the previous example, as a second step our researcher decides to randomly assign 50 of the 100 children chosen for the study to the "treatment group" (the group whose caretakers will receive followup phone calls between health center visits) with the remaining 50 children assigned to the "control group" (the group whose caretakers will not receive followup phone calls between visits). The technique described on page 12 will be used to make the random assignments. Each outcome of the second step is a random subset of size $n_1 = 50$ from the set of $n = 100$ children chosen for the study. The sample space $\mathcal{S}$ for the second step is the set of all possible subsets of size 50 from the set of children chosen for the study. Note that there are more than 10^{29} possible subsets.

Footnotes. The researcher given in the scenario above has explicitly introduced chance into the design of the study. This will allow the researcher to use the methods of statistical inference to generalize results from the sample to the study population, but *not* beyond the study population. If the study population is sufficiently like other potential study populations (for example, if the asthmatic children seen at the given health center are sufficiently like the asthmatic children seen at a different health center), then practitioners at other health centers may find the results interesting.

Suppose instead that our researcher decides to enroll into the study the first 100 asthmatic children seen at the health center after a given start date, and to randomly assign 50 children to the treatment group and 50 to the control group. Since chance has been used in the study design, methods of statistical inference can be used to generalize results. But the generalization does *not* go beyond the sample used by the researcher. If it is reasonable to believe that the 100 children enrolled in the study are representative of the children in the study population (that is, if the *convenience sample* is sufficiently like a *simple random sample* from the study population), then the results of the study can still be used to help determine if a followup phone call program should be instituted.

Note that both scenarios (using a simple random sample followed by randomization or using a convenience sample followed by randomization) assume that caretakers can be reached by phone and would be willing to take the phone calls.

4.2 Events, Operations on Events and Probability

An *event* is a subset of the sample space $\mathcal{S}$. We usually use capital letters (A, B, C, $\ldots$) to denote events and write "$A \subseteq \mathcal{S}$" (literally, "A is a subset of $\mathcal{S}$") when A is an event.

We usually use lower case letters (x, y, z, $\ldots$) to denote outcomes and write "$x \in A$" (literally, "x is an element of A") if the outcome x is a member of the event A. If $x \in A$ is observed, we say that event A has *occurred.*

The *probability* of event A, denoted by $P(A)$, is the proportion of times event A will occur in a sufficiently long series of repetitions of the experiment.

Here are two examples:

(1) In a simple chance experiment, you decide to toss a coin 5 times and record h (for head) or t (for tail) each time. Each outcome is a list of 5 h's and t's and the sample space is the set of all possible lists of 5 h's and t's:

$$\begin{aligned}\mathcal{S} = \quad & \{hhhhh, hhhht, hhhth, hhthh, hthhh, thhhh, hhhtt, hhtht, hthht, thhht,\\ & hhtth, hthth, thhth, htthh, ththh, tthhh, ttthh, tthth, thtth, httth, tthht,\\ & ththt, httht, thhtt, hthtt, hhttt, htttt, thttt, tthtt, tttht, tttth, ttttt\}.\end{aligned}$$

Let A be the event "exactly 3 heads are observed in 5 tosses," then

$$A = \{hhhtt, hhtht, hthht, thhht, hhtth, hthth, thhth, htthh, ththh, tthhh\}.$$

If the coin is fair (that is, if heads appears on the top face 50% of the time and tails appears 50% of the time), then $P(A)$ is the ratio of the number of elements in A to the number of elements in $\mathcal{S}$: $P(A) = 10/32 = 0.3125$.

(2) Among the 1250 asthmatic children regularly seen at a given health center, 600 are girls and 650 are boys. You decide to choose and review the case file of one of these children by implementing the following procedure:

1. Each child is assigned a unique number between 1 and 1250.
2. The computer is used to randomly select 1 number from the set $\{1, 2, \ldots, 1250\}$.
3. The case file of the child whose number is chosen is reviewed.

Let A be the event "the case file of a girl is chosen for review" and B be the event "the case file of a boy is chosen for review." Then $P(A)$ is the ratio of the number of girls to the number of children and $P(B)$ is the ratio of the number of boys to the number of children:

$$P(A) = \frac{600}{1250} = 0.48 \quad \text{and} \quad P(B) = \frac{650}{1250} = 0.52.$$

Footnotes. In the simple examples above, it was easy to determine the numbers of elements in the sample space and in the events of interest. Further, since each outcome was equally likely, the probability of each event was the ratio of the number of elements in the event to the number of elements in the sample space.

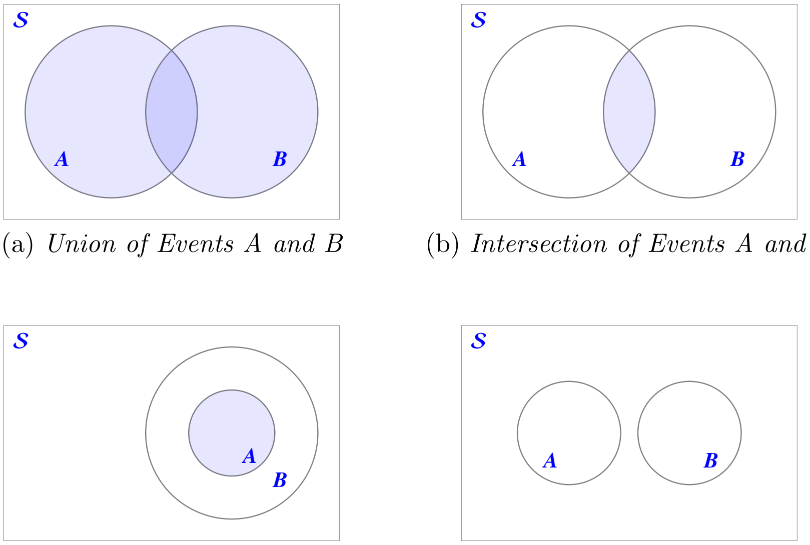
In many practical situations, computing probabilities can be challenging. The techniques introduced in the remaining sections of this chapter will help you compute and interpret probabilities in more challenging situations.

4.2.1 Set Operations

Sets and operations on sets are fundamental to the study of probability. Here we review the concepts of union, intersection and complement, discuss DeMorgan's laws for relating complements to unions and intersections, and introduce the notions of pairwise disjoint sets and partitions.

Union and intersection. The *union* of the events A and B (denoted by $A \cup B$) is the set of elements that belong to either A or B, as illustrated in Figure 4.1(a) (page 62). The *intersection* of the events A and B (denoted by $A \cap B$) is the set of elements that belong to both A and B, as illustrated in Figure 4.1(b).

Figure 4.1: *Venn diagrams for union and intersection. In part (a), the shaded area represents the union $A \cup B$, In part (b), the shaded area represents the intersection $A \cap B$. In part (c), the shaded area is $A \cap B = A$ since $A \subseteq B$. In part (d), there is no shaded area since $A \cap B = \emptyset$. In each part, the enclosing rectangular region represents the sample space $\mathcal{S}$.*



(a) *Union of Events A and B*

(b) *Intersection of Events A and B*

(c) *Intersection When A Is a Subset of B*

(d) *Intersection When A and B Have No Elements in Common*

For example, suppose there are 85 students in a course. After grading the first quiz, the professor decides to choose and review one paper (as a check on grading). The sample space for this experiment is the collection of 85 papers. Let A be the event "there is a perfect score on question 1" and B be the event "there is a perfect score on question 2." Then $A \cup B$ is the collection of papers with perfect scores on either question 1 or question 2 or on both questions, and $A \cap B$ is the collection of papers with perfect scores on both questions.

If A is a subset of B (denoted by $A \subseteq B$), then $A \cap B = A$, as illustrated in Figure 4.1(c). For example, if every paper with a perfect score on question 1 also has a perfect score on question 2 in the quiz example above, then $A \cap B = A$. Note that in this case $A \cup B = B$.

If A and B have no elements in common, then $A \cap B = \emptyset$, as illustrated in Figure 4.1(d). Note that the notation $\emptyset$ is used to denote the *empty set* (the set with no elements).

If A, B and C are events, then

1. the operations of union and intersection are *commutative*, that is,

$$A \cup B = B \cup A \quad \text{and} \quad A \cap B = B \cap A;$$

2. the operations of union and intersection are *associative*, that is,

$$(A \cup B) \cup C = A \cup (B \cup C) \quad \text{and} \quad (A \cap B) \cap C = A \cap (B \cap C);$$

3. the operations of union and intersection satisfy the following *distributive* laws:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \text{ and } A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

We will find each of these properties of union and intersection useful. Note that the second property says that the union of three events $A \cup B \cup C$ is well-defined and the intersection of three events $A \cap B \cap C$ is well-defined.

For example, if A is the event "the paper has a perfect score on the first question," B is the event "the paper has a perfect score on the second question," and C is the event "the paper has a perfect score on the third question" in the quiz review scenario, then $A \cup B \cup C$ is the set of papers with perfect scores on at least one of the first three questions and $A \cap B \cap C$ is the set of papers with perfect scores on all of the first three questions.

As an example of the distributive law $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, consider again the quiz review scenario. Let A be the event "the paper belongs to a male student in the class," B be the event "the paper has a perfect score on question 1," and C be the event "the paper has a perfect score on question 2." Then $A \cap (B \cup C)$ is the set of papers submitted by male students with perfect scores on either question 1 or question 2. And, $(A \cap B) \cup (A \cap C)$ is the set of papers either submitted by a male student with a perfect score on question 1 or submitted by a male student with a perfect score on question 2. From the descriptions, it is clear that these two sets are the same.

Complement. The *complement* of event A (denoted by A^c) is the set of all elements in the sample space that do not belong to A, as illustrated in Figure 4.2(a) (page 64).

For example, consider again the quiz review scenario above. If A is the event "the paper belongs to a male student in the class," then A^c is the event "the paper belongs to a female student in the class."

Note that, for each event A,

1. A and A^c have no elements in common: $A \cap A^c = \emptyset$.
2. Every element of the sample space is either in A or in A^c: $A \cup A^c = \mathcal{S}$.
3. The complement of A^c is the original set A: $(A^c)^c = A$.

De Morgan's laws. Let A and B be events. *De Morgan's laws* relate complements to unions and intersections. Specifically,

1. The complement of the union equals the intersection of the complements. In notation,

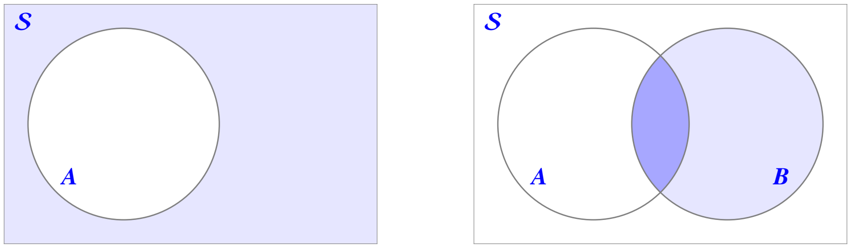
$$(A \cup B)^c = A^c \cap B^c;$$

2. The complement of the intersection equals the union of the complements. In notation,

$$(A \cap B)^c = A^c \cup B^c.$$

To illustrate DeMorgan's laws, let A be the event "the paper has a perfect score on the first question" and B be the event "the paper has a perfect score on the second question" in the quiz review scenario above. Then

Figure 4.2: *Venn diagrams for complement and disjoint union. In part (a), the shaded area represents A^c, the complement of event A. In part (b), the dark shaded area represents the intersection $B \cap A$, the light shaded area represents the intersection $B \cap A^c$ and the union of the two shaded areas represents B. The enclosing rectangular regions represent $\mathcal{S}$.*



(a) *Complement of Event A*

(b) *Event B as the Disjoint Union of $B \cap A$ and $B \cap A^c$*

1. The set of papers that do not have perfect scores on at least one of the first two questions (the set $(A \cup B)^c$) is the same as the set of papers that do not have a perfect score on the first question and do not have a perfect score on the second question (the set $A^c \cap B^c$).
2. The set of papers that do have perfect scores on both the first and second questions (the set $(A \cap B)^c$) is the same as the set of papers with less than perfect scores on the first question or less than perfect scores on the second question (the set $A^c \cup B^c$).

Disjoint or mutually exclusive events. The events A and B are *disjoint* if they have no elements in common, that is, if $A \cap B = \emptyset$. Disjoint events are often said to be *mutually exclusive* since the occurrence of one event prevents the occurrence of the other.

A useful special case is the following: If A and B are events, then the events $B \cap A$ and $B \cap A^c$ are mutually exclusive with union B. That is,

$$B = (B \cap A) \cup (B \cap A^c) \text{ where } (B \cap A) \cap (B \cap A^c) = \emptyset.$$

In this case, we say that we have written B as the *disjoint union* of the part of B that is in A and the part of B that is not in A, as illustrated in Figure 4.2(b) (page 64).

For example, if A is the event "the paper was submitted by a male student" and B is the event "the paper has a perfect score on the first question" in the quiz review scenario above, then B can be written as the disjoint union of the set of papers submitted by male students with perfect scores on the first question (the set $B \cap A$) and the set of papers submitted by female students with perfect scores on the first question (the set $B \cap A^c$).

Pairwise disjoint or pairwise mutually exclusive events. The events A_1, A_2, ..., A_k are said to be *pairwise disjoint* (or *pairwise mutually exclusive*) if

$$A_i \cap A_j = \emptyset \quad \text{whenever} \quad i \neq j.$$

Note that the use of subscripts in this definition is just a convenient way to name the sets in the list of k pairwise disjoint events. Other notations can be used.

For example, suppose there were 4 questions on the quiz. Let A_i be the event

$$\text{“there were perfect scores on exactly } i \text{ questions” for } i = 1, 2, 3, 4.$$

Then the events A_1, A_2, A_3, A_4 are pairwise mutually exclusive since a paper cannot have both exactly i perfect scores and exactly j perfect scores when $i \neq j$. The union of these four events in this example,

$$A_1 \cup A_2 \cup A_3 \cup A_4,$$

is the set of papers with at least one perfect score and the complement of this union,

$$(A_1 \cup A_2 \cup A_3 \cup A_4)^c$$

is the set of papers with no perfect scores.

Partitions or mutually exclusive and exhaustive events. A *partition* of the sample space $\mathcal{S}$ is a collection of pairwise disjoint (or pairwise mutually exclusive) events A_1, A_2, …, A_k whose union is $\mathcal{S}$. That is,

$$\mathcal{S} = A_1 \cup A_2 \cup \cdots \cup A_k \text{ where } A_i \cap A_j = \emptyset \text{ when } i \neq j.$$

Since each element in the sample space is a member of one and only one A_i, the collection of events is often said to be *mutually exclusive and exhaustive* (since all possibilities have been exhausted). Figure 4.3(a) (page 66) illustrates a partition of $\mathcal{S}$ using a collection of 4 mutually exclusive and exhaustive events.

If A_1, A_2, …, A_k is a partition of $\mathcal{S}$ and B is any event, then B can be written as the disjoint union of the intersections $B \cap A_i$ for $i = 1, 2, \ldots, k$. That is,

$$B = (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_k) \text{ where } (B \cap A_i) \cap (B \cap A_j) = \emptyset \text{ when } i \neq j.$$

Figure 4.3(b) illustrates B as the disjoint union of 4 mutually exclusive events.

Example 4.1 (Quiz Score Review) Continuing with the quiz score review scenario, suppose that the quiz had 4 questions. Let M be the event “the paper was submitted by a male student,” F be the event “the paper was submitted by a female student,” and

$$A_i \text{ be the event “there were perfect scores on exactly } i \text{ questions” for } i = 0, 1, 2, 3, 4.$$

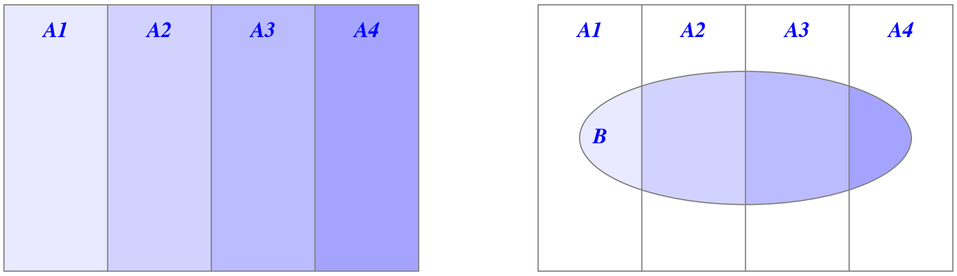
Assume that the following table cross-classifies the 85 student papers by the gender of the student and by the number of problems with perfect scores:

	0 Perfect Scores (A_0)	1 Perfect Score (A_1)	2 Perfect Scores (A_2)	3 Perfect Scores (A_3)	4 Perfect Scores (A_4)	Total
Male Student (M)	2	6	5	14	13	40
Female Student (F)	4	3	10	10	18	45
Total	6	9	15	24	31	85

The row factor corresponds to a partition of the sample space by gender:

$$\mathcal{S} = M \cup F \text{ where } M \cap F = \emptyset.$$

Figure 4.3: *Venn diagrams of partition and disjoint union. In part (a), the shaded areas represent 4 mutually exclusive and exhaustive events. The union of these events is the entire sample space: $\mathcal{S} = A_1 \cup A_2 \cup A_3 \cup A_4$. In part (b), the shaded areas represent 4 mutually exclusive events with union B: $B = (B \cap A_1) \cup (B \cap A_2) \cup (B \cap A_3) \cup (B \cap A_4)$.*



(a) *Partition of $\mathcal{S}$ Using Events A_1, A_2, A_3, A_4*

(b) *Event B as the Disjoint Union of $B \cap A_1$, $B \cap A_2$, $B \cap A_3$, $B \cap A_4$*

Note that there are 40 men and 45 women in the class.

The column factor corresponds to a partition of $\mathcal{S}$ by number of perfect scores:

$$\mathcal{S} = A_0 \cup A_1 \cup A_2 \cup A_3 \cup A_4 \text{ where } A_i \cap A_j = \emptyset \text{ when } i \neq j.$$

Note that 6 papers had no perfect scores, 9 had one perfect score, 15 had two perfect scores, 24 had three perfect scores and 31 had four perfect scores.

The first row corresponds to writing M as the disjoint union of 5 mutually exclusive events,

$$M = (M \cap A_0) \cup \cdots \cup (M \cap A_4) \text{ where } (M \cap A_i) \cap (M \cap A_j) = \emptyset \text{ when } i \neq j,$$

and recording the numbers of papers in each event. The frequencies 2, 6, 5, 14 and 13 are the numbers of papers submitted by male students with 0, 1, 2, 3 and 4 perfect scores, respectively. Similarly, the second row corresponds to writing F as the disjoint union of 5 mutually exclusive events,

$$F = (F \cap A_0) \cup \cdots \cup (F \cap A_4) \text{ where } (F \cap A_i) \cap (F \cap A_j) = \emptyset \text{ when } i \neq j,$$

and recording the numbers of papers in each event. The frequencies 4, 3, 10, 10, 18 are the numbers of papers submitted by female students with 0, 1, 2, 3 and 4 perfect scores, respectively.

Suppose that the professor decides to choose and review one paper (as a check on grading) by implementing the following procedure:

(a) Each student is assigned a unique number between 1 and 85.

(b) The computer is used to randomly select 1 number from the set $\{1, 2, \ldots, 85\}$.

(c) The paper of the student whose number is chosen is reviewed.

Since each choice of paper is equally likely, probabilities can be computed using simple ratios. By working with information that has been cross-classified (using partitions and disjoint unions), probabilities of unions can be computed using sums.

For example,

1. The probability that the chosen paper has 2 or more perfect scores is

   $P(\text{2 or more Perfect Scores}) =$

   $$P(A_2) + P(A_3) + P(A_4) = \frac{15}{85} + \frac{24}{85} + \frac{31}{85} = \frac{70}{85} = 0.8235.$$

2. The probability that the chosen paper was submitted by a female student and has between 1 and 3 perfect scores is

   $P(\text{Female and Between 1 and 3 Perfect Scores}) =$

   $$P(F \cap A_1) + P(F \cap A_2) + P(F \cap A_3) = \frac{3}{85} + \frac{10}{85} + \frac{10}{85} = \frac{23}{85} = 0.2706.$$

4.2.2 Probability Models on Finite Sample Spaces

A *probability model* (or *probability distribution*) on a finite sample space $\mathcal{S}$ is a specification of numbers $P(A)$ ("the probability of event A") satisfying the following rules:

1. ***0-1 Rule:*** The probability of the empty set is 0 and the probability of the full sample space is 1:

   $$P(\emptyset) = 0 \text{ and } P(\mathcal{S}) = 1.$$

2. ***Range Rule:*** The probability of any event is a number between 0 and 1:

   $$0 \leq P(A) \leq 1 \text{ when } A \subseteq \mathcal{S}.$$

3. ***Disjoint Union Rule:*** The probability of the union of mutually exclusive events is the sum of the probabilities of each event:

   $$P(A \cup B) = P(A) + P(B) \text{ when } A \cap B = \emptyset.$$

   More generally,

   $$P(A_1 \cup A_2 \cup \cdots \cup A_k) = P(A_1) + P(A_2) + \cdots + P(A_k)$$

   when events A_1, A_2, ..., A_k satisfy $A_i \cap A_j = \emptyset$ for $i \neq j$.

4. ***Complement Rule:*** The probability of the complement of an event is 1 minus the probability of the event:

   $$P(A^c) = 1 - P(A) \text{ when } A \subseteq \mathcal{S}.$$

5. ***Subset Rule:*** If A is a subset of B, then the probability of A must less than or equal to the probability of B:

$$P(A) \leq P(B) \text{ when } A \subseteq B \subseteq \mathcal{S}.$$

6. ***Union Rule:*** The probability of the union of two events that are not mutually exclusive is the sum of the probabilities of each event minus the probability of their intersection:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \text{ when } A \cap B \neq \emptyset.$$

Probability models are applied in situations where an outcome is always observed. The 0-1 rule tells us that it is impossible to observe no outcome (the probability is zero) and that you are certain to observe some outcome (the probability is one). Since the probability of A is the proportion of time we expect to observe an outcome in A, the range rule tells us that proportions are numbers between 0 and 1.

We have seen the disjoint union rule applied in the quiz score review example beginning on page 65. To illustrate the complement rule using the information in the same example, the probability that the chosen paper has fewer than 2 perfect scores is

$$P(\text{Fewer than 2 Perfect Scores}) =$$

$$1 - P(\text{2 or More Perfect Scores}) = 1 - \tfrac{70}{85} = \tfrac{15}{85} = 0.1765.$$

Note that this probability can also be computed as the sum of two probabilities:

$$P(\text{Fewer than 2 Perfect Scores}) =$$

$$P(\text{0 Perfect Scores}) + P(\text{1 Perfect Score}) = \tfrac{6}{85} + \tfrac{9}{85} = \tfrac{15}{85} = 0.1765.$$

To illustrate the subset rule using the information in the same example, note that

$$P(\text{Male and 2 or More Perfect Scores}) \leq P(\text{2 or More Perfect Scores})$$

since the first event is a subset of the second event. (The probabilities of the two events are 32/85 and 70/85, respectively.)

The following extension of the quiz score review example illustrates the union rule.

Example 4.2 (Quiz Score Review, continued) Let A be the event "the paper has a passing score" and B be the event "the paper was submitted by a male student." Assume that the 85 papers can be cross-classified by gender and passing status as follows:

	Passing Score (A)	Failing Score (A^c)	Total
Male Student (B)	29	11	40
Female Student (B^c)	32	13	45
Total	61	24	85

The probability of $A \cup B$ can be computed as the sum of 3 probabilities:

$P(\text{Male or Has a Passing Score}) =$

$$P(A \cap B) + P(A \cap B^c) + P(A^c \cap B) = \tfrac{29}{85} + \tfrac{32}{85} + \tfrac{11}{85} = \tfrac{72}{85} = 0.8471,$$

or by applying the union rule:

$P(\text{Male or Has a Passing Score}) =$

$$P(A) + P(B) - P(A \cap B) = \tfrac{61}{85} + \tfrac{40}{85} - \tfrac{29}{85} = \tfrac{72}{85} = 0.8471.$$

Footnotes. A probability model is *any* specification of numbers $P(A)$ satisfying the rules above. Researchers often use past experience as a guide to developing models. For example, the following tables show relative frequency distributions of ABO blood types in native Irish and native Japanese populations (from `www.bloodbook.com`):

Irish Population

O	A	B	AB	Total
52%	35%	10%	3%	100%

Japanese Population

O	A	B	AB	Total
30%	38%	22%	10%	100%

Based on these tables, it would be reasonable to assign probabilities 0.52, 0.35, 0.10 and 0.03 to the events that a randomly chosen individual from the Irish population has blood type O, A, B and AB, respectively. And, it would be reasonable to assign the probabilities 0.30, 0.38, 0.22 and 0.10 to the events that a randomly chosen individual from the Japanese population has blood type O, A, B and AB, respectively.

Students often apply the disjoint union rule when they should use the union rule. Whenever possible, you should use partitions and cross-classifications so that you can obtain the probability you want by computing a sum of probabilities of pairwise disjoint events.

4.2.3 Conditional Probability

A *conditional probability* is the probability of an event occurring given that certain conditions hold. More generally, we write the conditional probability that event A occurs given that event B has occurred is $P(A|B)$ (and say "the probability of A given B").

Example 4.3 (ABO Blood Types) (Source: `www.bloodbook.com`) Distributions of blood types differ by ethnic group, so it is natural to consider probabilities of blood types conditional on the random choice of an individual from a given population.

For example, assume that the left table below shows the conditional probabilities of blood types O, A, B and AB given that an individual is randomly selected from the native Irish population, and the right table shows similar probabilities given that an individual is randomly selected from the native Japanese population.

Conditional Distribution of ABO Blood Types among the Irish

O	A	B	AB	Total
0.52	0.35	0.10	0.03	1.00

Conditional Distribution of ABO Blood Types among the Japanese

O	A	B	AB	Total
0.30	0.38	0.22	0.10	1.00

Then, for example, there is a 52% chance that a randomly chosen individual from the native Irish population has blood type O compared to only a 30% chance that a randomly chosen individual from the native Japanese population has blood type O. In notation,

$$P(\text{Type O} \mid \text{Irish}) = 0.52 \text{ compared to } P(\text{Type O} \mid \text{Japanese}) = 0.30.$$

Formula for conditional probability. If $P(B) \neq 0$ and the probabilities $P(A \cap B)$ and $P(B)$ are known, then the probability of A given B can be found by applying the following formula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

That is, the probability that A occurs given that B has occurred is the ratio of the proportion of time we expect to observe an outcome in the intersection $A \cap B$ to the proportion of time we expect to observe an outcome in B.

Example 4.4 (Religion and Education) For example, assume that each individual in an adult community can be classified by religion and highest educational level and that the following two-way table gives the probabilities that a randomly chosen individual from the population falls into each classification and cross-classification:

	NoHS	SomeHS	HSGrad	SomeC	CGrad	AdvDeg	Total
Catholic	0.095	0.125	0.145	0.035	0.032	0.018	0.45
Jewish	0.012	0.035	0.036	0.023	0.027	0.017	0.15
Protestant	0.052	0.088	0.108	0.044	0.076	0.032	0.40
Total	0.159	0.248	0.289	0.102	0.135	0.067	1.00

Religion has three categories (Catholic, Jewish and Protestant), corresponding to three mutually exclusive and exhaustive events. The probability that a randomly chosen individual falls into each category is

$$P(\text{Catholic}) = 0.45,\ P(\text{Jewish}) = 0.15 \text{ and } P(\text{Protestant}) = 0.40,$$

respectively. Highest educational level has six categories (did not attend high school, some high school but not a high school graduate, high school graduate only, some college but not a college graduate, college graduate but no higher degree, earned an advanced degree), corresponding to six mutually exclusive and exhaustive events. The probability that a randomly chosen individual falls into each category is

$$P(\text{NoHS}) = 0.159,\ P(\text{SomeHS}) = 0.248,\ P(\text{HSGrad}) = 0.289,$$

$$P(\text{SomeC}) = 0.102,\ P(\text{CGrad}) = 0.135 \text{ and } P(\text{AdvDeg}) = 0.067,$$

respectively. The joint religion-educational level variable has eighteen categories (one for each choice of religion and highest educational level), corresponding to eighteen mutually exclusive and exhaustive events. The probability that a randomly chosen individual is Catholic and did not attend high school, for example, is

$$P(\text{Catholic and NoHS}) = 0.095.$$

The probability that an individual chosen at random from the subpopulation of Catholics did not attend high school can be computed as follows:

$$P(\text{NoHS} \mid \text{Catholic}) = \frac{P(\text{NoHS and Catholic})}{P(\text{Catholic})} = \tfrac{0.095}{0.45} = 0.2111.$$

(About 21.1% of the Catholic members of the community did not attend high school.) Similar computations can be done to find the probabilities that an individual chosen at random from the Catholic subpopulation falls into each of the remaining five categories of highest educational level. The results are summarized in the following table:

Conditional Distribution of Highest Educational Levels for the Catholic Members of the Adult Community

NoHS	SomeHS	HSGrad	SomeC	CGrad	AdvDeg	Total
0.2111	0.2778	0.3222	0.0778	0.0711	0.0400	1.0000

To compute the conditional distribution of highest educational levels for the Catholic subpopulation, each probability in the first row of the two-way table above is divided by 0.45; the numbers in the first row of the table have been reproportioned to sum to 1.

The conditional distributions of highest educational levels for the Jewish and Protestant subpopulations can be computed by reproportioning the second and third rows, respectively, and are displayed below.

Conditional Distribution of Highest Educational Levels for the Jewish Members of the Adult Community

NoHS	SomeHS	HSGrad	SomeC	CGrad	AdvDeg	Total
0.0800	0.2333	0.2400	0.1533	0.1800	0.1133	1.0000

Conditional Distribution of Highest Educational Levels for the Protestant Members of the Adult Community

NoHS	SomeHS	HSGrad	SomeC	CGrad	AdvDeg	Total
0.1300	0.2200	0.2700	0.1100	0.1900	0.0800	1.0000

Figure 4.4 (page 72) compares the three conditional distributions using polygon plots. Notice that the probability distributions for the Jewish and Protestant subpopulations are roughly similar and that the probability that an individual never attended college is much higher for Catholics living in the community than for Jews or Protestants since

$P(A|\text{Catholic}) = 0.8111$, compared to

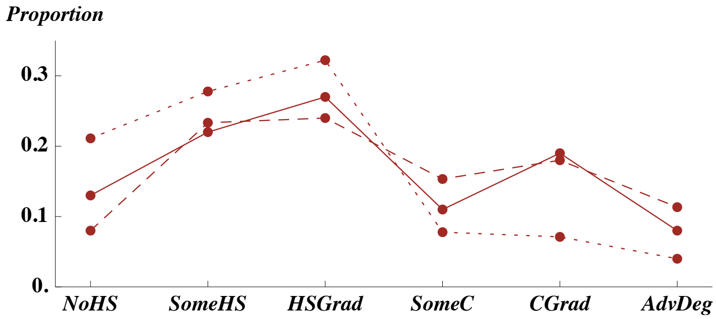
$$P(A|\text{Jewish}) = 0.5533 \text{ and } P(A|\text{Protestant}) = 0.62,$$

where $A = \text{NoHS} \cup \text{SomeHS} \cup \text{HSGrad}$.

Formula for intersection probability. If $P(B) \neq 0$ and the probabilities $P(A|B)$ and $P(B)$ are known, then the probability of the intersection $P(A \cap B)$ can be found by using the following formula:

$$P(A \cap B) = P(B)P(A|B).$$

Figure 4.4: *Comparison of conditional educational level distributions in an adult community. The dotted curve corresponds to the Catholic members of the community, the dashed curve to the Jewish members, and the solid curve to the Protestant members.*



That is, the probability of the intersection of events A and B is the product of the probability of B and the conditional probability of A given B.

Example 4.5 (ABO Blood Types and Eskimos) Distributions of blood types can differ in geographically distinct subpopulations of a given ethnic group.

For example, assume that 55% of Eskimos live in Alaska or Canada, 35% live in Greenland and 10% live in Russia. Further, assume that the following tables give the conditional distributions of ABO blood types in each subpopulation:

Conditional Distributions of ABO Blood Types in Three Subpopulations of Eskimos

	Alaska/ Canada	*Greenland*	*Russia*
Type O:	0.38	0.54	0.33
Type A:	0.44	0.26	0.36
Type B:	0.13	0.12	0.23
Type AB:	0.05	0.08	0.08
Total:	1.00	1.00	1.00

Then, for example, the probability that an individual chosen at random from the entire population of Eskimos lives in Alaska or Canada and has type O blood is

$$P(\text{AorC and Type O}) = P(\text{AorC})\ P(\text{Type O} \mid \text{AorC}) = (0.55)(0.38) = 0.209.$$

(20.9% of Eskimos live in Alaska or Canada and have type O blood.) Similarly, the probability that an individual chosen at random from the entire population of Eskimos lives in Alaska or Canada and has one of the remaining blood types is

$$P(\text{AorC and Type A}) = P(\text{AorC})\ P(\text{Type A} \mid \text{AorC}) = (0.55)(0.44) = 0.242,$$

$$P(\text{AorC and Type B}) = P(\text{AorC})\ P(\text{Type B} \mid \text{AorC}) = (0.55)(0.13) = 0.0715,$$

$$P(\text{AorC and Type AB}) = P(\text{AorC})\ P(\text{Type AB} \mid \text{AorC}) = (0.55)(0.05) = 0.0275,$$

respectively. Note that since event AorC can be written as the disjoint union of 4 events,

$$(\text{AorC and Type O}) \cup (\text{AorC and Type A}) \cup (\text{AorC and Type B}) \cup (\text{AorC and Type AB}),$$

the probability of event AorC must be the sum of the probabilities of these 4 events:

$$P(\text{AorC}) = 0.209 + 0.242 + 0.0715 + 0.0275 = 0.55.$$

The joint blood type-location variable has 12 categories (one for each combination of blood type and subpopulation location). Computations similar to those above can be done to find all 12 intersection probabilities, with results summarized in the following two-way table.

	Alaska/ Canada	*Greenland*	*Russia*	Total
Type O:	0.209	0.189	0.033	0.431
Type A:	0.242	0.091	0.036	0.369
Type B:	0.0715	0.042	0.023	0.1365
Type AB:	0.0275	0.028	0.008	0.0635
Total:	0.55	0.35	0.10	1.00

Note that the right column of the table is the probability distribution of blood types for the entire Eskimo population.

Footnotes. In the religion and education example, the formula for conditional probability was used to find conditional educational level distributions for three subpopulations of an adult community. In the blood types and Eskimos example, the formula for intersection probability was used to combine information initially gathered in each subpopulation to give a complete picture of the entire Eskimo population. We have used the same basic relationship among the probabilities $P(B)$, $P(A|B)$ and $P(A \cap B)$ in two different ways depending on how information was made available to us.

Students often equate $P(A|B)$ and $P(B|A)$, although these two probabilities are quite different both in value and meaning. For example, in the religion and education example,

$$P(\text{NoHS} \mid \text{Catholic}) = \frac{0.095}{0.45} = 0.2111 \text{ and } P(\text{Catholic} \mid \text{NoHS}) = \frac{0.095}{0.159} = 0.5975.$$

From the first calculation, we know that about 21.1% of Catholics living in the community did not go to high school. From the second calculation, we know that about 59.8% of individuals living in the community who did not go to high school are Catholics.

The assumption in the religion and education example was that we knew the exact proportions of individuals in each combination of religion and highest educational level. The assumptions in the blood types and Eskimos example were that we knew the exact proportions of individuals in each blood type category in each of three subpopulations and that we knew the exact proportions of individuals in each subpopulation. These exact proportions are examples of population parameters. In most practical situations, the values of population parameters are unknown and need to be estimated from samples.

4.2.4 Independent Events

Suppose that A and B have nonzero probabilities. Then A and B are said to be *probabilistically independent* (or *independent*) if any one of the following equivalent conditions holds:

1. The conditional probability of A given B is the same as the probability of A,
$$P(A|B) = P(A).$$
2. The conditional probability of B given A is the same as the probability of B,
$$P(B|A) = P(B).$$
3. The probability of the intersection of the two events is equal to the product of the probabilities of the events A, B,
$$P(A \cap B) = P(A)P(B).$$

Independent events are most easily illustrated using simple chance experiments.

Here are two examples:

(1) In a simple chance experiment you decide to toss a fair coin and roll a fair six-sided die. Then the events "a head appears on the top face of the coin" and "4 dots appear on the top face of the die" are independent. Further, the probability that both events occur is

$$P(\text{Head and 4 Dots}) = P(\text{Head})\ P(\text{4 Dots}) = \left(\tfrac{1}{2}\right)\left(\tfrac{1}{6}\right) = \tfrac{1}{12} = 0.0833.$$

(2) In the first step of a simple chance experiment you shuffle a standard deck of 52 cards, choose one card, record its value and return it to the deck. In the second step of the experiment you shuffle the same deck, choose one card, record its value and return it to the deck. Then the events "a red card was chosen in the first step" and "a spade was chosen in the second step" are independent. Further, the probability that both events occur is

$$P(\text{Red}_1 \text{ and Spade}_2) = P(\text{Red}_1)\ P(\text{Spade}_2) = \left(\tfrac{1}{2}\right)\left(\tfrac{1}{4}\right) = \tfrac{1}{8} = 0.125.$$

Positive and negative association. If A and B are not independent, then it is often interesting to know whether the probability of the intersection is greater than or less than what would be expected if the events were independent. That is, it is interesting to know if

$$P(A \cap B) > P(A)P(B) \ \text{ or } \ P(A \cap B) < P(A)P(B).$$

In the first case, we say that A and B are *positively associated*; in the second case, we say that they are *negatively associated*. For example, consider choosing an individual at random from the population of women in New York State who were diagnosed with breast cancer 5 years ago and who were between 55 and 60 years of age at the time of diagnosis. Let A be the event that the woman is alive today and B be the event that her disease was detected using mammography before clinical symptoms arose. A positive association between A and B in this population would provide evidence of the diagnostic value of mammography.

Footnote. Students often confuse mutually exclusive events and independent events. If A and B are mutually exclusive events, then the occurrence of one event prevents the occurrence of the other. If A and B are independent events, then the occurrence of one event has no effect on whether or not the other occurs.

4.2.5 Additional Probability Rules

A probability model must satisfy the six rules stated on page 67. Based on the work of the last two sections, we can add the following four rules:

7. ***Intersection Rule:*** If A and B are events and $P(B) \neq 0$, then the probability of the intersection $A \cap B$ is the probability of B times the probability of A given B:

   $$P(A \cap B) = P(B)P(A|B) \text{ when } P(B) \neq 0.$$

8. ***Intersection Rule for Independent Events:*** If A and B are independent events, then the probability of the intersection $A \cap B$ is the probability of A times the probability of B:

   $$P(A \cap B) = P(A)P(B) \text{ when } A \text{ and } B \text{ are independent.}$$

9. ***Rule of Average Conditional Probabilities:*** Let A_1, A_2, $\ldots$, A_k be a partition of the sample space $\mathcal{S}$, and assume that each $P(A_i) \neq 0$. If B is any event, then $P(B)$ is the weighted average of the conditional probabilities $P(B|A_i)$ with weights $P(A_i)$:

   $$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \cdots + P(A_k)P(B|A_k),$$

   when $P(A_i) \neq 0$ for each i.

10. ***Bayes' Rule for Conditional Probabilities:*** Let A_1, A_2, $\ldots$, A_k be a partition of the sample space $\mathcal{S}$, and assume that each $P(A_i) \neq 0$. If B is any event with $P(B) \neq 0$, then for each i, the conditional probability $P(A_i|B)$ can be found using the following formula:

    $$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1)+P(A_2)P(B|A_2)+\cdots+P(A_k)P(B|A_k)},$$

    when $P(B) \neq 0$ and $P(A_i) \neq 0$ for each i.

Example 4.6 (ABO Blood Types and Eskimos, continued) To illustrate the rule of average conditional probabilities and Bayes' rule, consider again the blood types and Eskimos example beginning on page 72. Let A_1 be the event that an individual lives in Alaska or Canada, A_2 be the event that an individual lives in Greenland, A_3 be the event that an individual lives in Russia, and B be the event that an individual has type O blood.

B can be written as the disjoint union of $B \cap A_1$, $B \cap A_2$ and $B \cap A_3$. Thus, by the disjoint union rule,

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_3).$$

Since each $P(B \cap A_i) = P(A_i)P(B|A_i)$ by the intersection rule, we can compute the probability of event B as follows:

$$\begin{aligned} P(B) = &\ P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3) \\ = &\ (0.55)(0.38) + (0.35)(0.54) + (0.10)(0.33) \\ = &\ 0.209 + 0.189 + 0.033 = 0.431. \end{aligned}$$

Thus, $P(B) = 0.431$ is the weighted average of the conditional probabilities of blood type O in the three subpopulations (0.38, 0.54 and 0.33), where the weights correspond to the proportions of individuals living in each subpopulation (0.55, 0.35 and 0.10).

The table below summarizes the computations needed to find $P(B)$ using the rule of average conditional probabilities and includes a column of the conditional probabilities $P(A_i|B)$ computed using Bayes' rule.

				$P(A_i \cap B) =$	$P(A_i\|B) =$
Event	$P(A_i)$	$P(B\|A_i)$		$P(A_i)P(B\|A_i)$	$P(A_i \cap B)/P(B)$
A_1	0.55	0.38		0.209	0.4849
A_2	0.35	0.54		0.189	0.4385
A_3	0.10	0.33		0.033	0.0766
			$P(B) =$	0.431	

In fact, once $P(A_i \cap B)$ and $P(B)$ are computed, Bayes' rule is just a restatement of the formula for finding the probability of A_i given B.

The last column of the table can be interpreted as follows: If an individual is chosen at random from those with blood type O, then there is a 48.49% chance that the person lives in Alaska or Canada, a 43.85% chance the person lives in Greenland and a 7.66% chance the person lives in Russia.

Example 4.7 (Employment Status and Hearing Impairment) (Pagano & Gauvreau, 2000, page 131) As part of the National Health Interview Survey, information was collected on employment status and hearing impairment due to injury for more than 160,000 individuals 17 years of age or older. These individuals form the study population for this example.

Let E_1 be the event that an individual was classified as currently employed, E_2 be the event that the individual was classified as currently unemployed, E_3 the event that the individual was classified as not in the labor force, and H be the event the individual has hearing impairment due to injury.

Assume that 60.63% of the study population was classified as currently employed, 4.57% as currently unemployed and 34.8% as not in the labor force. If an individual is chosen at random from the study population, then the probabilities are

$$P(E_1) = 0.6063,\ P(E_2) = 0.0457 \text{ and } P(E_3) = 0.3480$$

that the individual was classified as currently employed, currently unemployed or not in the labor force, respectively.

Further, assume that 0.56% of individuals classified as currently employed have hearing impairment due to injury, 0.36% of individuals classified as currently unemployed have hearing impairment due to injury and 0.65% of individuals classified as not in the labor force have hearing impairment due to injury. Then the conditional probabilities

$$P(H|E_1) = 0.0056,\ P(H|E_2) = 0.0036 \text{ and } P(H|E_3) = 0.0065,$$

are the probabilities that an individual chosen at random from a given subpopulation (currently employed, currently unemployed, not in the labor force) has hearing impairment due to injury.

The following table summarizes the computations needed to find $P(H)$ as the weighted average of the $P(H|E_i)$'s and to find the conditional employment status distribution given that an individual has hearing impairment.

Event	$P(E_i)$	$P(H\|E_i)$		$P(E_i \cap H) =$ $P(E_i)P(H\|E_i)$	$P(E_i\|H) =$ $P(E_i \cap H)/P(H)$
E_1	0.6063	0.0056		0.0034	0.5763
E_2	0.0457	0.0036		0.0002	0.0339
E_3	0.3480	0.0065		0.0023	0.3898
			$P(H) =$	0.0059	

About 0.59% of individuals in the study population were classified as hearing impaired due to injury. Among those so classified, 57.63% were classified as currently employed, 3.39% as currently unemployed and 38.98% as not in the labor force.

Notice, in particular, that the proportion of individuals not in the labor force is larger in the subpopulation of hearing impaired due to injury ($P(E_3|H) = 0.3898$) than in the study population as a whole ($P(E_3) = 0.3480$).

Footnotes. The intersection rule for independent events is a special case of the general intersection rule that only applies when $P(A) = P(A|B)$. Students often apply the intersection rule for independent events when they should use the general rule. Be sure to check assumptions before using each rule.

The rule of average conditional probabilities allows us to take knowledge of the chance of observing B within each subpopulation and, by averaging, to find the chance of observing B in the entire population. Bayes' rule takes that computation further, by allowing us to focus on probabilities given knowledge that B has occurred. An important application of these ideas is to the study of diagnostic tests, the subject of the next section.

4.3 Analysis of Diagnostic Tests

Much of the information gathered by taking a patient history, doing a physical exam, and analyzing laboratory and other tests, help medical practitioners diagnose a patient's condition. As each test result is observed, the physician adjusts the relative likelihood that the patient has one of several diseases under consideration.

In diagnostic screening, a test is applied to an individual who has not yet exhibited clinical symptoms of a particular disease with the goal of determining that individual's probability of having the disease. Those who test positive are considered more likely to have the disease and may be recommended to take further tests or to start treatment. Diagnostic screening is most often employed in situations where early detection would contribute to a more favorable prognosis for the individual or for the population as a whole. Examples of diagnostic screening tests are Pap smears for cervical cancer and mammograms for breast cancer.

4.3.1 Sensitivity and Specificity

A diagnostic test for a particular disease is said to be *sensitive* if all individuals who test positive for the disease actually have the disease; it is said to be *specific* if all individuals who test negative for the disease are actually disease free. If a diagnostic test is both sensitive and specific, then each individual would be correctly classified as having the disease or not.

Unfortunately, diagnostic tests are rarely perfectly sensitive and specific. Let D be the event that an individual has the disease, D^c be the event that the individual is disease free, POS be the event the individual tests positive for disease and NEG be the event the individual tests negative for disease.

The *sensitivity* of a diagnostic test is the probability that an individual chosen at random from those with disease tests positive for disease,

$$\text{Sensitivity} = P(POS|D).$$

The *specificity* of a diagnostic test is the probability that an individual chosen at random from those who are disease free tests negative for disease,

$$\text{Specificity} = P(NEG|D^c).$$

True-positive and false-negative rates. The sensitivity is also called the *true-positive rate*. If a diagnostic test is perfectly sensitive, then the true-positive rate is 1. Otherwise, the true-positive test is less than 1 and there is the possibility that a person with disease will test negative for disease. The probability that a person with disease tests negative is known as the *false-negative rate*,

$$\text{False-Negative Rate} = 1 - P(POS|D) = P(NEG|D).$$

True-negative and false-positive rates. The specificity is also called the *true-negative rate*. If a diagnostic test is perfectly specific, then the true-negative rate is 1. Otherwise, the true-negative rate is less than 1 and there is the possibility that a person who is disease free will test positive for disease. The probability that a person who is disease free tests positive is known as the *false-positive rate*,

$$\text{False-Positive Rate} = 1 - P(NEG|D^c) = P(POS|D^c).$$

Example 4.8 (Triple Blood Test and Down Syndrome) (Agresti & Franklin, 2007, page 216) Down syndrome is a genetic disorder arising from an error in cell division that results in a fetus having an extra copy of chromosome 21. The chance of giving birth to a child with Down syndrome increases as the mother's age increases.

The Triple Blood Test is used to screen a pregnant woman for this disorder. (A positive test would suggest that her baby has Down syndrome; a negative test would suggest that her baby does not have Down syndrome.) In a study of the effectiveness of the Triple Blood

Test (reported in the *New England Journal of Medicine* (1994) 330:1114-8), more than 5000 pregnant women 35 years of age or older were given the test prior to giving birth. These women form the study population for this example.

The following two-way table gives the probabilities that a randomly chosen woman from the study population falls into each cross-classification of disorder status (Down syndrome present or absent in her child) by blood test result (positive or negative):

	Blood Test Positive (POS)	Blood Test Negative (NEG)	Total
Down Syndrome Present (D)	0.0091	0.0011	0.0102
Down Syndrome Absent (D^c)	0.2475	0.7423	0.9898
Total	0.2566	0.7434	1.0000

In order to determine sensitivity and specificity, we need to consider the conditional blood test distributions in women whose babies did or did not have Down syndrome:

Down Syndrome Present:

POS	NEG	Total
0.8922	0.1078	1.0000

Down Syndrome Absent:

POS	NEG	Total
0.2501	0.7499	1.0000

In this study population, the sensitivity of the test is $P(POS|D) = 0.8922$, the specificity is $P(NEG|D^c) = 0.7499$, the false-negative rate is $P(NEG|D) = 0.1078$ and the false-positive rate is $P(POS|D^c) = 0.2501$.

Finally, note that the probability the test correctly predicted the outcome is

$$P(\text{Correct Prediction}) = P(POS \cap D) + POS(NEG \cap D^c) = 0.0091 + 0.7423 = 0.7514.$$

(The test was correct only 75.14% of the time.) The fairly low probability of a correct prediction is due, in part, to the fairly low specificity of the test (equivalently, to the fairly high false-positive rate).

4.3.2 Prevalence and Predictive Value

Let D be the event that an individual has the disease of interest. The *prevalence* of the disease is the probability that an individual chosen at random from the study population has the disease,

$$\text{Prevalence} = P(D).$$

For example, in the Triple Blood Test and Down syndrome example above, the prevalence of Down Syndrome was $P(D) = 0.0102$.

Predictive value of a positive test. The *predictive value of a positive test* (or the *positive predictive value of the test*) is the probability that an individual chosen at random from those testing positive for disease actually has the disease,

$$\text{Positive Predictive Value} = P(D|POS).$$

For example, in the Triple Blood Test and Down syndrome example, the positive predictive value is

$$P(D|POS) = \frac{P(D \cap POS)}{P(POS)} = \frac{0.0091}{0.2566} = 0.0355;$$

only 3.55% of women who tested positive gave birth to a child with Down syndrome.

Predictive value of a negative test. The *predictive value of a negative test* (or the *negative predictive value of the test*) is the probability that an individual chosen at random from those testing negative for disease is actually disease free,

$$\text{Negative Predictive Value} = P(D^c|NEG).$$

For example, in the Triple Blood Test and Down syndrome example, the negative predictive value is

$$P(D^c|NEG) = \frac{P(D^c \cap NEG)}{P(NEG)} = \frac{0.7423}{0.7434} = 0.9985;$$

99.85% of women who tested negative gave birth to a child who was disease free and only 0.15% gave birth to a child with Down syndrome ($P(D|NEG) = 0.0015$).

Footnote. In the example above, the Triple Blood Test for Down syndrome had very low positive predictive value but very high negative predictive value. In the study population, a woman has a 1.02% chance of giving birth to a child with Down syndrome. But, if she tests negative for Down syndrome, then she has only an 0.15% chance of giving birth to a child with Down syndrome. The chance of giving birth to a child with Down syndrome is reduced from a bit more than one in a hundred to a bit more than one in a thousand.

4.3.3 Use of Bayes' Rule

Recall that the sensitivity of a diagnostic test is the probability that an individual chosen at random from those with disease tests positive for disease and the specificity is the probability that an individual chosen at random from those who are disease free tests negative for disease,

$$\text{Sensitivity} = P(POS|D) \text{ and Specificity} = P(NEG|D^c).$$

Sensitivity and specificity do not depend on the prevalence of the disease in the population, $P(D)$. By contrast, the positive and negative predictive values of a test depend on the characteristics of the test (the sensitivity and specificity) and the prevalence of the disease in the population.

In many clinical situations, the sensitivity of a diagnostic test is established by applying the test to individuals who are known to have the disease and the specificity is established by applying the test to individuals who are known to be disease free. The prevalence of the disease is established in a separate study. The rule of average conditional probabilities and Bayes' rule are then applied to find the positive and negative predictive values of the test.

Example 4.9 (Carpal Tunnel Syndrome) (Pagano & Gauvreau, 2000, page 157) Carpal tunnel syndrome is an affliction of the wrist, which occurs when the median nerve (a nerve running from the forearm into the hand) becomes pressed or squeezed. The risk of developing carpal tunnel syndrome is not confined to individuals in a single industry, but is especially common in those performing assembly line work.

The National Institute for Occupational Safety and Health (NIOSH) has established a definition of this disorder that incorporates three criteria: symptoms of nerve damage, a history of occupational risk factors, and the presence of physical exam findings. The sensitivity of this definition as a test for the syndrome is 0.67, and the specificity is 0.58. That is,

$$P(POS|D) = 0.67 \text{ and } P(NEG|D^c) = 0.58,$$

where D is the event that a worker has carpal tunnel syndrome, D^c is the event that a worker does not have carpal tunnel syndrome, POS is the event that the worker satisfies all three criteria established by NIOSH and NEG is the event that the worker does not satisfy all three criteria.

In an industry where 20% of workers have carpal tunnel syndrome, the probability that an individual chosen at random from the population tests positive is

$$\begin{aligned} P(POS) = & \; P(POS \cap D) + P(POS \cap D^c) \\ = & \; P(D)P(POS|D) + P(D^c)P(POS|D^c) \\ = & \; (0.20)(0.67) + (0.80)(1 - 0.58) \\ = & \; 0.134 + 0.336 = 0.47, \end{aligned}$$

and the predictive value of a positive test is $P(D|POS) = 0.134/0.47 = 0.2851$.

In the same population of workers, the probability of a negative test is

$$\begin{aligned} P(NEG) = & \; P(NEG \cap D) + P(NEG \cap D^c) \\ = & \; P(D)P(NEG|D) + P(D^c)P(NEG|D^c) \\ = & \; (0.20)(1 - 0.67) + (0.80)(0.58) \\ = & \; 0.066 + 0.464 = 0.53 \end{aligned}$$

(or, simply, $P(NEG) = 1 - P(POS) = 1 - 0.47 = 0.53$), and the predictive value of a negative test is $P(D^c|NEG) = 0.464/0.53 = 0.8755$.

These results are summarized below:

Analysis of Carpal Tunnel Syndrome Test
When 20% of Workers Have the Disease

	POS	NEG	Total
D	0.134	0.066	0.20
D^c	0.336	0.464	0.80
Total	0.470	0.530	1.00

Predictive Value of
Positive Test is 0.2851.

Predictive Value of
Negative Test is 0.8755.

By contrast, in an industry where 5% of workers have carpal tunnel syndrome, computations similar to the ones above lead to the following summary table:

Analysis of Carpal Tunnel Syndrome Test
When 5% of Workers Have the Disease

	POS	NEG	Total
D	0.0335	0.0165	0.05
D^c	0.3990	0.5510	0.95
Total	0.4325	0.5675	1.00

Predictive Value of
Positive Test is 0.0775.

Predictive Value of
Negative Test is 0.9709.

Notice that the positive predictive value is much smaller when only 5% of workers have carpal tunnel syndrome, and the negative predictive value is much larger.

Footnote. The complement rule was used several times in the analyses of the carpal tunnel syndrome test. POS and NEG are complementary events in the population of all workers in a given industry, and in the subpopulations of workers with disease and of workers who are disease free. Thus, $P(NEG) = 1 - P(POS)$, $P(NEG|D) = 1 - P(POS|D)$ and $P(NEG|D^c) = 1 - P(POS|D^c)$. Be careful that you don't write something like

$$\text{“}P(POS|D^c) = 1 - P(POS|D)\text{” or “}P(NEG|D^c) = 1 - P(NEG|D)\text{”},$$

which are false in general.

4.4 Discrete Random Variables

A *random variable* is a variable whose value is a numerical outcome of a random phenomenon. We usually use capital letters (X, Y, Z, $\ldots$) to denote random variables and lower case letters (x, y, z, $\ldots$) to denote values of random variables.

The random variable is said to be *discrete* if its values form a finite set of numbers or a sequence of numbers such as 0, 1, 2, $\ldots$. For example, suppose that in a simple chance experiment you decide to toss a coin 5 times and let X be the total number of times a tail appeared on the top face. Then X is a discrete random variable whose values form the finite set $\{0, 1, 2, 3, 4, 5\}$.

4.4.1 Probability Distributions

If X is a discrete random variable, then the *probability distribution* of X specifies all possible outcomes of the random variable and gives the probability that each will occur. We use the notation $P(X = x)$ to denote the probability that outcome x occurs.

Example 4.10 (Number of Tails) Suppose that in a simple chance experiment you decide to toss a coin 5 times and let

X equal the total number of times a tail appeared on the top face.

If the coin is fair (that is, if a head appears 50% of the time and a tail appears 50% of the time), then the table on the left below shows the probability distribution of X:

x	$P(X = x)$
0	1/32
1	5/32
2	10/32
3	10/32
4	5/32
5	1/32
	32/32

Sequences with exactly x tails
hhhhh
hhhht, hhhth, hhthh, hthhh, thhhh
hhhtt, hhtht, hthht, thhht, hhtth, ththh, thhth, htthh, ththh, tthhh
ttthh, tthth, thtth, httth, tthht, ththt, httht, thhtt, hthtt, hhttt
htttt, thttt, tthtt, tttht, tttth
ttttt

The table on the right shows the work needed to construct the probability distribution in this case. There are a total of 32 sequences of 5 h's and t's: one sequence has 0 tails, five have 1 tail, ten have 2 tails, ten have 3 tails, five have 4 tails and one has 5 tails. Since the coin is assumed to be fair, each sequence is equally likely.

Note that the events "$X = x$" correspond to a partition of the sample space of all possible sequences of 5 h's and t's. Thus, the sum of the probabilities must be 1.

4.4.2 Probability Histograms

A *probability histogram* is a graphical display of the probability distribution of a discrete random variable with whole number values, where area is used to represent probability.

A probability histogram is constructed as follows:

1. For a given value x, a rectangle with base $[x-0.5, x+0.5]$ and height $P(X = x)$ is drawn. Since the width of the base is 1, the area equals $P(X = x)$.
2. The process is repeated for all values of X. The sum of the areas must be 1.

Example 4.11 (Numbers in Households) Distributions of numbers of individuals living in households differ, in general, by proximity to large cities.

For example, let X be the number of individuals living in a household chosen at random from households in the downtown area of a large midwestern city, Y be the number of individuals living in a household chosen at random from households in one of the adjacent suburban areas, and assume that the following tables represent probability distributions for the random variables X and Y:

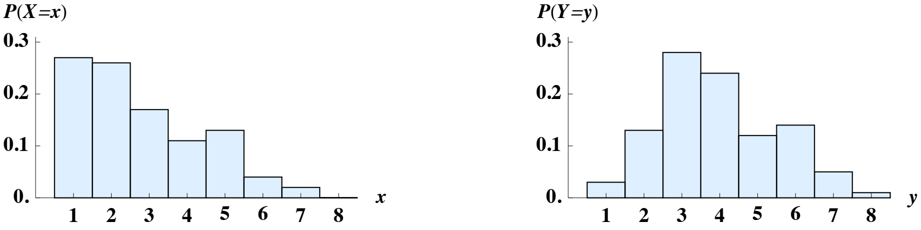
Probability Distribution for
Number of Individuals in Downtown Households

x	1	2	3	4	5	6	7	8	Total
$P(X = x)$	0.27	0.26	0.17	0.11	0.13	0.04	0.02	0.00	1.00

Probability Distribution for
Number of Individuals in Suburban Households

y	1	2	3	4	5	6	7	8	Total
$P(Y = y)$	0.03	0.13	0.28	0.24	0.12	0.14	0.05	0.01	1.00

Figure 4.5: *Probability histograms for numbers in households example.*



(a) *Downtown Households*

(b) *Suburban Households*

Figure 4.5 (page 84) shows probability histograms for these two distributions. Notice, in particular, that the most likely number of individuals living in a household in the downtown area is 1, compared to 3 in the suburban area.

4.4.3 Expected Value, Mean, Variance and Standard Deviation

An *expected value* is a weighted average, where the probability distribution of a discrete random variable serves as the weights in the weighted average.

If X is a discrete random variable, then

1. ***Mean:*** The *mean* of X is the weighted average of the values that X assumes:

$$E(X) = \sum_x x\ P(X=x), \text{ where the sum is over all outcomes } x.$$

   We often use the Greek letter μ ("mu") for the mean of X.

2. ***Variance:*** If $\mu = E(X)$ is the mean of X, then the *variance* of X is the weighted average of the squared deviations from the mean:

$$Var(X) = \sum_x (x-\mu)^2\ P(X=x), \text{ where the sum is over all outcomes } x.$$

3. ***Standard Deviation:*** The *standard deviation* of X is the positive square root of the variance:

$$SD(X) = \sqrt{Var(X)}, \text{ where } Var(X) \text{ is the variance of } X.$$

   We often use the Greek letter σ ("sigma") for the standard deviation of X.

The standard deviation of X can be thought of as an average distance from the mean.

Example 4.12 (Numbers in Households, continued) Consider again the numbers of individuals living in households in downtown and suburban areas.

The following tables show the work needed to compute the mean, variance and standard deviation of numbers for downtown households:

$x\ P(X=x)$	$(x-\mu)^2\ P(X=x)$
$(1)(0.27)=0.27$	$(3.1329)(0.27)=0.8459$
$(2)(0.26)=0.52$	$(0.5929)(0.26)=0.1542$
$(3)(0.17)=0.51$	$(0.0529)(0.17)=0.0090$
$(4)(0.11)=0.44$	$(1.5129)(0.11)=0.1664$
$(5)(0.13)=0.65$	$(4.9729)(0.13)=0.6465$
$(6)(0.04)=0.24$	$(10.4329)(0.04)=0.4173$
$(7)(0.02)=0.14$	$(17.8929)(0.02)=0.3579$
2.77	2.5972

From these computations,

$$E(X)=2.77,\ Var(X)=2.5972 \text{ and } SD(X)=\sqrt{2.5972}=1.62.$$

There are, on average, 2.77 individuals per household in the downtown area, with a standard deviation of 1.62 individuals.

Similarly, the following tables show the work needed to compute the mean, variance and standard deviation of numbers for suburban households:

$y\ P(Y=y)$	$(y-\mu)^2\ P(Y=y)$
$(1)(0.03)=0.03$	$(8.7616)(0.03)=0.0940$
$(2)(0.13)=0.26$	$(3.8416)(0.13)=0.0771$
$(3)(0.28)=0.84$	$(0.9216)(0.28)=0.0148$
$(4)(0.24)=0.96$	$(0.0016)(0.24)=0.3631$
$(5)(0.12)=0.60$	$(1.0816)(0.12)=0.5967$
$(6)(0.14)=0.84$	$(4.1616)(0.14)=1.4606$
$(7)(0.05)=0.35$	$(9.2416)(0.05)=0.8946$
$(8)(0.01)=0.08$	$(16.3216)(0.01)=0.2735$
3.96	2.3583

From these computations,

$$E(Y)=3.96,\ Var(Y)=2.3583 \text{ and } SD(Y)=\sqrt{2.3583}=1.54.$$

There are, on average, 3.96 individuals per household in the suburban area, with a standard deviation of 1.54 individuals.

Chebyshev's inequality. *Chebyshev's inequality* is a general theorem relating the mean and standard deviation of a random variable to probability.

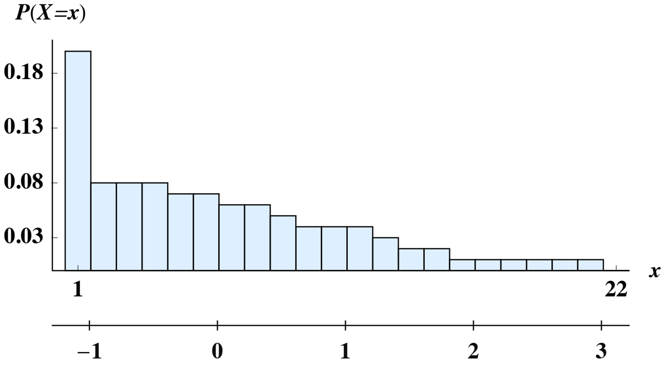
Specifically, let X be a discrete random variable with mean $\mu=E(X)$ and standard deviation $\sigma=SD(X)$ and let k be a constant greater than 1. Then the interval $[\mu-k\sigma,\mu+k\sigma]$ is guaranteed to contain at least $100(1-1/k^2)\%$ of the probability distribution of X,

$$P(\mu-k\sigma\leq X\leq\mu+k\sigma)\geq 1-\frac{1}{k^2}, \text{ when } k>1.$$

When $k=2$, for example, we can expect at least 75% of the distribution within 2 standard deviations of the mean (since $1-\frac{1}{2^2}=0.75$); when $k=2.5$, we can expect at least 84% of the distribution within 2.5 standard deviations of the mean (since $1-\frac{1}{2.5^2}=0.84$).

Example 4.13 (Life Expectancy on Dialysis) (Weinstein et al, 1980, page 224) Let X be the number of years a patient starting kidney dialysis treatments today survives and, for

Figure 4.6: *Probability distribution of the number of years surviving from the start of kidney dialysis treatment. The horizontal line below the probability histogram shows the standardized scores $z = (x - \mu)/\sigma$.*



purposes of illustration, assume that X takes whole number values between 1 and 21 with the following probability distribution:

Probability Distribution for
Number of Years a Kidney Dialysis Patient Survives After Starting Treatment

x	1	2	3	4	5	6	7	8	9	10	11
$P(X = x)$	0.20	0.08	0.08	0.08	0.07	0.07	0.06	0.06	0.05	0.04	0.04
x	12	13	14	15	16	17	18	19	20	21	22+
$P(X = x)$	0.04	0.03	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.00

For this random variable, the mean is

$$E(X) = 1(0.20) + 2(0.08) + \cdots + 20(0.01) + 21(0.01) = 6.44$$

and the standard deviation is

$$SD(X) = \sqrt{(1 - 6.44)^2(0.20) + \cdots + (21 - 6.44)^2(0.01)} = 4.99.$$

The life expectancy for a patient starting kidney dialysis treatment today is 6.44 years with a standard deviation of 4.99 years.

Figure 4.6 (page 86) shows the probability histogram for X. For convenience, a separate horizontal axis is drawn below the histogram marking locations of *standardized scores*,

$$z = (x - \mu)/\sigma, \text{ where } \mu = 6.44 \text{ and } \sigma = 4.99.$$

To illustrate Chebyshev's inequality, let $k = 2$ and consider the probability that X lies within 2 standard deviations of its mean (equivalently, between -2 and $+2$ on the standardized scores axis). Since $\mu - 2\sigma = -3.54$ and $\mu + 2\sigma = 16.42$,

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = P(X = 1) + P(X = 2) + \cdots + P(X = 16) = 0.95.$$

The exact probability is greater than the guaranteed lower bound of 0.75.

If we let $k = 2.5$, then $\mu - 2.5\sigma = -6.035$, $\mu + 2.5\sigma = 18.915$ and

$$P(\mu - 2.5\sigma \le X \le \mu + 2.5\sigma) = P(X = 1) + P(X = 2) + \cdots + P(X = 18) = 0.97,$$

which is greater than the guaranteed lower bound of 0.84.

Footnotes. The mean (μ), variance (σ^2) and standard deviation (σ) defined in this section are examples of population parameters. These population parameters can be computed if the probability distribution of the random variable is known exactly.

By contrast, the mean ($\overline{x}$), variance (s^2) and standard deviation (s) defined in Section 3.5.2 (beginning on page 39) are examples of estimates of population parameters (or statistics) based on samples taken from a population. In many practical situations, values of population parameters are unknown and need to be estimated from samples.

In the life expectancy on dialysis example, a *discrete* random variable was used for purposes of illustration only. In real applications, the number of years a patient survives is a *continuous* random variable since its values form an interval on the real line. Since analysis of continuous random variables (including finding probabilities, means, variances and standard deviations) requires knowledge of methods from calculus, a discrete approximation was used instead.

4.4.4 Joint and Conditional Distributions

A *joint distribution* is the probability distribution of two or more random variables. Two-way tables are often used to display the joint distribution of two variables.

Example 4.14 (Happiness in Marriage) (Olkin et al, 1994, page 544) In order to determine the extent to which agreement on dealing with in-laws could be used to predict happiness in a marriage, researchers asked more than 500 married couples to rate their agreement on dealing with in-laws using a six-point scale of whole numbers between 0 (lowest agreement) and 5 (highest agreement) and to rate the happiness of their marriage using a five-point scale of whole numbers between -2 (very unhappy) and 2 (very happy). These couples become the study population for this example.

Let X be the agreement score and Y be the happiness score of a couple chosen at random from the study population. The following two-way table shows the joint agreement-happiness distribution:

Joint Probability Distribution for Agreement-Happiness Scores

	$y = -2$	$y = -1$	$y = 0$	$y = 1$	$y = 2$	Total
$x = 0$	0.028	0.029	0.009	0.007	0.002	0.075
$x = 1$	0.009	0.022	0.009	0.002	0.002	0.044
$x = 2$	0.018	0.024	0.022	0.015	0.015	0.094
$x = 3$	0.011	0.028	0.033	0.029	0.042	0.143
$x = 4$	0.011	0.015	0.042	0.073	0.103	0.244
$x = 5$	0.015	0.011	0.046	0.077	0.251	0.400
Total	0.092	0.129	0.161	0.203	0.415	1.000

Notice, in particular, that 2.8% of couples in the study population gave the lowest scores on both questions and 25.1% gave the highest scores on both questions,

$$P(X = 0 \text{ and } Y = -2) = 0.028 \quad \text{and} \quad P(X = 5 \text{ and } Y = 2) = 0.251.$$

The right column of the table gives the probability distribution of X. For this random variable, the mean is $E(X) = 3.637$ and the standard deviation is $SD(X) = 1.544$.

The bottom row of the table gives the probability distribution of Y. For this random variable, the mean is $E(Y) = 0.72$ and the standard deviation is $SD(Y) = 1.357$.

Conditional distributions. A *conditional probability distribution* (or *conditional distribution*) is the distribution of one variable given a fixed value of another variable.

For example, if $P(X = x) \neq 0$, then to construct the *conditional distribution of Y given $X = x$* we would compute the conditional probability

$$P(Y = y|X = x) = \frac{P(Y = y \text{ and } X = x)}{P(X = x)} \text{ for each outcome } y.$$

Example 4.15 (Happiness in Marriage, continued) Continuing with the happiness in marriage example, the conditional distribution of Y given $X = 0$ is

Conditional Distribution of Y given X=0

$y = -2$	$y = -1$	$y = 0$	$y = 1$	$y = 2$	Total
0.3733	0.3867	0.1200	0.0933	0.0267	1.0000

Figure 4.7(a) (page 89) is a probability histogram of this conditional distribution.

The conditional distribution of Y given $X = 0$ is obtained by dividing each probability in the first row of the two-way table above by $P(X = 0) = 0.075$. Similarly, the conditional distribution of Y given $X = 1$ is obtained by dividing each probability in the second row of the two-way table by $P(X = 1) = 0.044$.

Conditional Distribution of Y given X=1

$y = -2$	$y = -1$	$y = 0$	$y = 1$	$y = 2$	Total
0.2045	0.5000	0.2045	0.0455	0.0455	1.0000

Figure 4.7(b) is a probability histogram of this conditional distribution.

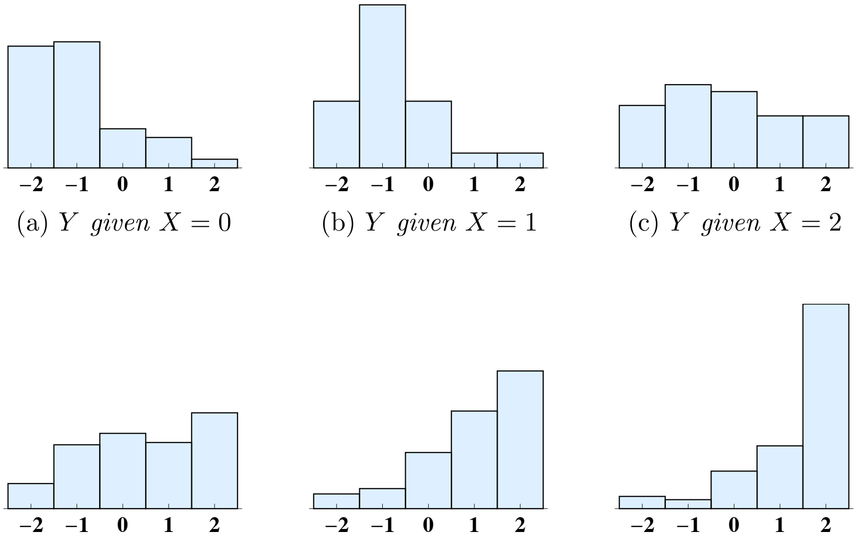
Figures 4.7(c) through (f) are probability histograms for conditional distributions given $X = 2$ through $X = 5$. Notice that as x increases, the conditional distribution of Y given $X = x$ becomes more concentrated near $Y = 2$.

Conditional expected value. A *conditional expected value* is an expected value computed using a conditional probability distribution.

For example, the *conditional mean of Y given $X = x$* is

$$E(Y|X = x) = \sum_y y \; P(Y|X = x), \text{ where the sum is over all outcomes } y.$$

Figure 4.7: *Probability histograms of conditional happiness score distributions.*



(a) Y *given* $X = 0$ (b) Y *given* $X = 1$ (c) Y *given* $X = 2$

(d) Y *given* $X = 3$ (e) Y *given* $X = 4$ (f) Y *given* $X = 5$

Example 4.16 (Happiness in Marriage, continued) Continuing with the happiness in marriage example, the conditional mean of Y given $X = 0$ is

$$E(Y|X=0) =$$

$$(-2)(0.3733) + (-1)(0.3867) + (0)(0.1200) + (1)(0.0933) + (2)(0.0267) = -0.9867$$

and the conditional mean of Y given $X = 1$ is

$$E(Y|X=1) =$$

$$(-2)(0.2045) + (-1)(0.5) + (0)(0.2045) + (1)(0.0455) + (2)(0.0455) = -0.7727.$$

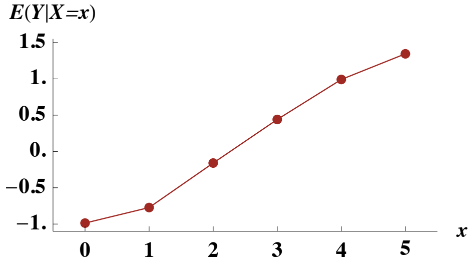
Using similar computations for the remaining agreement scores we get

x	2	3	4	5
$E(Y\|X=x)$	−0.1596	0.4406	0.9918	1.345

A plot of pairs $(x, E(Y|X=x))$, for $x = 0, 1, 2, 3, 4, 5$, is given in Figure 4.8 (page 90). Notice that as the agreement score increases, the average happiness score increases as well.

Footnotes. In the happiness in marriage example above, Y is the response variable and X is the explanatory variable. These variables are positively associated in the study population. To see this we compared conditional means $E(Y|X=x)$ for each x.

Figure 4.8: *Plot of expected happiness score $E(Y|X=x)$ versus agreement score x.*



If X is the response variable and Y is the explanatory variable, then it would be natural to compare the conditional means $E(X|Y=y)$ for each y. When we compute the conditional means $E(Y|X=x)$ for each x, we say that we are *regressing* Y *on* X. When we compute the conditional means $E(X|Y=y)$ for each y, we say that we are *regressing* X *on* Y.

4.5 Brief Summary and Additional Examples

This chapter introduces probability models, rules for working with probability models, methods for analyzing diagnostic tests and methods for working with experiments whose outcomes are coded as numbers.

Models are approximations. A probability model is *any* specification of numbers $P(A)$ to events of interest ($A \subseteq \mathcal{S}$) satisfying the rules for probability models. In general, researchers use past experience as a guide to developing models. Thus, probability models are approximations of reality only. But, if well-constructed, probability models can be very useful. A famous saying, attributed to many different statisticians, is

“All models are wrong, but some are useful.”

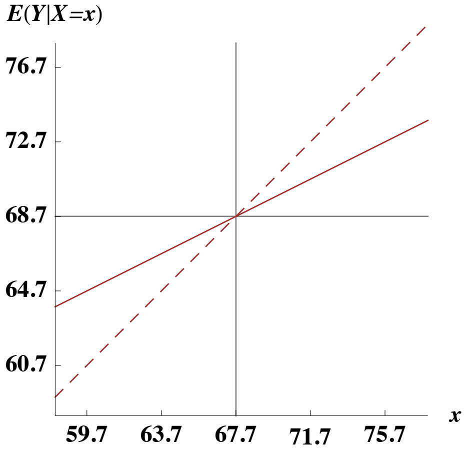
Sample used as study population. In several examples in this chapter — including the employment status and hearing impairment example (beginning on page 76) and the happiness in marriage example (beginning on page 87) — the sample gathered by the researchers was used as the study population. Since the study population was known exactly, we could produce *exact* probabilities based on the rules. When considering the information as sample information only (drawn from larger study populations), the results would no longer be exact. Instead, the results would be *estimated* probabilities.

Regression of Y on X. Many people associate the word “regression” with the process of estimating the slope and intercept of a least squares regression line (Section 3.7.3, page 48) and conducting statistical analyses of the results. In fact, the *regression of* Y *on* X refers to the computation of conditional expectations of the form

$$E(Y|X=x) \text{ for each } x.$$

The term “regression” was coined by the British statistician Francis Galton (1822–1911) who,

Figure 4.9: *Summary of the Galton-Pearson model for heights of fathers and their grown sons. The solid line is the expected height in inches, $E(Y|X=x)$, of a grown son who's father is x inches tall, for each x. A vertical gray line is drawn at the location of the mean height for fathers (67.7 inches). A horizontal gray line is drawn at the location of the mean height for sons (68.7 inches). The dashed line is $y = x + 1$.*



Fathers' Heights:
$E(X) = 67.7$ inches, $SD(X) = 2.7$ inches

Sons' Heights:
$E(Y) = 68.7$ inches, $SD(X) = 2.7$ inches

Regression of Y on X:
$E(Y|X=x) = 34.85 + 0.5x$, for each x.

together with his associate Karl Pearson (1857–1936), was interested in understanding how characteristics of children were related to those of their parents.

In one study, Galton and Pearson measured the heights of 1078 fathers and their grown sons and developed a model for the joint distribution of father-son heights. The model is summarized in Figure 4.9 (page 91).

Based on the Galton-Pearson model, the distribution of fathers' heights has a mean of $E(X) = 67.7$ inches and a standard deviation of $SD(X) = 2.7$ inches, while the distribution of sons' heights has a mean of $E(Y) = 68.7$ inches and a standard deviation of $SD(Y) = 2.7$ inches. Thus, on average, grown sons are 1 inch taller than fathers. Further, the expected height in inches of a grown son whose father is x inches tall is

$$E(Y|X=x) = 34.85 + 0.5x, \text{ for each } x.$$

Since sons are, on average, 1 inch taller than fathers and since the standard deviations of the height distributions are equal, Galton expected to find that the line $y = x+1$ would represent the conditional expectation. Instead, he got the line $y = 34.85 + 0.5x$ (the solid line in the figure). For each x, the value of y on this line is midway between the values on the lines $y = x + 1$ (the dashed line) and $y = 68.7$ (the horizontal line). This lead Galton to observe that the conditional expected values had "regressed" toward the mean for all sons.

Lastly, after studying the works of Galton and Gregor Mendel (1822–1884),[1] R.A. Fisher proved that Galton's idea of "regression to the mean" was a consequence of the probabilistic models of inheritance pioneered by Mendel (Freedman, et al, 1991, Chap. 25).

Other historical notes. Probability theory has a long and interesting history, beginning with a series of letters exchanged by the French mathematicians Blaise Pascal (1623–1662) and Pierre de Fermat (1601–1665). In these letters, Pascal and Fermat worked out the details of computing the probabilities of certain events in games of chance.

The British mathematician and Presbyterian minister Thomas Bayes (1702–1761) also deserves mention here. Bayes developed *Bayes' rule* (page 75) to solve what he called the "inverse probability" problem. Our main application of Bayes' rule was in the analysis of diagnostic tests, where information is available on the sensitivity and specificity of a test,

$$P(POS|D) \text{ and } P(NEG|D^c),$$

and our interest is in finding the positive and negative predictive values of the test,

$$P(D|POS) \text{ and } P(D^c|NEG).$$

(In each case, we would like to "invert" a probability of the form $P(A|B)$ to obtain a probability of the form $P(B|A)$.)

Bayes' rule for conditional probabilities is also the main method used in a branch of statistics known as *Bayesian statistics*. Bayesian statistical methods are now being applied to studies from many fields, including the health sciences.

Bayesian versus Frequentist debate. We have defined the probability of event A as the proportion of times event A occurs in a sufficiently long series of repetitions of an experiment. This is often called the *frequency definition* of probability since we are concerned with the relative frequency of A as the number of repetitions of the experiment grows large.

But, a probability model can be developed using *any* specification of numbers $P(A)$ consistent with the rules, including a person's personal beliefs about the probabilities of certain events. This fact has inspired a debate between those who believe that we should use methods based on the relative frequency definition only (the *Frequentists*), and those who believe that personal beliefs should be part of statistical analyses (the *Bayesians*).

We will say more about this debate at the end of the next chapter.

[1] Gregor Mendel is the Austrian-born Augustinian monk who is considered to be the father of modern genetics. He postulated the existence of entities now called *genes*, and developed simple chance models to explain how *genotypes* (the set of genes present in an individual) affect *phenotypes* (the observable physical characteristics of an individual).

See https://www.biography.com/scientist/gregor-mendel, for example, to learn more about Gregor Mendel and his work.

5 Families of Probability Distributions

This chapter continues our discussion of probability theory. Three families of discrete probability distributions (the hypergeometric, binomial and Poisson families) and one family of continuous distributions (the normal or Gaussian family) are introduced. Applications are stressed throughout the chapter.

5.1 Hypergeometric Distribution

The *hypergeometric distribution* gives the probability distribution of the number of individuals from a subpopulation of interest in a simple random sample from a study population.

Since formulas for working with this family of distributions depend on rules for counting the number of simple random samples, we begin with a discussion of counting rules.

5.1.1 Computing Numbers of Samples

Let N be the total number of individuals in a study population and assume that each individual is assigned a whole number between 1 and N. A sample of n individuals chosen from the population corresponds to a subset of size n from the set $\{1, 2, \ldots, N\}$.

For example, if $N = 12$ and $n = 5$, then the following graphic is a convenient way to illustrate the choice of individuals assigned the numbers 2, 8, 9, 10 and 11.

1	2	3	4	5	6	7	8	9	10	11	12
	○						○	○	○	○	

Formula for number of subsets of size n. The total number of subsets of size n from the set $\{1, 2, \ldots, N\}$ can be found by applying the following formula:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!},$$

where $\binom{N}{n}$ is read "N choose n" and "!" represents the factorial function.

Example 5.1 ($N = 12$) To illustrate finding the number of subsets by hand with the minimum amount of work, consider a population of size 12. Then the number of samples of size 4 is the same as the number of subsets of size 4 from the set $\{1, 2, \ldots, 12\}$,

$$\binom{12}{4} = \frac{(12\ 11\ 10\ 9\ \cancel{8}\ \cancel{7}\ \cancel{6}\ \cancel{5}\ \cancel{4}\ \cancel{3}\ \cancel{2}\ \cancel{1})}{(4\ 3\ 2\ 1)(\cancel{8}\ \cancel{7}\ \cancel{6}\ \cancel{5}\ \cancel{4}\ \cancel{3}\ \cancel{2}\ \cancel{1})} = \frac{(12\ 11\ 10\ 9)}{(4\ 3\ 2\ 1)} = 495.$$

Similarly, the number of samples of size 7 is

$$\binom{12}{7} = \frac{(12\ 11\ 10\ 9\ 8\ \cancel{7}\ \cancel{6}\ \cancel{5}\ \cancel{4}\ \cancel{3}\ \cancel{2}\ \cancel{1})}{(\cancel{7}\ \cancel{6}\ \cancel{5}\ \cancel{4}\ \cancel{3}\ \cancel{2}\ \cancel{1})(5\ 4\ 3\ 2\ 1)} = \frac{(12\ 11\ 10\ 9\ 8)}{(5\ 4\ 3\ 2\ 1)} = 792.$$

The following table lists the numbers of samples of size n for each possible n.

n	0	1	2	3	4	5	6	7	8	9	10	11	12	Total
$\binom{12}{n}$	1	12	66	220	495	792	924	792	495	220	66	12	1	4096

The table includes the possibility of choosing no individuals and the possibility of choosing all 12 individuals. The total number of samples of all sizes is $2^{12} = 4096$.

Subpopulation of interest. Let M be the total number of individuals in a subpopulation of interest to researchers and assume that these individuals have been assigned the whole numbers between 1 and M. For example, if M of N women in a study population have breast cancer, then the women who have breast cancer would be assigned the numbers between 1 and M and the women who do not have breast cancer would be assigned the numbers between $M + 1$ and N.

A sample of n individuals from the population corresponds to a subset of size n from the set $\{1, 2, \ldots, N\}$. A certain number of the chosen individuals will be from the subpopulation of interest; the remaining will be from the complementary subset.

For example, if $N = 12$, $M = 7$ and $n = 5$, then the following table gives the numbers of individuals from the subpopulation of interest in six different samples:

	1	2	3	4	5	6	7	8	9	10	11	12	*Number from Subpopulation*
Sample 1:		●						○	○	○	○		1
Sample 2:	●			●				○		○		○	2
Sample 3:				●				○	○	○		○	1
Sample 4:	●	●			●		●			○			4
Sample 5:		●			●	●				○	○		3
Sample 6:		●						○	○		○	○	1

(Filled circles correspond to individuals in the sample from the subpopulation of interest; open circles correspond to individuals from the complementary subset.)

Product formula. The total number of samples of size n with exactly x from the subpopulation of interest, and $n - x$ from the complementary subpopulation, can be found by computing the following product:

$$\binom{M}{x} \times \binom{N-M}{n-x} = \frac{M!}{x!(M-x)!} \times \frac{(N-M)!}{(n-x)!(N-M-(n-x))!}.$$

(x numbers are chosen from among the first M, and $n - x$ from among the last $N - M$.)

Example 5.2 $\boldsymbol{(N = 12,\ M = 7,\ n = 5)}$ Consider a population of 12 individuals and a subpopulation of interest of 7 individuals. Then the number of samples of size 5 with exactly 2 from the subpopulation of interest is the same as the number of subsets of size 5 from the set $\{1, 2, \ldots, 12\}$ with exactly 2 numbers chosen from among the first 7,

$$\binom{7}{2} \times \binom{5}{3} = \frac{(7\ 6\ \cancel{5}\ \cancel{4}\ \cancel{3}\ \cancel{2}\ \cancel{1})}{(2\ 1)(\cancel{5}\ \cancel{4}\ \cancel{3}\ \cancel{2}\ \cancel{1})} \times \frac{(5\ 4\ \cancel{3}\ \cancel{2}\ \cancel{1})}{(\cancel{3}\ \cancel{2}\ \cancel{1})(2\ 1)} = 21 \times 10 = 210.$$

Similarly, the number of samples of size 5 with exactly 3 chosen from the subpopulation of interest is

$$\binom{7}{3} \times \binom{5}{2} = \frac{(7\ 6\ 5\ \cancel{4}\ \cancel{3}\ \cancel{2}\ \cancel{1})}{(3\ 2\ 1)(\cancel{4}\ \cancel{3}\ \cancel{2}\ \cancel{1})} \times \frac{(5\ 4\ \cancel{3}\ \cancel{2}\ \cancel{1})}{(2\ 1)(\cancel{3}\ \cancel{2}\ \cancel{1})} = 35 \times 10 = 350.$$

The following table lists the numbers of samples of size 5 with exactly x from the subpopulation of interest for each possible x.

x	0	1	2	3	4	5	Total
$\binom{7}{x} \times \binom{5}{5-x}$	1	35	210	350	175	21	792

Note that the total number of samples is $\binom{12}{5} = 792$.

Example 5.3 **($N = 15$, $M = 7$, $n = 4$)** Consider a population of 15 individuals and a subpopulation of interest of 7 individuals. Then the number of samples of size 4 from this population is

$$\binom{15}{4} = \frac{(15\ 14\ 13\ 12\ \cancel{11}\ \cancel{10}\ \cancel{9}\ \cancel{8}\ \cancel{7}\ \cancel{6}\ \cancel{5}\ \cancel{4}\ \cancel{3}\ \cancel{2}\ \cancel{1})}{(4\ 3\ 2\ 1)(\cancel{11}\ \cancel{10}\ \cancel{9}\ \cancel{8}\ \cancel{7}\ \cancel{6}\ \cancel{5}\ \cancel{4}\ \cancel{3}\ \cancel{2}\ \cancel{1})} = \frac{(15\ 14\ 13\ 12)}{(4\ 3\ 2\ 1)} = 1365,$$

and the numbers of samples of size 4 with exactly x chosen from the subpopulation of interest for each possible x are as follows:

x	0	1	2	3	4	Total
$\binom{7}{x} \times \binom{8}{4-x}$	70	392	588	280	35	1365

Footnotes. A *factorial* is a descending product of integers defined for nonnegative integers. If $n = 0$, then $n! = 0! = 1$. If $n > 0$, then

$$n! = n\ (n-1)\ (n-2)\ \cdots\ 1.$$

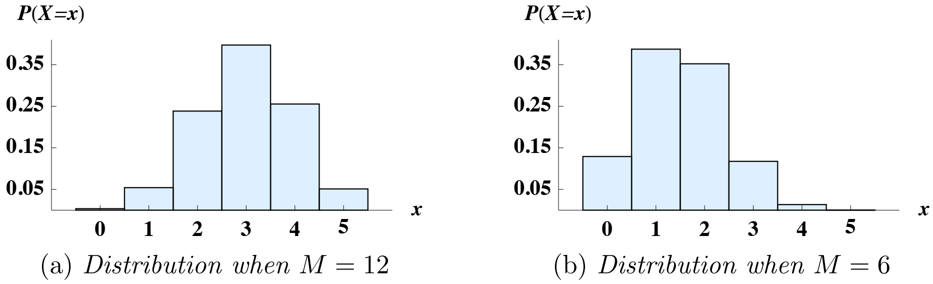
The numbers $\binom{n}{x}$ are often called the *binomial coefficients*. Binomial coefficients satisfy several properties that are useful when doing computations:

$$\binom{n}{0} = \binom{n}{n} = 1; \qquad \binom{n}{1} = \binom{n}{n-1} = n; \qquad \binom{n}{x} = \binom{n}{n-x} \text{ for each } x.$$

For each x, there are $\binom{n}{x}$ ways to choose x individuals (and leave $(n-x)$ behind) or to choose $(n-x)$ individuals (and leave x behind).

Note that most hand-held calculators provide shortcuts for finding binomial coefficients directly. If your calculator does not provide such shortcuts, then the examples above show you how to find the answers by hand with the minimum amount of work.

Figure 5.1: *Probability histograms for hypergeometric random variables based on simple random samples of size 5 from a population of size 20.*



(a) *Distribution when* $M = 12$ (b) *Distribution when* $M = 6$

5.1.2 Formulas for Probability and Summary Measures

Let N be the number of individuals in the study population, M the number of individuals in the subpopulation of interest, n the number of individuals in the simple random sample and X the number of individuals from the subpopulation of interest in the chosen sample.

Probability formula. Since each choice of sample is equally likely, the probability distribution of the hypergeometric random variable X can be computed using the formula

$$P(X = x) = \frac{\binom{M}{x} \times \binom{N-M}{n-x}}{\binom{N}{n}} \text{ for each possible } x.$$

[The probability that there are exactly x individuals from the subpopulation of interest in the sample is the ratio of the number of samples with exactly x from the subpopulation (and exactly $n - x$ from the complementary subpopulation) to the total number of samples.]

Formulas for model summaries. The mean and standard deviation of the hypergeometric distribution are

$$E(X) = np \quad \text{and} \quad SD(X) = \sqrt{np(1-p)\left(\frac{N-n}{N-1}\right)},$$

where $p = M/N$ is the proportion of individuals in the subpopulation of interest.

Example 5.4 ($N = 20$, $n = 5$) Consider choosing a simple random sample of size 5 from a population of size 20.

If there are 12 individuals in the subpopulation of interest, then the probability that there will be exactly x individuals from the subpopulation in the chosen sample is

$$P(X = x) = \frac{\binom{12}{x} \times \binom{8}{5-x}}{\binom{20}{5}} \text{ for each possible } x.$$

The following table gives the probability distribution of X:

x	0	1	2	3	4	5	Total
$P(X = x)$	0.0036	0.0542	0.2384	0.3973	0.2554	0.0511	1.000

Figure 5.2: *Probability histograms for hypergeometric distributions with $N = 20$, $n = 5$. The first row corresponds to $M = 3, 6, 9$, and the second row to $M = 12, 15, 18$. In each case, the mean of the distribution is marked on the horizontal axis.*

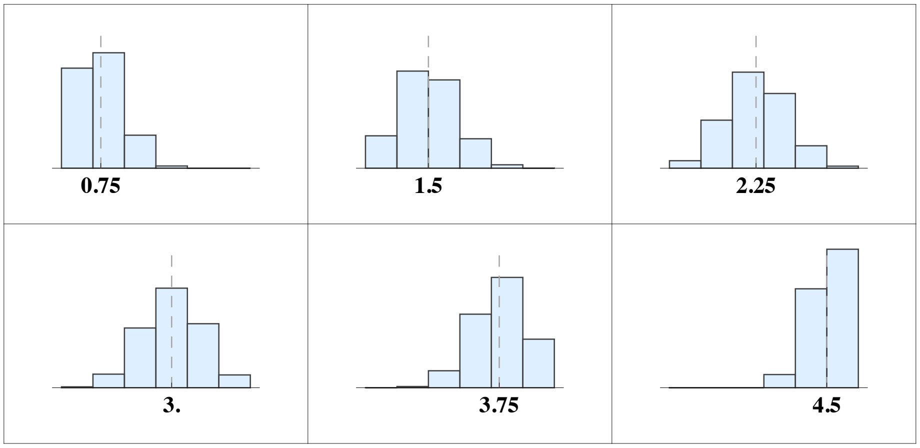


Figure 5.1(a) (page 96) is a probability histogram for X. For this random variable, the mean is $E(X) = 3$ and the standard deviation is $SD(X) = 0.9733$.

If there are only 6 individuals in the subpopulation of interest, then the probability that there will be exactly x individuals from the subpopulation in the chosen sample is

$$P(X = x) = \frac{\binom{6}{x} \times \binom{14}{5-x}}{\binom{20}{5}} \text{ for each possible } x.$$

The following table gives the probability distribution of X:

x	0	1	2	3	4	5	Total
$P(X = x)$	0.1291	0.3874	0.3522	0.1174	0.0135	0.0004	1.000

Figure 5.1(b) (page 96) is a probability histogram for X. For this random variable, the mean is $E(X) = 1.5$ and the standard deviation is $SD(X) = 0.9105$.

Figure 5.2 (page 97) takes this example further by examining how the probability distribution of X changes as the number of elements in subpopulation of interest changes. When $M = 3$, for example, the mean $E(X) = 0.75$ and most of the probability is concentrated near 0. As M increases, the distribution "moves to the right" and the mean increases as well.

Footnotes. Hypergeometric distributions are used to analyze the results of surveys, where the goal is to estimate M (the size of the subpopulation of interest to the researcher) using information from a simple random sample drawn from the study population.

Hypergeometric distributions are also used to analyze *capture-recapture* experiments. Here, a double sampling scheme is applied to a population whose total size (N) is not known.

The goal is to estimate N. Capture-recapture methods have been applied by naturalists to estimate the total size of an animal population, by the U.S. Census Bureau to estimate the total size of the homeless population in a large urban area, and by public health researchers to estimate the total number of individuals with a certain disease. See page 113 for an example in the public health setting.

5.2 Binomial Distribution

The *binomial distribution* gives the probability distribution of the number of successes in a fixed number of trials of an experiment. The underlying assumptions are:

1. ***Fixed Number of Trials:*** There are a total of n trials (or repetitions) of an experiment. Each trial ends in one of two complementary events, A or A^c, where event A corresponds to *success* and event A^c corresponds to *failure*.
2. ***Independence:*** The trials are mutually independent. The result of any one trial has no effect on the results of any of the other trials.
3. ***Constant Probabilities:*** The success probability, $p = P(A)$, is constant from trial to trial.

For example, consider the experiment of choosing a card from a well-shuffled deck of 52 cards and recording its value, and let A be the event that a diamond is chosen. If the experiment is repeated 8 times (which includes returning the card to the deck and reshuffling after every trial) and X is the total number of times a diamond is observed, then X is a binomial random variable based on 8 trials with success probability 1/4.

5.2.1 Formulas for Probability and Summary Measures

Assume that X satisfies the assumptions for a binomial distribution based on n independent trials of an experiment with success probability p.

Probability formula. Since the trials are mutually independent, the probability distribution for the binomial random variable X can be computed using the formula

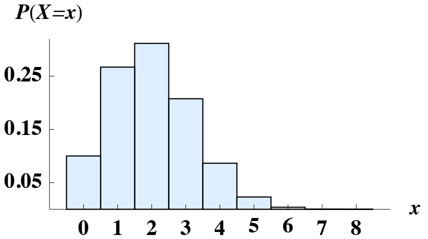
$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for each } x.$$

[The probability that there are exactly x successes in n trials is the number of ways to choose the trial numbers of the successes times the probability that a given list of n results has exactly x successes (and exactly $n - x$ failures).]

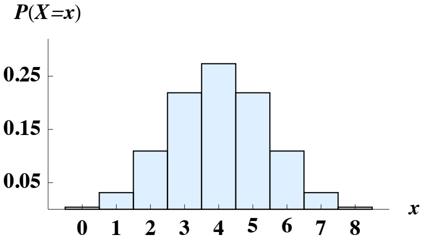
Formulas for model summaries. The mean and standard deviation of the binomial distribution are

$$E(X) = np \quad \text{and} \quad SD(X) = \sqrt{np(1-p)}.$$

Figure 5.3: *Probability histograms for binomial random variables based on 8 trials.*



(a) *Distribution when* $p = \frac{1}{4}$



(b) *Distribution when* $p = \frac{1}{2}$

Example 5.5 (Choosing Cards) Consider 8 repetitions of the following experiment:

> *"Choose a card from a well-shuffled deck of 52 cards, record the value of the card and return the card to the deck."*

If the event of interest is that a diamond is chosen, then the success probability is 1/4 and the probability that exactly x trials end in success is

$$P(X = x) = \binom{8}{x} \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{8-x} \quad \text{for each } x.$$

The following table gives the probability distribution of X:

x	0	1	2	3	4	5	6	7	8	Total
$P(X = x)$	0.1001	0.2670	0.3115	0.2076	0.0865	0.0231	0.0038	0.0004	0.0000	1.000

(The probability of the event "$X = 8$" is zero to 4 decimal places of accuracy.)

Figure 5.3(a) (page 99) is a probability histogram for X. For this random variable, the mean is $E(X) = 2$ and the standard deviation is $SD(X) = 1.2247$.

If the event of interest is that a red card is chosen, then the success probability is 1/2 and the probability that exactly x trials end in success is

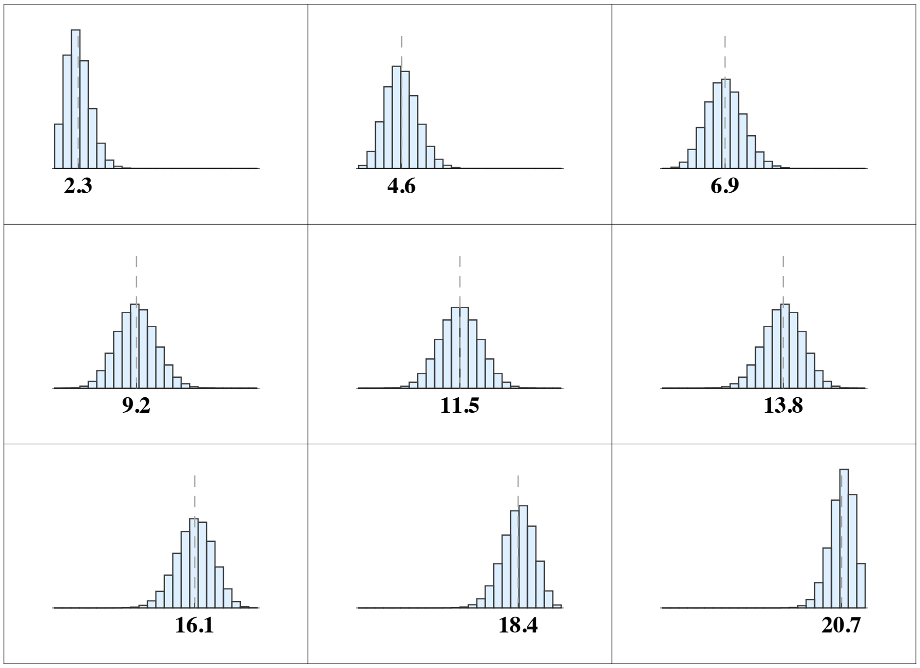
$$P(X = x) = \binom{8}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{8-x} = \binom{8}{x} \left(\frac{1}{2}\right)^8 \quad \text{for each } x.$$

The following table gives the probability distribution of X:

x	0	1	2	3	4	5	6	7	8	Total
$P(X = x)$	0.0039	0.0312	0.1094	0.2188	0.2734	0.2188	0.1094	0.0312	0.0039	1.000

Figure 5.3(b) is a probability histogram for X. For this random variable, the mean is $E(X) = 4$ and the standard deviation is $SD(X) = 1.4142$.

Figure 5.4: *Probability histograms for binomial distributions based on 23 independent trials of an experiment. The first row corresponds to success probabilities $p = 0.1, 0.2, 0.3$, the second row to $p = 0.4, 0.5, 0.6$, and the last row to $p = 0.7, 0.8, 0.9$. In each case, the mean of the distribution is marked on the horizontal axis.*



Example 5.6 **($n = 23$, $p = 0.1, 0.2, \ldots, 0.9$)** Figure 5.4 (page 100) considers how binomial distributions based on $n = 23$ independent trials changes as p changes.

When $p = 0.1$, for example, most of the probability is concentrated near 0 and the mean is $E(X) = 2.3$. At the other extreme, when $p = 0.9$, most of the probability is concentrated near $n = 23$ and the mean is $E(X) = 20.7$. The distribution when $p = 0.5$ is perfectly symmetry around its mean $E(X) = 11.5$.

Relationship to sampling from a population. Consider choosing a simple random sample of size n from a population of N individuals. If there are M individuals in a subpopulation of interest, then the probability that exactly x individuals from the subpopulation of interest are in the chosen sample is

$$P(X = x) = \frac{\binom{M}{x} \times \binom{N-M}{n-x}}{\binom{N}{n}} \text{ for each } x,$$

since X is a hypergeometric random variable.

If N is large, then the binomial distribution based on n trials of an experiment with success probability $p = M/N$ can be used to approximate the hypergeometric distribution. That is,

$$P(X = x) = \frac{\binom{M}{x} \times \binom{N-M}{n-x}}{\binom{N}{n}} \approx \binom{n}{x} p^x (1-p)^{n-x} \text{ for each } x,$$

where $p = M/N$. The success probability in the binomial approximation is the probability that an individual chosen at random from the population is in the subpopulation of interest. The approximation is good as long as the population size is greater than 20 times the sample size ($N > 20n$) and the success probability lies in the interval $0.05 < p < 0.95$.

Footnotes. Binomial probabilities are easy to work with. Further, the binomial approximation to the hypergeometric distribution requires knowledge of the proportion of individuals in the subpopulation of interest (p), but does not require precise knowledge of the total population size (N). For these reasons, the binomial approximation to the hypergeometric distribution is very useful in practice.

Binomial distributions are used to analyze the results of surveys in situations where the binomial distribution is a good approximation to the hypergeometric distribution. The goal is to estimate the proportion p of individuals in a subpopulation of interest. Applications include estimating the proportion of registered voters who support a certain candidate or proposal, or estimating the prevalence of a disease in a given population. Statistical methods for estimating the success probability p are discussed in the next chapter.

5.3 Poisson Distribution

The *Poisson distribution* gives the probability distribution of the number of events that occur in an interval of time or a region of space. The underlying assumptions are:

1. ***Proportionality:*** The probability that a single event occurs is proportional to the length of the time interval or the area of the region of space.
2. ***Unlimited Number of Events:*** Within a single time interval or region of space, any number of events are possible.
3. ***Independence:*** Events occur independently. An event occurring in a particular subinterval of time or subregion of space has no effect on whether an event will occur in any other subinterval or subregion.

Poisson distributions have been used, for example, to model the number of angina attacks a patient suffers in a certain period of time, the number of industrial accidents occurring in a certain period of time, and the number of leukemia cases occurring within a certain distance of a toxic waste site.

A Poisson distribution can be completely specified once the average number of events per unit time or space is known. The Greek letter λ ("lambda") is used to denote the average number of events of a Poisson distribution.

5.3.1 Formulas for Probability and Summary Measures

Let X be the number of events occurring in one unit of time or space, where the occurrences of events satisfy the three assumptions above, and let λ be the average number of events.

Probability formula. The probability distribution for the Poisson random variable X can be computed using the following formula:

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \text{ for each } x.$$

The formula can be used to compute the probability that exactly x events occur for any number of events (0, 1, 2, ...), although the values are virtually zero when x is very large. The number e in the formula is Euler's constant, $e \approx 2.71828$.

Formulas for model summaries. The mean and standard deviation of the Poisson distribution are

$$E(X) = \lambda \text{ and } SD(X) = \sqrt{\lambda}.$$

Example 5.7 (Traffic Patterns) The numbers of cars passing a particular intersection in one hour's time depends on the time of day.

For example, let X be the number of cars passing the intersection in one hour's time in the evening and assume that X has a Poisson distribution with $\lambda = 1.8$ (on average, 1.8 cars will pass the intersection in the hour). Then the probability that exactly x cars pass the intersection in one hour is

$$P(X = x) = \frac{e^{-1.8}(1.8)^x}{x!} \text{ for each } x.$$

The following table gives the probability distribution of X when $x \leq 9$.

x	0	1	2	3	4	5	6	7	8	9
$P(X = x)$	0.1653	0.2975	0.2678	0.1607	0.0723	0.026	0.0078	0.002	0.0005	0.0001

When $x \geq 10$, $P(X = x)$ is zero to 4 decimal places of accuracy.

By contrast, let X be the number of cars passing the intersection in one hour's time in mid-afternoon and assume that X has a Poisson distribution with $\lambda = 6.2$ (on average, 6.2 cars will pass the intersection in the hour). Then the probability that exactly x cars pass the intersection in one hour is

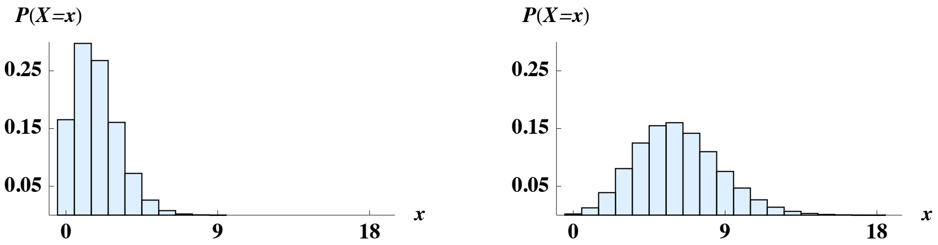
$$P(X = x) = \frac{e^{-6.2}(6.2)^x}{x!} \text{ for each } x.$$

The following table gives the probability distribution of X when $x \leq 18$.

x	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$P(X = x)$	0.0020	0.0126	0.0390	0.0806	0.1249	0.1549	0.1601	0.1418	0.1099	0.0757	0.0469	0.0265	0.0137	0.0065	0.0029	0.0012	0.0005	0.0002	0.0001

When $x \geq 19$, $P(X = x)$ is zero to 4 decimal places of accuracy.

Figure 5.5: *Probability histograms for Poisson random variables.*



(a) *Distribution when* $\lambda = 1.8$ (b) *Distribution when* $\lambda = 6.2$

Figure 5.5(a) (page 103) is a probability histogram for the model for evening traffic (where the average is 1.8 cars per hour) and Figure 5.5(b) is a probability histogram for the model for mid-afternoon traffic (where the average is 6.2 cars per hour). Since the standard deviation for the first model is $\sqrt{1.8} = 1.34$ cars per hour and the standard deviation for the second model is $\sqrt{6.2} = 2.49$ cars per hour, the distribution for evening traffic is much more concentrated than the distribution for mid-afternoon traffic.

Example 5.8 $(\boldsymbol{\lambda = 5, 15, 25, 35})$ Figure 5.6 (page 104) considers how Poisson distributions change as λ changes.

In each case, probability is concentrated near the mean of the distribution and falls off (that is, approaches 0) as you go further from the mean in either direction. As λ increases, the distribution becomes wider, flatter and more symmetric.

Relationship to binomial distribution. Let X be the number of successes in n independent trials of an experiment and let p be the probability of success in each trial. Then, the probability that exactly x trials end in success is

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for each } x,$$

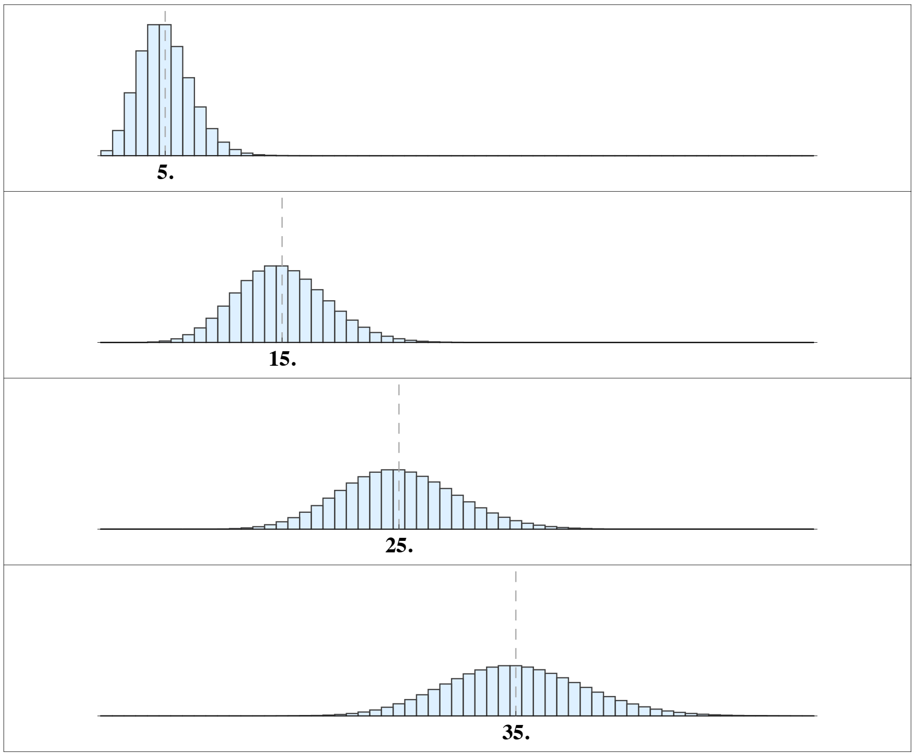
since X is a binomial random variable.

If n is large and p is small, then the Poisson distribution with $\lambda = np$ can be used to approximate the binomial distribution. That is,

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \approx \frac{e^{-\lambda}\lambda^x}{x!} \text{ for each } x,$$

where $\lambda = np$. The average of the Poisson approximation is the same as the average of the binomial distribution. The approximation is good as long as $n > 100$ and $np < 5$.

Figure 5.6: *Probability histograms for Poisson random variables with means 5, 15, 25, 35. In each case, the mean of the distribution is marked on the horizontal axis.*



Example 5.9 (Leukemia Clusters) (Larsen & Marx, 1986, page 196)

> "Leukemia is a rare form of cancer whose cause and mode of transmission remain largely unknown. While evidence abounds that excessive exposure to radiation can increase a person's risk of contracting the disease, it is at the same time true that most cases occur among persons whose history contains no such overexposure. A related issue, one maybe even more basic than the causality question, concerns the *spread* of the disease. It is safe to say that the prevailing medical opinion is that most forms of leukemia are not contagious – still, the hypothesis persists that *some* forms of the disease, particularly the childhood variety, may be. What continues to fuel this speculation are the discoveries of so-called 'leukemia clusters,' aggregations in time and space of unusually large numbers of cases.
>
> "[A frequently cited leukemia clusters case, reported in *The American Journal of Medicine* (1963) 34:796-812,] occurred during the 1950s and early 1960s in Niles, Illinois, a suburb of Chicago. In the $5\frac{1}{3}$-year period from 1956 to the first four months of 1961, physicians in Niles reported a total of eight cases of leukemia

among children less than 15 years of age. The number at risk (that is, the number of residents in that age range) was 7076. To assess the likelihood of that many cases occurring in such a small population, it is necessary to look first at the leukemia incidence in neighboring towns. For all of Cook county, excluding Niles, there were 1,152,695 children less than 15 years of age – and among those, 286 diagnosed cases of leukemia. That gives an average $5\frac{1}{3}$-year leukemia rate of 24.8 cases per 100,000 population:

$$\frac{\text{286 cases for } 5\frac{1}{3} \text{ years}}{\text{1,152,695 children}} \times \frac{100,000}{100,000} = \text{24.8 cases per 100,000 per } 5\frac{1}{3} \text{ years.}$$

"Now, imagine the 7076 children in Niles to be a series of $n = 7076$ [independent trials, each with probability]

$$p = \frac{24.8}{100,000} = 0.000248 \text{ of contracting leukemia.}$$

The question then becomes, given an n of 7076 and a p of 0.000248, how likely is it that eight *or more* 'successes' would occur? (The expected number, of course, would be $7076 \times 0.000248 = 1.75$.) . . . If the probability associated with this event is very small, it could be argued that leukemia did not occur randomly in Niles and that, perhaps, contagion was a factor.

"Using the binomial distribution, we can express the probability of eight or more cases as

$$P(\text{8 or more cases}) = \sum_{x=8}^{7076} \binom{7076}{x} (0.000248)^x (0.999752)^{7076-x}.$$

[Using the Poisson approximation with $\lambda = np = 1.75$, this probability becomes]

$$P(X \geq 8) = 1 - P(X \leq 7) \approx 1 - \sum_{x=0}^{7} \frac{e^{-1.75}(1.75)^x}{x!} = 1 - 0.99951 = 0.00049.$$

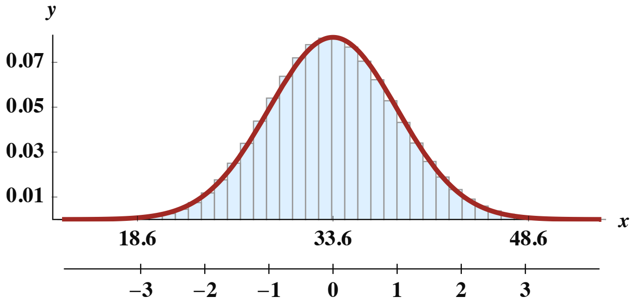
. . . The fact that the probability is so very small tends to denigrate the hypothesis that leukemia in Niles occurred at random. On the other hand, rare events, such as clusters, *do* happen by chance. The basic difficulty in putting the probability associated with a given cluster in any meaningful perspective is not knowing in how many similar communities leukemia did not exhibit a tendency to cluster. That there is no obvious way to do this is one reason the leukemia controversy is still with us."

Footnotes. The Poisson distribution is often called the *distribution of rare events*, since it is used as an approximation to the binomial distribution when p is small (that is, when success is a rare event). Since Poisson probabilities are easier to compute than binomial probabilities when n is large and p is small, the Poisson approximation to the binomial distribution is very useful in practice.

In the leukemia clusters example above, researchers used the value of p for Cook county excluding Niles to compute the complementary probabilities

$$P(\text{8 or more cases}) \text{ and } P(\text{Fewer than 8 cases}).$$

Figure 5.7: *Probability histogram for the binomial distribution based on 120 trials with success probability 0.28, with the normal approximation curve superimposed. The horizontal line below the plot shows standardized scores $z = (x - \mu)/\sigma$.*



That is, they compared the probability of seeing an event (number of cases) as extreme or more extreme than the observed event to the probability of seeing an event that was closer to what might be expected for the given value of p. Since 8 or more cases would occur only 0.049% of the time (compared to 99.951% of the time for fewer than 8 cases), the event is unusual enough to warrant further study. The process of comparing complementary probabilities is part of *significance testing*, a subject we will cover in the next chapter.

5.4 Normal Distribution

The *normal distribution*, also known as the *Gaussian distribution*, is an idealized continuous distribution that is fundamental to much of statistical inference. The normal distribution is symmetric, "bell shaped," and characterized by its mean μ and standard deviation σ. The formula for the normal curve is

$$y = f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/(2\sigma^2)} \text{ for all } x.$$

5.4.1 Relationship to Binomial Distribution

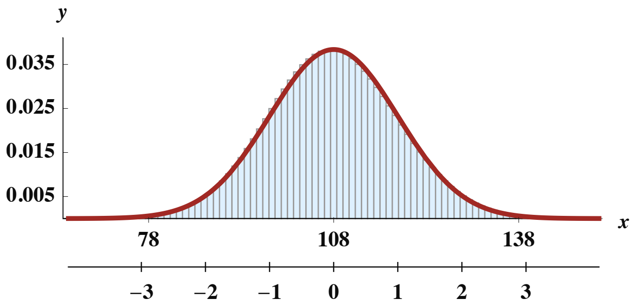
The normal distribution can be used to approximate the binomial distribution. The approximation is good when n is large enough so that the expected number of successes is at least 15 and the expected number of failures is at least 15,

$$np \geq 15 \text{ and } n(1-p) \geq 15.$$

For example, let X be a binomial random variable based on 120 trials with success probability 0.28. For this random variable, the mean and standard deviation are

$$E(X) = np = 33.6 \text{ and } SD(X) = \sqrt{np(1-p)} = 4.92.$$

Figure 5.8: *Probability histogram for the Poisson distribution with mean 108, with the normal approximation curve superimposed. The horizontal line below the plot shows standardized scores $z = (x - \mu)/\sigma$.*



Since both the expected number of successes ($np = 33.6$) and the expected number of failures ($n(1-p) = 86.4$) exceed 15, the normal approximation will be quite good.

Figure 5.7 (page 106) shows the probability histogram of X for values of the random variable satisfying $13 \le x \le 54$. (Probabilities for values of x outside this range are zero to 4 decimal places of accuracy.) The normal approximation curve with $\mu = 33.6$ and $\sigma = 4.92$ is superimposed; the curve approximates the shape of the probability histogram very well. Note that, for convenience, a separate horizontal axis of *standardized scores*,

$$z = (x - \mu)/\sigma, \text{ where } \mu = 33.6 \text{ and } \sigma = 4.92,$$

is drawn below the plot.

5.4.2 Relationship to Poisson Distribution

The normal distribution can also be used to approximate the Poisson distribution. The approximation is good when the average number of events is greater than 100 ($\lambda > 100$).

For example, let X be a Poisson random variable with mean 108 (on average, there are 108 events in one unit of time or space). For this random variable, the mean and standard deviation are

$$E(X) = \lambda = 108 \text{ and } SD(X) = \sqrt{\lambda} = 10.39.$$

Figure 5.8 (page 107) shows the probability histogram of X for values of the random variable satisfying $65 \le x \le 151$. (Probabilities for values of x outside this range are zero to 4 decimal places of accuracy.) The normal approximation curve with $\mu = 108$ and $\sigma = 10.39$ is superimposed; the curve approximates the shape of the probability histogram very well. Note that, for convenience, a separate horizontal axis of *standardized scores*,

$$z = (x - \mu)/\sigma, \text{ where } \mu = 108 \text{ and } \sigma = 10.39,$$

is drawn below the plot.

5.4.3 Computing Normal Probabilities

Area under the normal curve is used to represent probability. Although each choice of mean μ and standard deviation σ gives a different distribution, we can compute probabilities by converting any measurement x into a standardized score $z = (x-\mu)/\sigma$ and using Tables 5.1-5.2 (pages 109-110). These tables give probabilities of the form

$$P(Z \leq z) \text{ where } Z \text{ is the normal random variable with } \mu = 0 \text{ and } \sigma = 1.$$

Standard normal distribution. Z is called the *standard normal random variable*. The distribution of Z is called the *standard normal distribution* (the distribution of standardized scores when observations are normally distributed). A probability of the form $P(Z \leq z)$ is called a *cumulative probability* (or *left-tail probability* or *lower-tail probability*).

The tables include cumulative probabilities for the following standardized scores:

$$z = -3.89, -3.88, -3.87, -3.86, \ldots, 3.86, 3.87, 3.88, 3.89.$$

Each z is obtained by adding a number in the left column to a number in the top row. The table entry for z is the cumulative probability. For example, the cumulative probability $P(Z \leq -3.89) = 0.0001$ is the entry in row -3.8 and column -0.09 of Table 5.1.

Probabilities of the form $P(z_1 \leq Z \leq z_2)$ are obtained by subtracting table entries:

$$P(z_1 \leq Z \leq z_2) = P(Z \leq z_2) - P(Z \leq z_1).$$

Technically, the difference of table entries is $P(z_1 < Z \leq z_2)$, but the results are the same since we are working with a continuous random variable. For example,

$$P(-3.89 \leq Z \leq 3.89) = 0.9999 - 0.0001 = 0.9998.$$

(This computation tells us that the central 99.98% of the distribution is represented in the tables. $P(Z \leq 3.89) = 0.9999$ is the entry in row 3.8 and column 0.09 of Table 5.2.)

A *right-tail probability* (or *upper-tail probability*) is a probability of the form $P(Z \geq z)$. Right-tail probabilities are obtained using the complement rule:

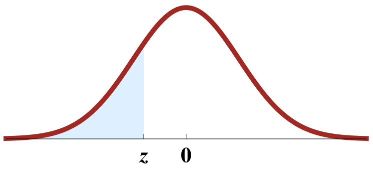
$$P(Z \geq z) = 1 - P(Z \leq z).$$

Technically, the complement rule gives us $P(Z > z)$, but the results are the same since we are working with a continuous random variable. For example,

$$P(Z \geq 1.54) = 1 - P(Z \leq 1.54) = 1 - 0.9382 = 0.0618.$$

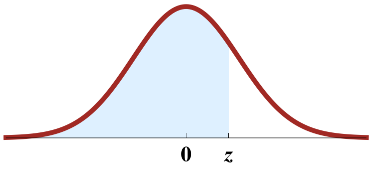
($P(Z \leq 1.54) = 0.9382$ is the entry in row 1.5 and column 0.04 of Table 5.2.)

Table 5.1: *Standard normal cumulative probabilities, $P(Z \leq z)$, when $z \leq 0$.*



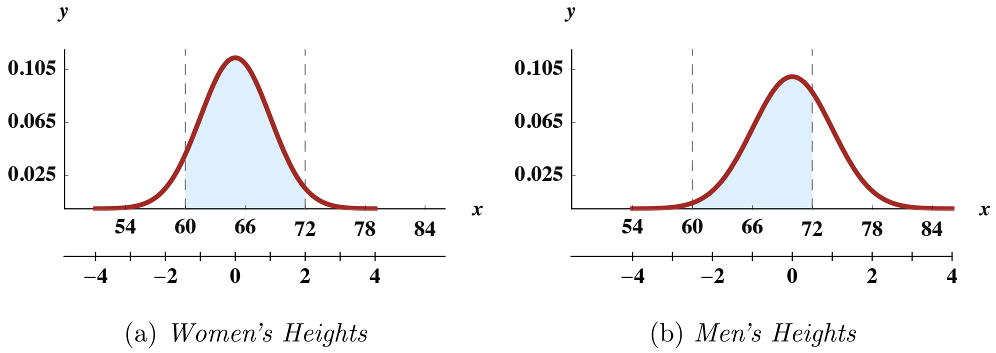
z	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01	-0.00
-3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.6	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002
-3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.4	0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
-3.3	0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005
-3.2	0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007
-3.1	0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010
-3.0	0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013
-2.9	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019
-2.8	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026
-2.7	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035
-2.6	0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047
-2.5	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062
-2.4	0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082
-2.3	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107
-2.2	0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139
-2.1	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179
-2.0	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228
-1.9	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287
-1.8	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359
-1.7	0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446
-1.6	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548
-1.5	0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668
-1.4	0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808
-1.3	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968
-1.2	0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131	0.1151
-1.1	0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357
-1.0	0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587
-0.9	0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841
-0.8	0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119
-0.7	0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389	0.2420
-0.6	0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743
-0.5	0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085
-0.4	0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446
-0.3	0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821
-0.2	0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168	0.4207
-0.1	0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602
-0.0	0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960	0.5000

Table 5.2: *Standard normal cumulative probabilities, $P(Z \le z)$, when $z \ge 0$.*



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999

Figure 5.9: *Normal distributions of heights in inches of adults living in North America, with heights between 5 feet (60 inches) and 6 feet (72 inches) are highlighted. The distribution of women's heights has mean 65 inches and standard deviation 3.5 inches. The distribution of men's heights has mean 70 inches and standard deviation 4 inches. Horizontal axes of standardized scores are shown below each distribution.*



(a) *Women's Heights*

(b) *Men's Heights*

Example 5.10 (Adult Heights) (Agresti & Franklin, 2007, page 306) Distributions of heights for adult men and for adult women are often well-approximated by normal distributions. In North America, for example,

1. *Women's Heights:* the distribution of heights for adult women is well-approximated by a normal distribution with mean 65 inches and standard deviation 3.5 inches and
2. *Men's Heights:* the distribution of heights for adult men is well-approximated by a normal distribution with mean 70 inches and standard deviation 4 inches.

To use the normal model for women's heights to find the probability that a woman chosen at random from the population of adult women living in North America is between 5 feet (60 inches) and 6 feet (72 inches) tall, we convert the heights to standardized scores,

$$z_1 = (60-65)/3.5 = -1.43 \text{ and } z_2 = (72-65)/3.5 = 2.00,$$

and use the tables to compute the probability:

$$P(\text{Height between 60 and 72 in.}) = P(-1.43 \le Z \le 2.00) = 0.9772 - 0.0764 = 0.9008.$$

Figure 5.9(a) (page 111) shows the model for women's heights in inches, with heights between 5 feet and 6 feet highlighted. The area of the gray region is 0.9008.

To use the normal model for men's heights to find the probability that a man chosen at random from the population of adult men living in North America is between 5 feet (60 inches) and 6 feet (72 inches) tall, we convert the heights to standardized scores,

$$z_1 = (60-70)/4 = -2.50 \text{ and } z_2 = (72-70)/4 = 0.50,$$

and use the tables to compute the probability:

$$P(\text{Height between 60 and 72 in.}) = P(-2.50 \leq Z \leq 0.50) = 0.6915 - 0.0062 = 0.6853.$$

Figure 5.9(b) shows the model for men's heights in inches, with heights between 5 feet and 6 feet highlighted. The area of the gray region in this case is 0.6853.

Using the normal models, we find that 90.08% of adult women living in North America have heights between 5 and 6 feet, compared to only 68.53% of adult men.

Footnotes. Normal distributions can be used to model physical characteristics (such as height, weight, cholesterol level, blood pressure level) in specific populations, and to approximate binomial and Poisson distributions. More generally, normal distributions can be used to approximate distributions of sums and averages, as we will learn in the next chapter.

One way to compare the height distributions in the last example is to consider the probabilities of the three mutually exclusive and exhaustive events:

Height under 5 feet, Height between 5 and 6 feet, Height over 6 feet.

Based on this partition of heights, we have the following probability distributions:

Probability Distribution for
Women's Heights in North America

Under 5ft	Between 5ft & 6ft	Over 6ft	Total
0.0764	0.9008	0.0228	1.0000

Probability Distribution for
Men's Heights in North America

Under 5ft	Between 5ft & 6ft	Over 6ft	Total
0.0062	0.6853	0.3085	1.0000

The tables include the central probability of interest (the probability that the height is between 5 and 6 feet), and lower and upper tail probabilities.

Since the standard normal random variable is *continuous*, the following probabilities are equal for all values of z_1 and z_2:

$$P(z_1 < Z < z_2) = P(z_1 < Z \leq z_2) = P(z_1 \leq Z < z_2) = P(z_1 \leq Z \leq z_2).$$

This result is true because probability corresponds to area under the normal curve, and the area is the same whether we include the endpoints or not. By contrast, when working with *discrete* random variables, probabilities for intervals *with* endpoints are generally different than for intervals *without* endpoints.

5.5 Brief Summary and Additional Examples

This chapter introduces four important families of probability distributions and discusses the relationships among these families. The four families will be used extensively in later chapters.

All models are wrong, but... A famous saying, attributed to many statisticians and mentioned at the end of the last chapter, is

"All models are wrong, but some are useful."

The four models studied in this chapter have proven to be useful in many applications.

Although the term "hypergeometric distribution" appeared in the literature for the first time in the early 1900's, the distribution has been used in scientific applications since the 1700's. The Swiss mathematician James Bernoulli (1655-1705) is credited with introducing the binomial distribution and studying its properties. The Poisson distribution is named for the French mathematician Simeon Denis Poisson (1781-1840), who introduced the distribution in the context of studying the occurrence of wrongful convictions of prisoners in a given country. Many names have been associated with the introduction of the normal distribution, most prominently the German mathematician Carl Friedrich Gauss (1777-1855) and the French mathematician Pierre-Simon Laplace (1749-1827). The normal distribution is often called the "Gaussian distribution" of honor of Carl Friedrich Gauss.

Finally, it should be noted that each mathematician named in the previous paragraph is known for contributions in many scientific disciplines.

Estimating population characteristics. In applications of probability models to real world problems, we are often interested in using sample data to estimate an unknown population characteristic, such as the probability of success (p) of a certain intervention, the mean level (μ) of blood pressure in the population, or the even the total size of the population (N).

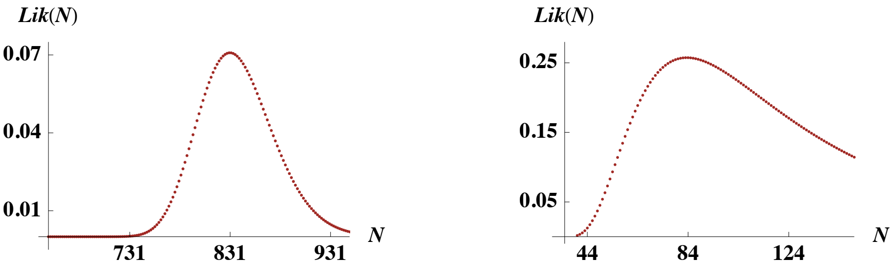
One of the most important estimation methods in statistics (known as the *method of maximum likelihood*), was developed by the British statistician R.A. Fisher (page 19). An interesting application of Fisher's method uses hypergeometric probabilities to estimate the total size of a population (N) based on information from two overlapping but incomplete sources:

1. *First source:* Information from a first main source is used to identify a subpopulation of interest, whose size is M.
2. *Second source:* Information from a second independent source is used to identify a sample from the population, whose size is n.
3. *Overlap:* Individuals identified using both sources form the overlap. Let x be the size of the overlap.

Example 5.11 (Spina Bifida) (Regal & Hook, *Biometrics* (1999) 55:1241-46) Spina bifida is a rare spinal column defect that can be treated but not cured. Treatments for spina bifida include surgery, medication and physical therapy. This example considers estimating the number of babies born with spina bifida in upstate New York between 1969 and 1974, using information from birth and death certificates as the primary source of information on the disorder, and rehabilitation files as the secondary source.

The researchers identified $M = 566$ cases of spina bifida by examining the birth and death certificates of all live births in upstate New York between 1969 and 1974. Rehabilitation files included information on $n = 188$ cases, with an overlap of $x = 128$ cases. Since $M - x = 438$ cases were identified from birth and death certificates only and $n-x = 60$ cases were identified from rehabilitation files only, the sources give incomplete information about the total number of spina bifida cases.

Figure 5.10: *Likelihood plots for spina-bifida example.*



(a) *All Babies* (b) *African-American Babies*

To estimate N, we evaluate the probability of the event

"There are 128 overlapping cases when the total population size is N"

for each possible N, and choose the event with the maximum probability (or maximum *likelihood*) based on the observed. Let $Lik(N)$ ("likelihood of N") represent the probability for a given total population size. When $M = 566$, $n = 188$ and $x = 128$,

$$Lik(N) = \frac{\binom{566}{128}\binom{N-566}{60}}{\binom{N}{188}} \text{ for each } N.$$

Figure 5.10(a) (page 114) is a graph of this likelihood function for values of N between 650 and 950. The function is maximized when N equals 831, the maximum likelihood estimate of N from the observed data.

The researchers were also interested in estimating the number of spina bifida cases among African-Americans born in upstate New York between 1969 and 1974. They identified $M = 28$ cases by examining birth and death certificates and $n = 12$ cases by examining rehabilitation files, with an overlap of $x = 4$ cases. To estimate N, we evaluate the probability of the event

"There are 4 overlapping cases when the total population size is N"

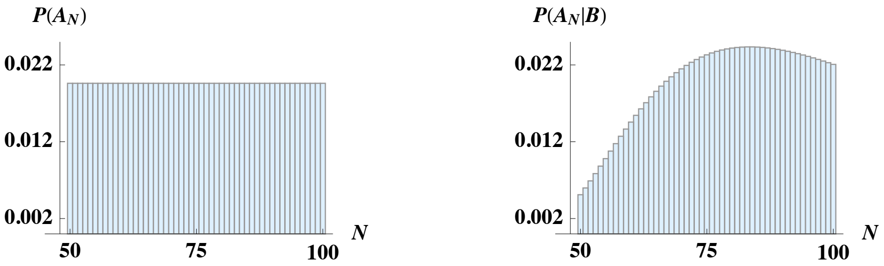
for each possible N, and choose the event with the maximum probability. The likelihood function now becomes

$$Lik(N) = \frac{\binom{28}{4}\binom{N-28}{8}}{\binom{N}{12}}, \text{ for each } N.$$

Figure 5.10(b) (page 114) is a graph of this likelihood function for values of N between 40 and 150. The function is maximized when N equals 84, the maximum likelihood estimate of N from the observed data.

Bayesian versus Frequentist debate. As stated in the previous chapter, we have defined the probability of event A as the proportion of times event A occurs in a sufficiently long series of repetitions of an experiment. This is often called the *frequency definition* of probability

Figure 5.11: *Probability histograms for total population size distributions.*



(a) *Prior Distribution* (b) *Posterior Distribution*

since we are concerned with the relative frequency of A as the number of repetitions of the experiment grows large.

But, a probability model can be developed using *any* specification of numbers $P(A)$ consistent with the rules, including a person's personal beliefs about the probabilities of certain events. This fact has inspired a debate between those who believe that we should use methods based on the relative frequency definition only (the *Frequentists*), and those who believe that personal beliefs should be part of statistical analyses (the *Bayesians*).

Example 5.12 (Spina Bifida, continued) As an example of the Bayesian approach, consider again the problem of estimating the number (N) of spina bifida cases among African-Americans born in upstate New York between 1969 and 1974, using birth and death certificates as the primary source, and rehabilitation files as the secondary source, of information on spina bifida.

Let A_N be the event "the total population size is N." Assume that, before gathering all the information, the researchers believed that N was equally likely to be any whole number between 50 and 100. That is, assume that

$$P(A_N) = \tfrac{1}{51} \text{ for } N = 50, 51, \ldots, 99, 100.$$

Thus, prior to gathering all the information, the researchers treated N as a random variable whose distribution (called the *prior distribution* of N) is shown in Figure 5.11(a) (page 115). The mean of the prior distribution,

$$E(N) = \tfrac{1}{51}(50 + 51 + \ldots + 99 + 100) = 75,$$

is the researchers' estimate of N before gathering all the information.

Assume, as before, that the researchers identified $M = 28$ cases by examining birth and death certificates and $n = 12$ cases by examining rehabilitation files, with an overlap of $x = 4$ cases. Let B be the event "there are 4 overlapping cases." Then the conditional probability of B given each choice of N is

$$P(B|A_N) = \frac{\binom{28}{4}\binom{N-28}{8}}{\binom{N}{12}} \text{ for } N = 50, 51, \ldots, 99, 100.$$

Bayes' rule can now be used to modify the researchers' prior beliefs to produce a new distribution for N. Specifically, Bayes' rule is used compute the probabilities

$$P(A_N|B) \text{ for } N = 51, 52, \ldots, 99, 100.$$

Computations for four values of N are shown below:

				$P(A_N \cap B) =$	$P(A_N\|B) =$
Event	$P(A_N)$	$P(B\|A_N)$		$P(A_N)P(B\|A_N)$	$P(A_N \cap B)/P(B)$
A_{50}	1/51	0.0539		0.0011	0.0051
A_{51}	1/51	0.0632		0.0012	0.0060
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_{99}	1/51	0.2357		0.0046	0.0223
A_{100}	1/51	0.2333		0.0046	0.0221
			$P(B) =$	0.2073	

Thus, after gathering all the information, the researchers now treat N as a random variable whose distribution (called the *posterior distribution* of N) is shown in Figure 5.11(b). The mean of the posterior distribution,

$$E(N|4 \text{ in overlap}) = 50(0.0051) + 51(0.0060) + \ldots + 99(0.0223) + 100(0.0221) = 78.59,$$

is the researchers' estimate of N after gathering all the information. If more information were to become available, the researchers' could further update their estimate, using the posterior distribution as a new prior distribution.

Relationships among the families. The relationships among the four families of distributions studied in this chapter are important in applications. For example,

1. The large sample binomial approximation to hypergeometric probabilities,

   $$P(X = x) = \frac{\binom{M}{x} \times \binom{N-M}{n-x}}{\binom{N}{n}} \approx \binom{n}{x} p^x (1-p)^{n-x} \text{ for each } x,$$

   where $p = M/N$, will be used when the population size is greater than 20 times the sample size ($N > 20n$) and the success probability lies in the interval $0.05 < p < 0.95$.

2. The large sample Poisson approximation to binomial probabilities,

   $$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \approx \frac{e^{-\lambda}\lambda^x}{x!} \text{ for each } x,$$

   where $\lambda = np$, will be used when the number of trials is greater than 100 ($n > 100$) and the expected number of success is less than 5 ($np < 5$).

In addition, the large sample normal approximations to binomial probabilities (Section 5.4.1, page 106) and to Poisson probabilities (Section 5.4.2, page 107) are key components of our discussion of statistical inference in the next chapter.

6 Introduction to Statistical Inference

The term *statistical inference* refers to a collection of methods for generalizing from a sample to a population. This chapter is the first of three chapters introducing statistical inference. Additional applications are given in later chapters. References for these chapters include the texts by Agresti & Franklin (2007, Chaps 6–9), Baldi & Moore (2009, Chaps 13–20), Freedman et al (1991, Parts V–VIII), Moore & McCabe (1999, Chaps 5–8), Moore & Notz (2006, Part IV) and Pagano & Gauvreau (2000, Chaps 8–11, 14).

6.1 Sampling Distributions

Our study of statistical inference begins with the concept of a sampling distribution.

Variables, statistics and parameters. Recall that a *variable* is a characteristic of an individual, a *statistic* is a numerical summary of the values of a variable based on a sample of individuals, and a *parameter* is a numerical summary of values for the entire population.

For example, suppose that researchers are interested in the mean survival time in years from diagnosis for individuals in a certain population of cancer patients. Then the characteristic of interest is survival time in years from diagnosis, the parameter of interest is the mean for all individuals in the population, and the statistic of interest is the sample mean for individuals in a particular sample drawn from the population.

Sampling distribution of a statistic. The *sampling distribution* of a statistic is the probability distribution of values taken by the statistic in all possible samples of the same size from the same population.

Example 6.1 ($N = 7,\ n = 2, 3, 4, 5$) To illustrate the concept of sampling distribution, assume that a population has 7 individuals (numbered 1 through 7), and that the values of a variable of interest for these individuals are given in the second row of the following table:

Individual Number	1	2	3	4	5	6	7
Variable Value	1	3	5	6	6	10	11

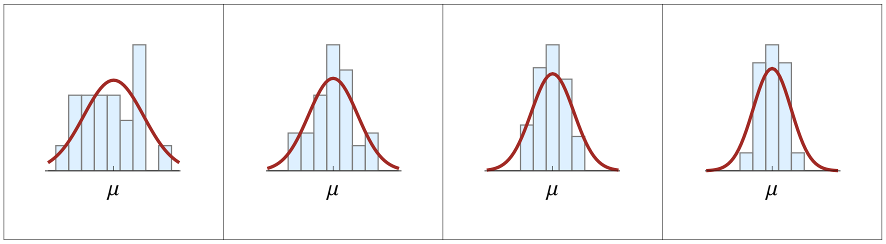
In addition, assume that the parameter of interest is the mean value for all individuals,

$$\mu = \tfrac{1}{7}\,(1 + 3 + 5 + 6 + 6 + 10 + 11) = 6.$$

Let $\overline{X}$ be the sample mean of a simple random sample of size 2 from the population. If, for example, the sample consists of the 3rd and 5th individuals, then the value of the sample mean is $\overline{x} = \frac{1}{2}(5 + 6) = 5.5$. If the sample consists of the 4th and 6th individuals, then the value of the sample mean is $\overline{x} = \frac{1}{2}(6 + 10) = 8$.

Since a simple random sample corresponds to a subset chosen in such a way that each choice of subset is equally likely (using, for example, the method described on page 8), the table on

Figure 6.1: *Sampling distributions of $\overline{X}$ when $N = 7$ and $n = 2, 3, 4, 5$, respectively, are represented as histograms with unit bases. In each case, a normal distribution with the same mean and standard deviation as $\overline{X}$ is superimposed.*



the left below shows the probability distribution of the sample mean $\overline{X}$ when $n = 2$, and the table on the right shows the work needed to construct the distribution.

$\overline{x}$	$P(\overline{X} = \overline{x})$	Samples of size 2
2.0	1/21	$\{1, 2\}$
3.0	1/21	$\{1, 3\}$
3.5	2/21	$\{1, 4\}$, $\{1, 5\}$
4.0	1/21	$\{2, 3\}$
4.5	2/21	$\{2, 4\}$, $\{2, 5\}$
5.5	3/21	$\{1, 6\}$, $\{3, 4\}$, $\{3, 5\}$
6.0	2/21	$\{1, 7\}$, $\{4, 5\}$
6.5	1/21	$\{2, 6\}$
7.0	1/21	$\{2, 7\}$
7.5	1/21	$\{3, 6\}$
8.0	3/21	$\{3, 7\}$, $\{4, 6\}$, $\{5, 6\}$
8.5	2/21	$\{4, 7\}$, $\{5, 7\}$
10.5	1/21	$\{6, 7\}$
	21/21	

There are a total of $\binom{7}{2} = 21$ samples of size 2 from the population of 7 individuals. The mean is 2.0 for exactly 1 sample (the sample with the 1st and 2nd individuals), the mean is 3.0 for exactly 1 sample (the sample with the 1st and 3rd individuals), and so forth. Since the choice of each subset is equally likely, the probabilities

$$P(\overline{X} = \overline{x}) \text{ can be computed as simple ratios.}$$

The left part of Figure 6.1 (page 118) represents the sampling distribution of the sample mean $\overline{X}$ when $n = 2$ as a histogram, with probabilities grouped into 1-unit intervals, and the remaining parts of the figure are representations for samples of sizes $n = 3, 4, 5$, respectively. Each plot is superimposed with a normal curve with same the mean and standard deviation as the $\overline{X}$ distribution. Notice that the distributions become more concentrated around $\mu = 6$, and more bell shaped, as n increases.

6.2 Large Sample Approximate Distributions

This section introduces the sampling distributions we will need to answer statistical questions about the population mean (μ), the population proportion (p) and the Poisson mean rate (λ) when sample sizes are large.

6.2.1 Sampling Distribution of the Sample Mean

Let X represent the value of a characteristic of interest for an individual chosen at random from a study population and suppose that the mean and standard deviation of X are

$$E(X) = \mu \text{ and } SD(X) = \sigma.$$

Further, let $\overline{X}$ be the sample mean of a simple random sample of size n from the population.

Central limit theorem. The *central limit theorem* is a general theorem relating sampling distributions of sample means to normal distributions.

Specifically, if the sample size (n) is large, but small relative to the population size (N), then the sampling distribution of the sample mean of a simple random sample of size n from the population can be well-approximated by a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.

Example 6.2 (Educational Levels) Let X be the number of years an individual chosen at random from a population of 4,000 adults attended school, and assume that X takes whole number values between 8 and 20 with the following probability distribution:

x	8	9	10	11	12	13	14
$P(X = x)$	0.035	0.035	0.070	0.070	0.140	0.072	0.090
x	15	16	17	18	19	20	
$P(X = x)$	0.108	0.180	0.040	0.060	0.050	0.050	

Figure 6.2(a) (page 120) shows a probability histogram for X. For this random variable, the mean and standard deviation are

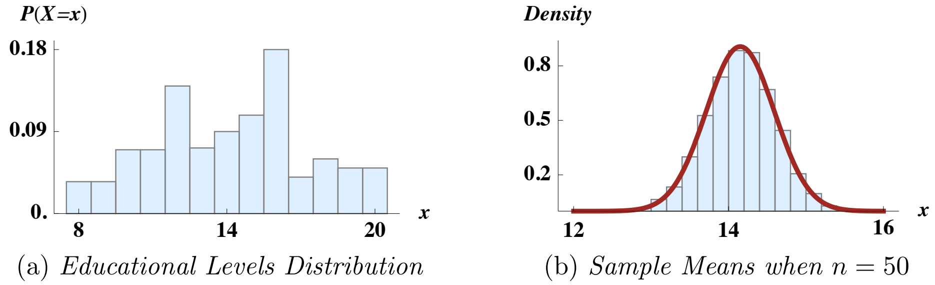
$$\mu = E(X) = 14.15 \text{ and } \sigma = SD(X) = 3.11.$$

An adult living in this population attended school an average of 14.15 years, with a standard deviation of 3.11 years.

Let $\overline{X}$ be the sample mean of a simple random sample of size 50 from this population. Figure 6.2(b) is a histogram of 10,000 simulated sample means. (The computer was used to generate a simple random sample of size 50 from the population and compute the sample mean a total of 10,000 times.) A computer simulation with 10,000 samples was used since it would be impossible to generate and summarize all samples of size 50 from the population; there are more than 3×10^{115} samples of size 50 from a population of size 4,000.

The normal curve with mean $\mu = 14.15$ and standard deviation $\sigma/\sqrt{n} = 0.440$ is superimposed on the histogram in Figure 6.2(b). The superimposed normal distribution is a close approximation to the histogram.

Figure 6.2: *Histograms of educational levels in an adult population and of mean educational levels in simple random samples of size 50 from the population. Part (a) is a probability histogram for X. Part (b) is a histogram of 10,000 simulated values of $\overline{X}$. The normal approximation curve with mean 14.15 years and standard deviation $3.11/\sqrt{50}$ = 0.440 years is superimposed on the histogram in part (b).*



(a) *Educational Levels Distribution*

(b) *Sample Means when* $n = 50$

Standard error of the sample mean. The *standard error of the sample mean*, denoted by $SE(\overline{X})$, is the standard deviation of the normal approximation to the sampling distribution of the sample mean $\overline{X}$:

$$SE(\overline{X}) = \sigma/\sqrt{n}.$$

For example, in the educational levels example above, $SE(\overline{X}) = 3.11/\sqrt{50} = 0.440$.

Computing approximate probabilities. Tables 5.1–5.2 (pages 109–110) can be used to find approximate probabilities for $\overline{X}$. For a given value $\overline{x}$, the standardized score is the ratio of the difference between $\overline{x}$ and μ to the standard error of the sample mean,

$$z = (\overline{x} - \mu)/(\sigma/\sqrt{n}).$$

Example 6.3 (Educational Levels, continued) To use the normal approximation to find the probability that the sample mean lies between 13.65 and 14.35, for example, we convert these values to standardized scores,

$$z_1 = (13.65 - 14.15)/0.440 = -1.14 \text{ and } z_2 = (14.35 - 14.15)/0.440 = 0.45,$$

and use the tables to compute the probability:

$$P(13.65 \leq \overline{X} \leq 14.35) = P(-1.14 \leq Z \leq 0.45) = 0.6736 - 0.1271 = 0.5465.$$

Thus, the mean educational level of individuals in a simple random sample of size 50 from the population is between 13.65 years and 14.15 years about 54.65% of the time.

6.2.2 Sampling Distribution of the Sample Proportion

Consider choosing a simple random sample of size n from a population with proportion p individuals in a subpopulation of interest. Let X be the number of individuals from the subpopulation of interest in the sample, and $\widehat{p}$ ("p-hat") be the proportion of individuals from the subpopulation of interest in the sample,

$$\widehat{p} = X/n.$$

The statistic $\widehat{p}$ is called the *sample proportion*.

Example 6.4 (Educational Levels, continued) Suppose, for example, that we are interested in the subpopulation of adults from the educational levels example who attended school for 12 or fewer years. From the distribution given on page 119, we know that 35% of individuals are members of this subpopulation.

Further, suppose that in a simple random sample of 50 from the population, exactly 15 attended school for 12 or fewer years. Then $p = 0.35$ and $\widehat{p} = 15/50 = 0.3$.

Approximate normal sampling distribution. If the sample size (n) is large, but small relative to the population size (N), and if both $np \geq 15$ and $n(1-p) \geq 15$, then the sampling distribution of the sample proportion of a simple random sample of size n from a population can be well-approximated by a normal distribution with mean p and standard deviation $\sqrt{p(1-p)/n}$.

This result is a special case of the central limit theorem mentioned earlier. The number of individuals from the subpopulation of interest in the sample (X) is a hypergeometric random variable whose distribution is well-approximated by a binomial distribution. The binomial distribution is, in turn, well-approximated by a normal distribution. Finally, when X is divided by n to form the sample proportion, the resulting distribution remains approximately normal with the mean and standard deviation given above.

Note that Figure 5.7 (page 106) illustrates how well a normal curve approximates the binomial distribution when $n = 120$ and $p = 0.28$.

Standard error of the sample proportion. The *standard error of the sample proportion*, denoted by $SE(\widehat{p})$, is the standard deviation of the normal approximation to the sampling distribution of $\widehat{p}$:

$$SE(\widehat{p}) = \sqrt{p(1-p)/n}.$$

Computing approximate probabilities. Tables 5.1–5.2 (pages 109–110) can be used to find approximate probabilities for $\widehat{p}$. For a given value of the statistic, the standardized score is the ratio of the difference between $\widehat{p}$ and p to the standard error of the sample proportion,

$$z = (\widehat{p} - p)/\sqrt{p(1-p)/n}.$$

Example 6.5 (Educational Levels, continued) Continuing with the educational levels example above, recall that 35% of individuals attended school for 12 or fewer years.

Let $\widehat{p}$ be the proportion of individuals who attended school for 12 or fewer years in a simple random sample of 50 individuals from the population. The sampling distribution of $\widehat{p}$ is approximately normal with mean $p = 0.35$ and standard deviation

$$\sqrt{p(1-p)/n} = \sqrt{(0.35)(0.65)/50} = 0.0675.$$

To use the normal approximation to find the probability that the sample proportion lies between 0.31 and 0.43, for example, we convert these values to standardized scores,

$$z_1 = (0.31 - 0.35)/0.0675 = -0.59 \text{ and } z_2 = (0.43 - 0.35)/0.0675 = 1.19,$$

and use the tables to compute the probability:

$$P(0.31 \leq \widehat{p} \leq 0.43) = P(-0.59 \leq Z \leq 1.19) = 0.8830 - 0.2776 = 0.6054.$$

Thus, the sample proportion of a simple random sample of size 50 from the population is between 0.31 and 0.43 about 60.54% of the time.

6.2.3 Sampling Distribution of the Sample Mean Rate

Assume that the occurrence of events follows the assumptions for the Poisson distribution (Section 5.3, page 101). Let λ be the average number of events in one unit of time or space, X the number of events observed in n units of time or space, and $\widehat{\lambda}$ ("lambda-hat") the sample mean number of events observed per unit time or space,

$$\widehat{\lambda} = X/n.$$

The statistic $\widehat{\lambda}$ is called the *sample mean rate* for the Poisson distribution.

Example 6.6 (Medical Hotline) Suppose that one of the services provided by a local community health center is a medical hotline and that, on average, $\lambda = 3$ calls are received per hour during the late-night shift (the 10pm to 6am shift).

Further, suppose that the numbers of calls received during the late-night shift on 5 consecutive nights were 19, 20, 22, 29 and 22, for a total of 112 calls over 40 hours of late-night service. Then $\widehat{\lambda} = 112/40 = 2.8$.

Approximate normal sampling distribution. If $n\lambda > 100$, then the sampling distribution of the sample mean rate for a Poisson distribution can be well-approximated by a normal distribution with mean λ and standard deviation $\sqrt{\lambda/n}$.

This result is a special case of the central limit theorem mentioned earlier. The number of events observed in n units of time or space has a Poisson distribution with mean $E(X) = n\lambda$ and the Poisson distribution is well-approximated by a normal distribution as long as its mean is more than 100. Finally, when X is divided by n to form the sample mean rate, the resulting sampling distribution remains approximately normal with the mean and standard deviation given above.

Note that Figure 5.8 (page 107) illustrates how well a normal curve approximates the Poisson distribution when the mean is 108.

Standard error of the sample mean rate. The *standard error of the sample mean rate* for the Poisson distribution, denoted by $SE(\widehat{\lambda})$, is the standard deviation of the normal approximation to the sampling distribution of $\widehat{\lambda}$:

$$SE(\widehat{\lambda}) = \sqrt{\lambda/n}.$$

Computing approximate probabilities. Tables 5.1–5.2 (pages 109–110) can be used to find approximate probabilities for $\widehat{\lambda}$. For a given value of the statistic, the standardized score is the ratio of the difference between $\widehat{\lambda}$ and λ to the standard error of the sample mean rate,

$$z = (\widehat{\lambda} - \lambda)/\sqrt{\lambda/n}.$$

Example 6.7 (Medical Hotline, continued) Continuing with the medical hotline example above, let $\widehat{\lambda}$ be the sample mean rate for a Poisson distribution over 40 hours of late-night service. The sampling distribution of $\widehat{\lambda}$ is approximately normal with mean $\lambda = 3$ and standard deviation

$$\sqrt{\lambda/n} = \sqrt{3/40} = 0.274.$$

To use the normal approximation to find the probability that the sample mean rate lies between 2.9 and 3.8, for example, we convert these values to standardized scores,

$$z_1 = (2.9 - 3)/0.274 = -0.36 \text{ and } z_2 = (3.8 - 3)/0.274 = 2.92,$$

and use the tables to compute the probability:

$$P(2.8 \leq \widehat{\lambda} \leq 3.9) = P(-0.36 \leq Z \leq 2.92) = 0.9982 - 0.3594 = 0.6388.$$

Thus, the sample mean rate averaged over 40 hours of late-night service lies between 2.9 calls per hour and 3.8 calls per hour about 63.88% of the time.

Footnotes. The sampling distributions introduced in this section are similar in that each is well-approximated by a normal distribution in large sample settings. The distributions for the sample mean and sample proportion are based on drawing simple random samples of size n from a population of size N, where N is assumed to be *much larger* than n. The exact value of N is not used, and is often not known exactly.

When working with the Poisson distribution, there is no mention of a population, although one is implied in most applications. For example, in the medical hotline example above, the implied population is anyone who might call with a medical question. Since the chance of a particular person calling is quite small, the use of a Poisson distribution is justified as an approximation to a binomial distribution based a large number of trials (one "trial" for each person who might call) and a small probability of success (the probability that a person actually does call).

In each case, the observed value of the statistic ($\overline{X}$, $\widehat{p}$ or $\widehat{\lambda}$) can be used to estimate the parameter of interest (μ, p or λ) in statistical applications. The mean of each random variable is the parameter of interest,

$$E(\overline{X}) = \mu, \;\; E(\widehat{p}) = p \text{ and } E(\widehat{\lambda}) = \lambda;$$

we say that each is an *unbiased estimator* of the parameter of interest.

Table 6.1: *Critical values for two-sided analyses when probability distributions are well-approximated by normal distributions.*

$100\alpha\%$:	50%	40%	30%	20%	10%	5%	1%	0.1%
$100(1-\alpha)\%$:	50%	60%	70%	80%	90%	95%	99%	99.9%
Critical Value, $z_{\alpha/2}$:	0.674	0.842	1.036	1.282	1.645	1.960	2.576	3.291

6.3 Critical Values and Central Intervals

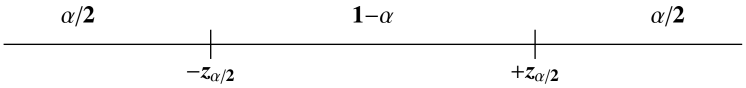
The Greek letter α ("alpha") is used to represent a small probability, and the notation

$$z_{\alpha/2} \quad (\text{“}z\text{-sub-alpha-over-2”})$$

is used to represent the position along the z-axis where the standard normal random variable Z satisfies the following probability statements:

$$P(Z < -z_{\alpha/2}) = \alpha/2,\ P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1-\alpha \text{ and } P(Z > z_{\alpha/2}) = \alpha/2.$$

These statements can be illustrated graphically as follows:



Using this notation, the central $100(1-\alpha)\%$ of the standard normal distribution lies in the interval $[-z_{\alpha/2},\ z_{\alpha/2}]$, and the remaining $100\alpha\%$ of the distribution lies outside the interval.

The numbers $z_{\alpha/2}$ are called the *two-sided critical values* (or *critical values*) for the standard normal distribution. Table 6.1 (page 124) lists critical values for commonly used values of α. Critical values are used in many statistical applications. For example, they are used to find the endpoints of the interval containing the central $100(1-\alpha)\%$ of a normal distribution.

Central intervals. Suppose that the probability distribution of X is well-approximated by a normal distribution with mean μ and standard deviation σ. The interval containing the central $100(1-\alpha)\%$ of the X distribution is

$$[\mu - z_{\alpha/2}(\sigma),\ \mu + z_{\alpha/2}(\sigma)].$$

Example 6.8 (Adult Heights, Revisited) Consider again the probability distributions of heights for adult men and women from the example on page 111. Each distribution is well-approximated by a normal distribution.

The interval containing the central 80% of the distribution of women's heights is

$$[65 - 1.282(3.5),\ 65 + 1.282(3.5)] \ \Rightarrow\ [65 - 4.487,\ 65 + 4.487] \ \Rightarrow\ [60.513,\ 69.487].$$

The mean height for women living in North America is 65 inches. The heights of 80% of the women living in North America are within about 4.5 inches of the mean; that is, the heights are between about 60.5 and 69.5 inches.

The interval containing the central 80% of the distribution of men's heights is

$$[70 - 1.282(4),\ 70 + 1.282(4)] \ \Rightarrow\ [70 - 5.128,\ 70 + 5.128] \ \Rightarrow\ [64.872,\ 75.128].$$

The mean height for men living in North America is 70 inches. The heights of 80% of the men living in North America are within about 5.1 inches of the mean; that is, the heights are between about 64.9 and 75.1 inches.

Footnote: large sample approximate central intervals. Since the sampling distributions of the sample mean, the sample proportion and the Poisson mean rate are well-approximated by normal distributions in large sample settings, a generalization of the formula above can be used to find large sample approximate central intervals. Specifically:

1. Let $\overline{X}$ be the sample mean of a simple random sample of size n from a population of interest. If the sample size is large, but small relative to the population size, then the interval containing the central $100(1-\alpha)\%$ of the $\overline{X}$ distribution is

$$[\mu - z_{\alpha/2}(\sigma/\sqrt{n}),\ \mu + z_{\alpha/2}(\sigma/\sqrt{n})].$$

2. Let $\widehat{p}$ be the sample proportion of a simple random sample of size n from a population with proportion p individuals in a subpopulation of interest. If the sample size is large, but small relative to the population size, and if both $np \geq 15$ and $n(1-p) \geq 15$, then the interval containing the central $100(1-\alpha)\%$ of the $\widehat{p}$ distribution is

$$\left[p - z_{\alpha/2}(\sqrt{p(1-p)/n}),\ p + z_{\alpha/2}(\sqrt{p(1-p)/n})\right].$$

3. Let $\widehat{\lambda}$ be the sample mean rate for the Poisson distribution based on observations taken over n units of time or regions of space, and let λ be the average number of events for one unit of time or region of space. If $n\lambda > 100$, then the interval containing the central $100(1-\alpha)\%$ of the $\widehat{\lambda}$ distribution is

$$\left[\lambda - z_{\alpha/2}(\sqrt{\lambda/n}),\ \lambda + z_{\alpha/2}(\sqrt{\lambda/n})\right].$$

6.4 Confidence Intervals

A *point estimate* (or *estimate*) is a single number that represents our best guess of a parameter of interest. For example, the observed value of the sample mean for a given sample of n individuals is a point estimate of a population mean μ.

By contrast, an *interval estimate* is an interval of numbers within which the parameter is believed to fall. Interval estimates are preferred to point estimates since they can be designed to incorporate information from the sampling distribution of the statistic used to estimate the parameter.

This section introduces methods for constructing interval estimates known as "confidence intervals" for population means, population proportions, and Poisson mean rates when sample sizes are large.

6.4.1 Introductory Concepts

A *confidence interval* is an interval estimate of an unknown parameter together with a fixed quantity known as the confidence level of the interval. The *confidence level* is the probability that the interval contains the parameter in repeated samples.

The Greek letter α ("alpha") is used to denote the probability that a confidence interval does *not* contain the parameter in repeated samples. Thus,

$$\alpha = 1 \; - \; \text{Confidence Level.}$$

Equivalently, the confidence level is $(1 - \alpha)$ for some α. A confidence interval with confidence level $(1 - \alpha)$ is often called a "$100(1 - \alpha)\%$ confidence interval."

Common choices for confidence levels are

$$(1 - \alpha) = 0.50, 0.60, 0.70, 0.80, 0.90, 0.95, 0.99, 0.999.$$

6.4.2 Large Sample Confidence Interval Methods

Formulas for constructing large sample approximate confidence intervals for population means, population proportions and Poisson mean rates are given below:

1. ***Population Mean:*** If $\overline{x}$ is the sample mean and s is the sample standard deviation of a simple random sample of size n from the population of interest, then an approximate $100(1 - \alpha)\%$ confidence interval for the population mean μ is the interval

   $$\overline{x} \;\pm\; z_{\alpha/2}\;(s/\sqrt{n}).$$

   The approximation can be used when the sample size is large, but small relative to the population size.

2. ***Population Proportion:*** If $\widehat{p}$ is the sample proportion of a simple random sample of size n from the population, then an approximate $100(1 - \alpha)\%$ confidence interval for the population proportion p is the interval

   $$\widehat{p} \;\pm\; z_{\alpha/2}\;\sqrt{\widehat{p}(1 - \widehat{p})/n}.$$

   The approximation can be used when the sample size is large, but small relative to the population size, and both $np \geq 15$ and $n(1 - p) \geq 15$.

3. ***Poisson Mean Rate:*** If $\widehat{\lambda}$ is the sample mean rate for a Poisson distribution observed over n units of time or space, then an approximate $100(1-\alpha)\%$ confidence interval for λ is the interval
$$\widehat{\lambda} \;\pm\; z_{\alpha/2}\,\sqrt{\widehat{\lambda}/n}.$$
The approximation can be used when $n\lambda > 100$.

Note that each formula is written as "Estimate ± Error," where

$$\text{Error} = (\text{Critical Value})(\text{Estimated Standard Error});$$

once the "Error" is computed, the answer is "[Estimate − Error, Estimate + Error]." In addition, note that an estimated standard error is used in each formula since the true standard error of the statistic used to estimate the parameter of interest is not known.

For example, suppose that the sample mean of a simple random sample of size 85 from a large population is 28.73, with a standard deviation of 2.25, and that we are interested in estimating the population mean with 90% confidence.

Since the estimated standard error of the sample mean is $2.25/\sqrt{85} = 0.244$, and the critical value for 90% confidence is 1.645 (from Table 6.1, page 124), the approximate 90% confidence interval is
$$28.73 \pm (1.645)(0.244) \;\Rightarrow\; 28.73 \pm 0.401 \;\Rightarrow\; [28.329, 29.131].$$
With 90% confidence, we believe that the population mean is between about 28.3 and 29.1.

If, instead, we are interested in estimating the population mean with 99% confidence, then the critical value becomes 2.576 (from Table 6.1, page 124) and the 99% confidence interval is
$$28.73 \pm (2.576)(0.244) \;\Rightarrow\; 28.73 \pm 0.629 \;\Rightarrow\; [28.101, 29.359].$$
With 99% confidence, we believe that the population mean is between about 28.1 and 29.4.

Note that the 99% confidence interval is wider than the 90% confidence interval. In general, as the confidence level increases, so does the width of the corresponding interval.

Example 6.9 (Mean Systolic Blood Pressure) (Pagano & Gauvreau, 2000, page 230) As part of a study on labor and delivery characteristics of low birthweight infants, researchers gathered information on a simple random sample of 100 low birthweight babies born in two teaching hospitals in Boston.

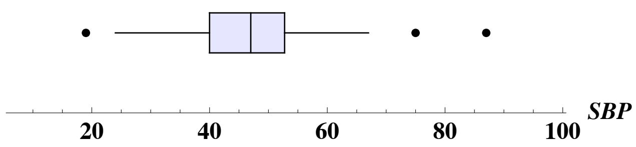
The following table gives the systolic blood pressures of these newborns, arranged in increasing order. The units are millimeters of mercury (mmHg).

Systolic Blood Pressures ($n = 100$)

19	24	25	26	27	29	29	30	31	31	34	35	35	36	36	36	36	37	39	39
39	40	40	40	40	40	42	42	42	42	42	43	43	43	43	44	44	44	44	44
44	45	45	45	45	45	46	46	46	47	47	47	47	48	48	48	48	48	48	48
48	49	49	49	50	50	50	51	51	51	51	51	52	52	52	53	53	53	54	54
56	57	58	59	59	61	62	62	62	63	63	64	64	64	64	66	66	67	75	87

Numerical and graphical summaries of these data are given in Figure 6.3 (page 128).

Figure 6.3: *Summaries of systolic blood pressures for 100 low birthweight infants.*



Q1: 40.0 mmHg	*Median:* 47.0 mmHg	*Mean:* 47.08 mmHg
Q2: 47.0 mmHg	*IQR:* 12.75 mmHg	*SD:* 11.403 mmHg
Q3: 52.75 mmHg	*Outliers:* 19, 75, 87	*Sample Size:* 100

To estimate the mean systolic blood pressure at birth for all low birthweight infants born at these two hospitals with 95% confidence, for example, we use the formula:

$$47.08 \pm (1.960)(11.403/\sqrt{100}) \;\Rightarrow\; 47.08 \pm 2.235 \;\Rightarrow\; [44.845, 49.315].$$

With 95% confidence, we believe that the mean systolic blood pressure for this population of low birthweight babies is between about 44.8 and 49.3 mmHg.

Example 6.10 (Proportion with Toxemia) (Pagano & Gauvreau, 2000, page 340) Continuing with the study on labor and delivery characteristics of low birthweight infants from the last example, researchers also gathered information on whether or not the mother of the low birthweight child was diagnosed with toxemia during pregnancy. Toxemia is a condition characterized by high blood pressure and other potentially serious complications.

The researchers found that 21% (21/100) of mothers in the sample were diagnosed with toxemia during pregnancy. To estimate the proportion of all mothers of low birthweight infants born at these two teaching hospitals who are diagnosed with toxemia during pregnancy with 95% confidence, for example, we use the formula:

$$0.21 \pm (1.960)\sqrt{(0.21)(0.79)/100} \;\Rightarrow\; 0.21 \pm 0.0798 \;\Rightarrow\; [0.1302, 0.2898].$$

With 95% confidence, we believe that between 13.02% and 28.98% of mothers of low birthweight babies are diagnosed with toxemia during pregnancy.

Example 6.11 (Mean Rate of Asbestos Fibers) (Rice, 1995, page 247) As part of a study done at the National Institute of Science and Technology (NIST), researchers counted the number of asbestos fibers on filters. Their goal was to establish standards for measuring asbestos concentration.

By using an electron microscope, the researchers found the following numbers of asbestos fibers on 23 small grid squares chosen at random from a large filter:

$$16, 17, 18, 18, 18, 19, 21, 22, 24, 24, 24, 26, 27, 27, 27, 28, 28, 29, 30, 31, 31, 34, 34.$$

The numbers are listed in increasing order. The sample mean rate is $\widehat{\lambda} = 573/23 = 24.913$ fibers per small grid square (fibers/sq).

The Poisson distribution would be a reasonable model for describing the variability of counts from grid square to grid square in this situation. To estimate the mean rate for the Poisson distribution with 95% confidence, for example, we use the formula:

$$24.913 \pm (1.960)\sqrt{24.913/23} \;\Rightarrow\; 24.913 \pm 2.040 \;\Rightarrow\; [22.873, 26.953].$$

With 95% confidence, we believe that the mean concentration of asbestos fibers per small grid square is between about 22.9 and 27.0 fibers/sq.

Footnotes. The confidence interval formulas introduced in this section are examples of *two-sided confidence intervals*. In each case, the interval we compute has both a lower limit ("Estimate − Error") and an upper limit ("Estimate + Error").

In some situations, we may be interested in finding an upper limit only at a certain confidence level or in finding a lower limit only at a certain confidence level. Intervals constructed with a single limit only (either upper or lower) are called *one-sided confidence intervals*.

Two-sided confidence intervals are used more often than one-sided intervals. In fact, the term "confidence interval" refers to a two-sided confidence interval unless there is an explicit statement to the contrary.

6.4.3 Interpretation of Confidence Level

People often mistake the confidence level with the probability that the parameter of interest lies in the computed interval. For example, in the toxemia example above, either the proportion of mothers of low birthweight babies diagnosed with toxemia is between 0.1302 and 0.2898 or it is not. There is no probability associated with the computed interval. Instead, probability is associated with the procedure used to choose the sample. If we were to repeatedly choose simple random samples of size 100 from the study population, and compute 95% confidence intervals for the proportion of mothers diagnosed with toxemia from each sample, then we can expect that the true proportion lies in about 95% of the computed intervals.

Example 6.12 (Computer Simulation) A good way to help understand the concept of confidence level is to examine the results of a computer simulation.

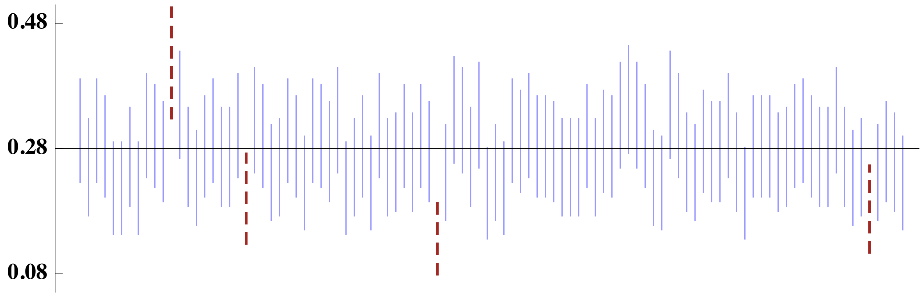
The computer was used to generate 100 simple random samples of size 120 from a large population, where 28% of individuals have a characteristic of interest. For each sample, a sample proportion $\widehat{p}$ was computed. The sample proportions ranged from 0.14 to 0.42 with the following distribution:

$[0, 0.20)$	$[0.20, 0.25)$	$[0.25, 0.30)$	$[0.30, 0.35)$	$[0.35, 0.40)$	$[0.40, 1.00)$	Total
2	20	45	29	3	1	100

Next, a 95% confidence interval for the proportion of individuals with the characteristic of interest was computed for each sample. In 96 of 100 cases, the interval contained 0.28.

Figure 6.4 (page 130) illustrates the results. Each confidence interval is represented as a vertical line segment. Solid segments represent intervals containing 0.28; dashed segments are used otherwise. There are 96 solid segments and 4 dashed segments in the figure.

Figure 6.4: *Plot of 100 simulated 95% confidence intervals for $p = 0.28$, using simple random samples of size $n = 120$ from a large population. Each interval is represented as a vertical line segment from its lower endpoint $(\widehat{p} - 1.96\sqrt{\widehat{p}(1-\widehat{p})/n})$ to its upper endpoint $(\widehat{p} + 1.96\sqrt{\widehat{p}(1-\widehat{p})/n})$. Intervals containing the population proportion are represented as solid line segments; the remaining are represented as dashed line segments. In this simulation, 96% (96/100) of intervals contain the population proportion.*



6.4.4 Sample Size Computations

Researchers are often interested in estimating a parameter to within a certain margin of error. The *margin of error* is the value of

$$\text{Error} = (\text{Critical Value})(\text{Estimated Standard Error})$$

in the formula for the confidence interval: "Estimate ± Error."

The value of the error term in the confidence interval formula depends on both the chosen confidence level and the estimated standard error. The estimated standard error depends, in turn, on the sample size. Thus, once a margin of error is chosen for a particular confidence level, the researcher can then determine an appropriate sample size n for the study.

Sample size for a population proportion. The error term in the confidence interval formula for a population proportion is

$$\text{Error} = (\text{Critical Value})(\sqrt{\widehat{p}(1-\widehat{p})/n}) = (\text{Critical Value})(\sqrt{\widehat{p}(1-\widehat{p})})/(\sqrt{n}).$$

The term $\sqrt{\widehat{p}(1-\widehat{p})}$ is never more than 0.5, and it is common practice to substitute 0.5 in the formula when computing a sample size.

Let m be the margin of error we would like in a particular study. Then

$$m = (\text{Critical Value})(0.5)/\sqrt{n} \ \Rightarrow\ n = \Big((\text{Critical Value})(0.5)/m\Big)^2.$$

The formula on the right is used to compute the sample size.

For example, suppose that we would like to estimate the percentage of likely voters in Massachusetts who support a given ballot proposal to within 4 percentage points, with 99%

confidence. Then the margin of error is $m = 0.04$, the critical value is 2.576 (from Table 6.1, page 124), and the sample size is

$$n = ((2.576)(0.5)/0.04)^2 = (32.2)^2 = 1036.84 \nearrow 1037.$$

Thus, a simple random sample of size 1,037 (or more) from the population of likely voters will be needed to estimate the level of support for the ballot initiative to within 4 percentage points with 99% confidence.

Sample size for a population mean. The error term in the confidence interval formula for a population mean is

$$\text{Error} = (\text{Critical Value})(s/\sqrt{n}) = (\text{Critical Value})(s)/(\sqrt{n}).$$

Since the value of s (the estimated population standard deviation) is unknown, it is common practice to substitute an educated guess in the formula when computing sample sizes. A safe approach is to use a number larger than estimated population standard deviations from previous similar studies.

Let m be the margin of error and let s_g be our educated guess for the population standard deviation. Then

$$m = (\text{Critical Value})(s_g)/\sqrt{n} \;\Rightarrow\; n = \Big((\text{Critical Value})(s_g)/m\Big)^2.$$

The formula on the right is used to compute the sample size.

For example, suppose that we would like to estimate the mean serum cholesterol level in milligrams per deciliter (mg/dL) of adult men living in the United States to within 4 mg/dL with 95% confidence, and that we are willing to use 50 mg/dL as our educated guess for the population standard deviation. Then the margin of error is $m = 4$, the critical value is 1.960 (from Table 6.1, page 124), the estimated population standard deviation is $s_g = 50$ and the sample size is

$$n = ((1.960)(50)/4)^2 = (24.5)^2 = 600.25 \nearrow 601.$$

Thus, a simple random sample of size 601 (or more) from the population of adult men living in the United States will be needed to estimate the mean serum cholesterol level to within 4 mg/dL with 95% confidence.

Sample size for a Poisson mean rate. The error term in the confidence interval formula for a Poisson mean rate is

$$\text{Error} = (\text{Critical Value})\left(\sqrt{\widehat{\lambda}/n}\right) = (\text{Critical Value})\left(\sqrt{\widehat{\lambda}}\right)/\sqrt{n}.$$

Since the value of $\widehat{\lambda}$ is unknown, it is common practice to substitute an educated guess in the formula when computing sample sizes. A safe approach is to use a number larger than estimates of the Poisson mean rate from previous similar studies.

Let m be the margin of error and let $\widehat{\lambda}_g$ be our educated guess for the Poisson mean rate. Then

$$m = (\text{Critical Value})\left(\sqrt{\widehat{\lambda}_g}\right)/\sqrt{n} \;\Rightarrow\; n = \left((\text{Critical Value})\left(\sqrt{\widehat{\lambda}_g}\right)/m\right)^2.$$

The formula on the right is used to compute the sample size.

For example, suppose that one of the services provided by a local community health center is a medical hotline. In addition, suppose that we would like to estimate the mean hourly rate of phone calls to the hotline during the late-night shift (the 10pm to 6am shift) to within 0.5 calls/hr with 95% confidence, and that we are willing to use 9 calls/hr as our educated guess for the unknown Poisson mean rate. Then the margin of error is $m = 0.5$, the critical value is 1.960 (from Table 6.1, page 124), the estimated sample mean rate is $\widehat{\lambda}_g = 9$ and the sample size is

$$n = \left((1.960)(\sqrt{9})/(0.5)\right)^2 = (11.76)^2 = 138.298 \nearrow 139.$$

Thus, we will need to count the number of calls received by the medical hotline for 139 late-night hours (or more) in order to estimate the Poisson mean rate to within 0.5 calls/hr with 95% confidence.

6.5 Hypothesis Tests

An *hypothesis* is a statement about a population, often written in the form that a parameter takes a particular value or falls in a certain range of values.

This section introduces methods for conducting "two-sided hypothesis tests" for population means, population proportions, and Poisson mean rates when sample sizes are large. The methods are parallel to the "two-sided confidence interval" methods of the last section.

6.5.1 Introductory Concepts

In hypothesis testing, we first determine competing hypotheses, known as the

1. *Null hypothesis*, H_0 ("H-naught"), and the
2. *Alternative hypothesis*, H_A ("H-A").

The null hypothesis often represents no difference or no effect, while the alternative hypothesis represents a difference or effect of some kind.

For example, suppose that the mean serum cholesterol level for adult men living in the United States is known to be 211 milligrams per deciliter (mg/dL) and that we are interested in testing whether the mean (μ) for the subpopulation of hypertensive male smokers is the same as the mean for the entire population or not. Then, our null and alternative hypotheses would be set up as follows:

$$H_0 : \mu = 211 \quad \text{versus} \quad H_A : \mu \neq 211.$$

Once null and alternative hypotheses are stated, an hypothesis test can be performed to determine if differences observed in sample information are due to chance alone (supporting the null hypothesis) or due to something else (supporting the alternative hypothesis).

Test statistic. The *test statistic* used in an hypothesis test measures the distance between the point estimate of the parameter of interest and its null hypothesis value. In many cases, the measure is the number of standard errors between the values.

If the sampling distribution of the statistic used to estimate the parameter of interest is well-approximated by a normal distribution, then a *z-statistic* can be used. For example, in the serum cholesterol scenario above, the z-statistic is

$$z = (\overline{x} - 211)/(s/\sqrt{n}),$$

where $\overline{x}$ is the sample mean and s is the sample standard deviation in a simple random sample of size n from the population of adult men who are hypertensive smokers, and 211 is the null hypothesis value of the population mean. The use of the z-statistic is justified if the sample size is large.

Acceptance and rejection regions; significance level. The *acceptance region* associated with an hypothesis test is the range of values of the test statistic that would lead us to conclude that we believe observed differences are due to chance alone, and the *rejection region* is the complementary range of values.

The *significance level*, α, is the probability that the test statistic falls in the rejection region when the null hypothesis is true. Common choices for significance levels are

$$\alpha = 0.20, 0.10, 0.05, 0.01, 0.001.$$

If the significance level is α, we often say that we have a "$100\alpha\%$ test" or that we are conducting the test at the "$100\alpha\%$ significance level."

For example, in the serum cholesterol scenario above, the acceptance and rejection regions for a $100\alpha\%$ test of $\mu = 211$ mg/dL versus $\mu \neq 211$ mg/dL are

Acceptance Region:	Rejection Region:
$\lvert z\rvert < z_{\alpha/2}$	$\lvert z\rvert \geq z_{\alpha/2}$

where $z = (\overline{x} - 211)/(s/\sqrt{n})$ and $z_{\alpha/2}$ is the critical value for two-sided analyses from Table 6.1 (page 124). If the population mean is 211 mg/dL and n is large, then the probability that the value of the test statistic falls in the rejection region is α.

Statistical significance. If the value of the test statistic falls in the rejection region for a $100\alpha\%$ test, then we say that the result is *statistically significant at the* $100\alpha\%$ *level*.

If the test result is statistically significant at the $100\alpha\%$ level, then our conclusion is that we believe the observed difference did not occur by chance. Otherwise, we believe the observed difference did occur by chance.

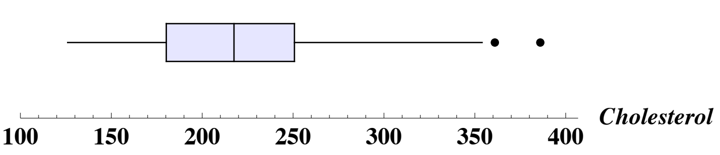
6.5.2 Large Sample Test of a Population Mean

Suppose that we are interested in testing the null hypothesis that a population mean equals a fixed known constant μ_0 ("mu-naught") versus the alternative hypothesis that the mean is not equal to μ_0,

$$H_0\colon\ \mu = \mu_0 \text{ versus } H_A\colon\ \mu \neq \mu_0,$$

using information from a simple random sample of size n from the population.

Figure 6.5: *Summaries of serum cholesterol levels for 140 hypertensive male smokers.*



Q1: 180.25 mg/dL	*Median:* 217.5 mg/dL	*Mean:* 222.586 mg/dL
Q2: 217.5 mg/dL	*IQR:* 70.5 mg/dL	*SD:* 52.347 mg/dL
Q3: 250.75 mg/dL	*Outliers:* 361, 386	*Sample Size:* 140

If n is large, but small compared to the population size, then the test statistic is

$$z = (\overline{x} - \mu_0)/(s/\sqrt{n}),$$

where $\overline{x}$ is the sample mean and s is the sample standard deviation, and the rejection region for a two-sided test conducted at the $100\alpha\%$ significance level is $|z| \geq z_{\alpha/2}$.

Example 6.13 (Mean Serum Cholesterol Level) Continuing with the serum cholesterol scenario above, suppose that the researchers were interested in testing

$$H_0\colon\ \mu = 211 \text{ versus } H_A\colon\ \mu \neq 211\ ,$$

at the 1% significance level, using information from a simple random sample of 140 men from the subpopulation of hypertensive male smokers. The rejection region for a two-sided test conducted at the 1% significance level is $|z| \geq 2.576$ (from Table 6.1, page 124).

Assume that the following table lists the serum cholesterol levels for these men:

Serum Cholesterol Levels ($n = 140$)*:*

126	132	138	141	147	148	153	154	154	156	157	160	160	160	163	163
165	167	169	169	169	170	170	171	173	175	176	177	177	177	178	180
180	180	180	181	183	184	184	186	189	191	193	193	194	198	199	199
200	200	201	201	202	202	203	203	204	205	206	207	207	209	209	211
212	214	216	216	216	217	218	219	220	220	221	222	222	223	223	224
225	229	229	230	230	231	231	231	231	233	234	237	237	239	239	240
240	241	241	241	243	244	245	248	250	251	251	252	254	254	255	258
261	262	265	267	268	272	273	278	278	280	281	288	289	292	292	293
299	321	324	325	327	330	340	348	352	354	361	386				

Numerical and graphical summaries of these data are given in Figure 6.5 (page 134). Using information in the figure, the observed value of the test statistic is

$$z = (222.586 - 211)/(52.347/\sqrt{140}) = 2.619.$$

Since the observed value of z lies in the rejection region, the test result is statistically significant at the 1% level. Thus, we believe that the observed difference did not occur by chance.

Further, we have reason to believe that the population mean is actually greater than 211 milligrams per deciliter.

Footnote. It is common practice to include a $100(1-\alpha)\%$ confidence interval for the population mean with the results from an hypothesis test about the mean. For example, in the mean serum cholesterol example, a 99% confidence interval for the population mean is

$$222.586 \pm (2.576)(52.347/\sqrt{140}) \;\Rightarrow\; 222.586 \pm 11.397 \;\Rightarrow\; [211.189, 233.983].$$

Since 211 is less than the lower limit of the confidence interval, we have reason to believe that the mean serum cholesterol level for hypertensive men is greater than 211 mg/dL. Note, however, that the lower limit of the confidence interval is only slightly greater than 211.

6.5.3 Large Sample Test of a Population Proportion

Suppose that we are interested in testing the null hypothesis that a population proportion equals a fixed known constant p_0 ("p-naught") versus the alternative hypothesis that the proportion is not equal to p_0,

$$H_0\colon\ p = p_0 \text{ versus } H_A\colon\ p \neq p_0,$$

using information from a simple random sample of size n from the population.

If n is large, but small compared to the population size, $np_0 \geq 15$ and $n(1-p_0) \geq 15$, then the test statistic is

$$z = (\widehat{p} - p_0)/\sqrt{p_0(1-p_0)/n},$$

where $\widehat{p}$ is the sample proportion, and the rejection region for a two-sided test conducted at the $100\alpha\%$ significance level is $|z| \geq z_{\alpha/2}$.

Example 6.14 (Gender Bias) (Agresti & Franklin, 2007, page 386) A women's group associated with a large supermarket chain in Florida claimed that female employees were passed over for management training in favor of their male colleagues.

The company denied their claim, saying that they picked the employees from the eligible pool at random to receive this training. Statewide, the large pool of more than 1000 eligible employees is 40% female and 60% male. Since the program began, a total of 40 employees were chosen for management training; 12 were women and 28 were men.

In order to test the company's claim of a lack of gender bias in choosing employees for management training, the women's group decided to test

$$H_0\colon\ p = 0.40 \text{ versus } H_A\colon\ p \neq 0.40$$

at the 5% significance level, where p is the proportion of women in the pool of eligible employees, using information on the 40 employees chosen for management training since the program began. The rejection region for a two-sided test conducted at the 5% significance level is $|z| \geq 1.960$ (from Table 6.1, page 124).

For these data, $\widehat{p} = 12/40 = 0.30$ and the observed value of the test statistic is

$$z = (0.30 - 0.40)/\sqrt{(0.40)(0.60)/40} = -1.291.$$

Since the observed value of z does not fall in the rejection region, the test result is not statistically significant at the 5% level. Although the proportion of women chosen for management training (0.30) was less than the proportion of women eligible for training (0.40), a difference of this size could have occurred by chance.

Footnotes. The z-statistic used to test $p = p_0$ versus $p \neq p_0$ in large sample settings uses the standard error of the sample proportion under the null hypothesis,

$$SE(\widehat{p}) = \sqrt{p_0(1-p_0)/n},$$

since the test is conducted assuming that the population proportion is p_0. By contrast, an estimate of the standard error is used when constructing confidence intervals for population proportions since the true proportion is assumed to be unknown.

It is common practice to include a $100(1-\alpha)\%$ confidence interval for the population proportion with the results of an hypothesis test about the proportion. For example, in the gender bias example above, a 95% confidence interval for the value of p is

$$0.30 \pm (1.960)\sqrt{(0.30)(0.70)/40} \;\Rightarrow\; 0.30 \pm 0.142 \;\Rightarrow\; [0.158, 0.442].$$

Since 0.40 is in the 95% confidence interval, we have no reason to doubt that the sample of employees chosen for management training was a simple random sample from a population of eligible employees with 40% women and 60% men.

6.5.4 Large Sample Test of a Poisson Mean Rate

Suppose that we are interested in testing the null hypothesis that a Poisson mean rate equals a fixed known constant λ_0 ("lambda-naught") versus the alternative hypothesis that the rate is not equal to λ_0,

$$H_0\text{: } \lambda = \lambda_0 \text{ versus } H_A\text{: } \lambda \neq \lambda_0,$$

using information gathered over n intervals of time or regions of space.

If $n\lambda_0 > 100$, then the test statistic is

$$z = (\widehat{\lambda} - \lambda_0)/\sqrt{\lambda_0/n},$$

where $\widehat{\lambda}$ is the sample mean rate, and the rejection region for a two-sided test conducted at the $100\alpha\%$ significance level is $|z| \geq z_{\alpha/2}$.

Example 6.15 (Suicide Rates) The National Center for Health Statistics (NCHS) routinely publishes information about suicide rates in the United States.

Suppose that a recent NCHS publication reported that the suicide rate is 15.6 suicides per hundred thousand population per year, but gave no information about regional differences in rates. After reading the report, a health care professional living in a rural community decided to consult other government reports to determine if the suicide rate among individuals in the rural subpopulation of the country is the same as the rate for the country as a whole. The researcher found that there were 294 suicides among the 2.2 million individuals living in rural communities last year.

In order to determine if the yearly suicide rate for the rural subpopulation is the same as for the country as a whole, the researcher decided to test

$$H_0\colon\ \lambda = 15.6 \text{ versus } H_A\colon\ \lambda \neq 15.6,$$

at the 5% significance level. The rejection region for a two-sided test conducted at the 5% significance level is $|z| \geq 1.960$ (from Table 6.1, page 124).

Since the 2.2 million individuals can be thought of as 22 groups of 100,00, the sample mean rate is $294/22 = 13.364$ suicides per hundred thousand population in the last year. The observed value of the test statistic is

$$z = (13.364 - 15.6)/\sqrt{15.6/22} = -2.655.$$

Since the observed value of z lies in the rejection region, the test result is statistically significant at the 5% level. Thus, we believe that the observed difference between did not occur by chance. Further, we have reason to believe that the population mean suicide rate is actually smaller than 15.6 per hundred thousand per year.

Footnotes. The analysis done in the last example can be justified by the fact that the Poisson distribution is a reasonable model for describing the variability of numbers of suicides in each of 22 groups of 100,000 individuals each. Since the researcher was working with approximate information from government reports only, the results should be considered rough approximations only.

The z-statistic used to test $\lambda = \lambda_0$ versus $\lambda \neq \lambda_0$ in large sample settings uses the standard error of the sample mean rate under the null hypothesis,

$$SE(\widehat{\lambda}) = \sqrt{\lambda_0/n},$$

since the test is conducted assuming that the Poisson mean rate is λ_0. By contrast, an estimate of the standard error is used when constructing confidence intervals for Poisson mean rates since the true mean rate is assumed to be unknown.

It is common practice to include a $100(1-\alpha)\%$ confidence interval for the Poisson mean rate with the results of an hypothesis test about the rate. For example, in the suicide rate example above, a 95% confidence interval for the value of λ is

$$13.364 \pm (1.960)\sqrt{13.364/22} \ \Rightarrow\ 13.364 \pm 1.528 \ \Rightarrow\ [11.836, 14.892].$$

Since 15.6 is above the upper limit of the 95% confidence interval, we have reason to believe that the suicide rate in the rural subpopulation is less than 15.6 per 100,000 per year.

6.5.5 Interpretation of Significance Level

People often mistake the significance level with the probability that the null hypothesis is false. In fact, the significance level represents an error rate when the null hypothesis is true.

Specifically, we expect to *correctly* conclude that an observed difference is due to chance alone $100(1-\alpha)\%$ of the time when sampling is done from a population satisfying the null hypothesis and to *incorrectly* conclude that the difference is due to something other than chance $100\alpha\%$ of the time. The probability of making a mistake is the significance level, α.

Example 6.16 (Computer Simulation) A good way to help understand the concept of significance level is to examine the results of a computer simulation.

The computer was used to generate 500 simple random samples of size 120 from a large population, where 28% of individuals have a characteristic of interest. For each sample, a sample proportion $\widehat{p}$ was computed. The sample proportions ranged from 0.175 to 0.408 with the following distribution:

$[0, 0.20)$	$[0.20, 0.25)$	$[0.25, 0.30)$	$[0.30, 0.35)$	$[0.35, 0.40)$	$[0.40, 1.00)$	Total
12	91	224	144	28	1	500

Next, a test of $p = 0.28$ versus $p \neq 0.28$ was conducted at the 5% significance level for each sample. The rejection region for each test was $|z| \geq 1.960$. Values of z-statistics ranged from -2.562 to 3.131 with the following distribution:

$z \leq -1.960$	$-1.960 < z < 1.960$	$z \geq 1.960$	Total
12	474	14	500

Observed differences were correctly judged to be due to chance in 94.8% (474/500) of the simulated samples. Observed differences were incorrectly judged to be due to something other than chance in 5.2% (26/500) of the samples.

6.5.6 Observed Significance Level

The *observed significance level* (or *p-value*) is the probability that the test statistic equals the observed value or a value even more extreme. It is calculated assuming that the null hypothesis is true.

If the observed significance level is less than or equal to α, then the result of the test is statistically significant at the $100\alpha\%$ level. Otherwise, the result is not statistically significant at the $100\alpha\%$ level.

For two-sided z-tests, the observed significance level is

$$\text{Observed Significance Level} = P(|Z| \geq |z|).$$

Equivalently, the observed significance level is

$$\text{Observed Significance Level} = \begin{cases} 2P(Z \leq z) & \text{when } z < 0, \\ 2(1 - P(Z \leq z)) & \text{when } z \geq 0. \end{cases}$$

Example 6.17 (Mean Serum Cholesterol Level, continued) In the mean serum cholesterol example (beginning on page 134), the test was conducted at the 1% significance level and the observed value of z was 2.62 (using 2 decimal places of accuracy). The observed significance level for this test is

$$2(1 - P(Z \leq 2.62)) = 2(1 - 0.9956) = 0.0088$$

(using Table 5.2, page 110). Since the observed significance level is less than 0.01, the result is statistically significant at the 1% level.

Example 6.18 (Gender Bias, continued) In the gender bias example (beginning on page 135), the test was conducted at the 5% level and the observed value of z was -1.29 (using 2 decimal places of accuracy). The observed significance level for this test is

$$2P(Z \leq -1.29) = 2(0.0985) = 0.1970$$

(using Table 5.1, page 109). Since the observed significance level is greater than 0.05, the result is not statistically significant at the 5% level.

Example 6.19 (Suicide Rates, continued) In the suicide rates example (beginning on page 136), the test was conducted at the 5% level and the observed value of z was -2.66 (using 2 decimal places of accuracy). The observed significance level for this test is

$$2P(Z \leq -2.66) = 2(0.0039) = 0.0078$$

(using Table 5.1, page 109). Since the observed significance level is less than 0.05, the result is statistically significant at the 5% level.

Footnote: computer generated reports. It is common practice for statistical programs to include observed significance levels (p-values) in reports on the results of hypothesis tests, but to exclude explicit statements about whether the results are statistically significant at a given significance level. It is also common practice to include a confidence interval in the report. For example, a report for the gender bias example might look like the following:

Large sample test of $p = 0.40$ versus $p \neq 0.40$:

x	n	$\widehat{p}$	95% CI	z-Statistic	p-Value
12	40	0.30	[0.158,0.442]	-1.291	0.197

Since the observed significance level is greater than 0.05, the results are not statistically significant at the 5% level. Although the observed population proportion was less than the hypothesized proportion, a difference of this size could have occurred by chance. This conclusion is confirmed by examining the 95% confidence interval for the proportion. Since 0.40 is in this interval, we cannot rule out the possibility that the true proportion is 0.40.

6.5.7 Errors of Types I and II

If the null hypothesis is true but we incorrectly conclude that observed differences are due to something other than chance, then we have made an *error of type I*. If the null hypothesis is false but we incorrectly conclude that observed differences are due to chance, then we have made an *error of type II*.

In two-sided tests of a population parameter, the probability of making an error of type I is the same as the significance level,

$$P(\text{Error of Type I}) = \text{Significance Level} = \alpha.$$

The probability of making an error of type II depends on the true population parameter. The Greek letter β ("beta") is used to denote the probability of making an error of type II at a specific value of the population parameter.

Errors of type II when testing $p = p_0$ versus $p \neq p_0$. Suppose that we are interested in testing the null hypothesis $p = p_0$ versus the alternative $p \neq p_0$ at the $100\alpha\%$ significance level using information from a simple random sample of size n from a large population, and suppose that the true population proportion is p_1.

The acceptance region for the test is $|z| < z_{\alpha/2}$, which we rewrite as follows:

$$\left|\frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}\right| < z_{\alpha/2} \;\Rightarrow\; \frac{|\widehat{p} - p_0|}{\sqrt{p_0(1-p_0)/n}} < z_{\alpha/2} \;\Rightarrow\; |\widehat{p} - p_0| < z_{\alpha/2}\sqrt{p_0(1-p_0)/n},$$

where $\widehat{p}$ is the sample proportion.

If we let $\text{error}_0 = z_{\alpha/2}\sqrt{p_0(1-p_0)/n}$, then we can rewrite the acceptance region as follows:

$$|\widehat{p} - p_0| < \text{error}_0 \;\Rightarrow\; p_0 - \text{error}_0 < \widehat{p} < p_0 + \text{error}_0.$$

The probability of making a type II error, β, is the probability that the z-statistic falls in the acceptance region when the true population proportion is p_1. Using the work above,

$$\beta = P(z_1 < Z < z_2), \text{ where } z_1 = \frac{((p_0 - \text{error}_0) - p_1)}{\sqrt{p_1(1-p_1)/n}} \text{ and } z_2 = \frac{((p_0 + \text{error}_0) - p_1)}{\sqrt{p_1(1-p_1)/n}}.$$

This analysis is valid when n is large.

Finally, since

$$\beta = P(z_1 < Z < z_2) = P(Z \leq z_2) - P(Z \leq z_1),$$

we can use Tables 5.1–5.2 (pages 109–110) to find β when sample sizes are large.

Example 6.20 ($n = 120$, $p_0 = 0.28$) For example, suppose that we are interested in testing $p = 0.28$ versus $p \neq 0.28$ at the 5% significance level, using a simple random sample of size 120 from a large population. Then

$$p_0 = 0.28 \text{ and } \text{error}_0 = (1.960)\sqrt{(0.28)(0.72)/120} = 0.0803.$$

If the true population proportion is 0.33, then

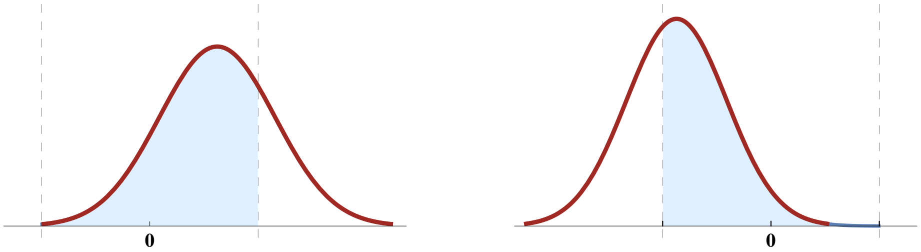
$$z_1 = \frac{((0.28 - 0.0803) - 0.33)}{\sqrt{(0.33)(0.67)/120}} = -3.04, \quad z_2 = \frac{((0.28 + 0.0803) - 0.33)}{\sqrt{(0.33)(0.67)/120}} = 0.71$$

and the probability of making a type II error is

$$\beta = P(-3.04 \leq Z \leq 0.71) = 0.7611 - 0.0012 = 0.7599.$$

Figure 6.6(a) (page 141) shows the distribution of the z-statistic when the true population proportion is 0.33. The area of the shaded region is 0.7599.

Figure 6.6: *Approximate sampling distributions of* $z = (\widehat{p} - 0.28)\,/\sqrt{(0.28)(0.72)/n}$ *when the sample size is* $n = 120$. *The true population proportion is 0.33 in part (a) and 0.21 in part (b). In each case, the area of the shaded region represents the probability of making a type II error for a 5% test of the null hypothesis* $p = 0.28$ *versus the alternative that* $p \neq 0.28$.



(a) *Distribution when* $p_1 = 0.33$ (b) *Distribution when* $p_1 = 0.21$

If instead the true population proportion is 0.21, then

$$z_1 = \frac{((0.28 - 0.0803) - 0.21)}{\sqrt{(0.21)(0.79)/120}} = -0.28, \quad z_2 = \frac{((0.28 + 0.0803) - 0.21)}{\sqrt{(0.21)(0.79)/120}} = 4.04$$

and the probability of making a type II error is

$$\beta = P(-0.28 \leq Z \leq 4.04) = 1 - 0.3897 = 0.6103.$$

Figure 6.6(b) shows the distribution of the z-statistic when the true population proportion is 0.21. The area of the shaded region is 0.6103.

Recall that Tables 5.1–5.2 (pages 109–110) give cumulative probabilities

$$P(Z \leq z) \text{ for } z = -3.89, -3.88, -3.87, \ldots, 3.87, 3.88, 3.89.$$

When $z < -3.89$, the cumulative probability is 0 (with 4 decimal-place accuracy). When $z > 3.89$, the cumulative probability is 1 (with 4 decimal-place accuracy).

In particular, the probability of the event "$Z \leq 4.04$" in the example above is equal to 1 with 4 decimal places of accuracy.

6.5.8 Power at a Specific Alternative Parameter Value

Suppose that the null hypothesis is false. If we *incorrectly* conclude that observed differences are due to chance, then we have made a type II error, but if we *correctly* conclude that observed differences are due to something other than chance, then no error has been made.

In two-sided tests of a population parameter, the probability of correctly concluding that observed differences are due to something other than chance at a specific value of the population parameter is called the *power* of the test at that parameter value.

Power is complementary to the probability of making a type II error. Specifically, if β is the probability making a type II error at a specific value of the population parameter, then

$$\text{Power} = 1 - P(\text{Type II Error}) = 1 - \beta.$$

Example 6.21 ($n = 120$, $p_0 = 0.28$, continued) Using the computations in the example beginning on page 140, we know that the

- Power equals $0.2401 = 1 - 0.7599$ when the true proportion is 0.33 and the
- Power equals $0.3897 = 1 - 0.6103$ when the true proportion is 0.21.

Thus, there is about a 24.01% chance of correctly concluding that an observed difference is due to something other than chance when the true proportion is 0.33, and about a 38.97% chance of correctly concluding that an observed difference is due to something other than chance when the true proportion is 0.21.

The following table gives the probability of a type II error and the power at eight different alternative population proportions:

True population proportion, p_1	0.14	0.18	0.22	0.26	0.30	0.34	0.38	0.42
P(type II error), β	0.0298	0.2875	0.7045	0.9280	0.9172	0.6803	0.3286	0.0927
Power, $1 - \beta$	0.9702	0.7125	0.2955	0.0720	0.0828	0.3197	0.6714	0.9073

Computations were done using the formulas described earlier, with $n = 120$ and $p_0 = 0.28$. Notice that as the true population proportion moves further from the null hypothesis value of 0.28, the type II error decreases and the power increases.

Footnotes. The power of an hypothesis test can be informally described as the ability of the test to detect a difference if there truly is one. Clinical investigations often have poor power because researchers fail to consider both the significance level and the power at an alternative of clinical importance when designing a study. The next section of this chapter considers study designs for two-sided tests of population proportions.

A good way to summarize hypothesis test decisions (accept or reject), potential errors and associated probabilities when testing the null hypothesis $p = p_0$ versus the alternative hypothesis that $p \neq p_0$ is with the following table:

	Accept $p = p_0$	*Reject* $p = p_0$
Population proportion is p_0*:*	A correct decision is made with probability $1 - \alpha$	A type I error is made with probability α
Population proportion is p_1*:*	A type II error is made with probability β	A correct decision is made with probability $1 - \beta$

If the true population proportion is p_0 (first row of table), then accepting the null hypothesis is the correct decision and rejecting the null hypothesis results in an error of type I; an error of type I is made with probability α. If the true population proportion is p_1 (second row of table), then rejecting the null hypothesis is the correct decision and accepting the null hypothesis results in an error of type II; an error of type II is made with probability β when the true population proportion is p_1.

6.5.9 Sample Size Computations

Researchers are often interested in designing a two-sided test of a population parameter with significance level α, and with power equal to $(1-\beta)$ at a specific value of the parameter.

Common choices for power are given in the second row of the following table:

P(type II error), β	0.20	0.15	0.10	0.05	0.01
Power, $(1-\beta)$	0.80	0.85	0.90	0.95	0.99

Sample size computations when testing $p = p_0$ versus $p \neq p_0$. Suppose that we are interested in constructing a $100\alpha\%$ test of $p = p_0$ versus $p \neq p_0$, with power equal to $(1-\beta)$ when the true population proportion is p_1. We assume that the computed sample size will be large enough so that the sampling distribution of the sample proportion $\widehat{p}$ is well-approximated by a normal distribution, and large enough so that

1. The probability that the test statistic falls below the negative of the critical value for the test is the power when $p_1 < p_0$,

   $$P(\text{Test Statistic} \leq -z_{\alpha/2}) = 1-\beta,$$

   as illustrated in Figure 6.7(a) (page 144), and
2. The probability that the test statistic falls above the critical value for the test is the power when $p_1 > p_0$,

   $$P(\text{Test Statistic} \geq z_{\alpha/2}) = 1-\beta,$$

   as illustrated in Figure 6.7(b).

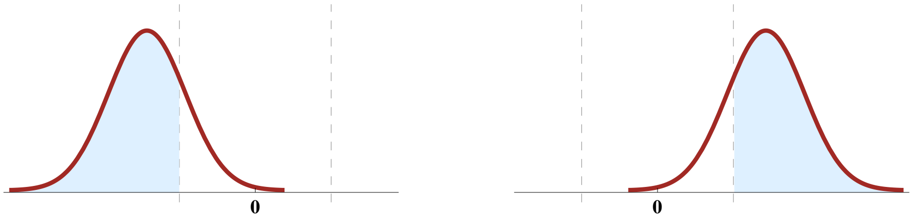
Let $z_{\alpha/2}$ be the critical value for two-sided analyses from Table 6.1 (page 124) and z_β be the critical value for one-sided analyses from Table 6.2 (page 145). Then the assumptions listed above for two-sided tests of population proportions lead to the following sample size formula:

$$n = \left(\frac{z_{\alpha/2}\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p_1(1-p_1)}}{p_1 - p_0} \right)^2 .$$

The following example illustrates the computations in a setting of current interest to researchers in the health sciences.

Example 6.22 (Staph Infections) A staph infection is caused by the staphylococcus bacteria, a type of germ commonly found on the skin or in the nose of even healthy individuals. In most cases, staph infections remain at skin level and are relatively harmless. Staph infections can be deadly, however, if the bacteria invade deeper into an individual's body.

Figure 6.7: *Approximate sampling distributions of $z = (\widehat{p} - p_0) / \sqrt{p_0(1-p_0)/n}$. The true population proportion is less than the null hypothesis value in part (a), and greater than the null hypothesis value in part (b). In each case, the area of the shaded region represents the power of a size α test of the null hypothesis $p = p_0$ versus the alternative that $p \neq p_0$.*



(a) *Distribution when $p_1 < p_0$* (b) *Distribution when $p_1 > p_0$*

Serious staph infections are common in surgical patients. In fact, it is estimated that 22% of surgical patients carry the staphylococcus infection in their nasal cavities at the time of admission. If those infected with the bacteria are treated at the time of admission, then it is possible that serious post-surgical infections could be reduced.

A team of researchers has developed a nasal ointment they believe will reduce the chance of serious post-surgical infection if applied at the time of admission to those who test positive for a staph infection. A clinical difference of 10% would be important.

To be safe, the team decides to design a test of

$$p = 0.22 \text{ versus } p \neq 0.22 \text{ at the 1\% significance level,}$$

using a sample size large enough so that there is at least an 80% chance of rejecting $p = 0.22$ if the post-surgical infection rate is 12% (the infection rate is reduced by 10%), and if the post-surgical infection rate is 32% (the infection rate is increased by 10%).

Let $p_0 = 0.22$, $\alpha = 0.01$ and $\beta = 0.20$. Now,

(1) If $p_1 = 0.12$, then the sample size is

$$n = \left(\frac{(2.576)\sqrt{(0.22)(0.78)} + (0.842)\sqrt{(0.12)(0.88)}}{(0.12 - 0.22)} \right)^2 \approx 179.752 \nearrow 180.$$

(2) If $p_1 = 0.32$, then the sample size is

$$n = \left(\frac{(2.576)\sqrt{(0.22)(0.78)} + (0.842)\sqrt{(0.32)(0.68)}}{(0.32 - 0.22)} \right)^2 \approx 213.122 \nearrow 214.$$

The larger of these two numbers is 214. Thus, the team will need to use a sample of size 214 (or more) to conduct a test of $p = 0.22$ versus $p \neq 0.22$ at the 1% significance level, with a power of 80% or more when the true infection rate is between 12% and 32%.

Table 6.2: *Critical values for one-sided analyses when probability distributions are well-approximated by normal distributions. These critical values are used for sample size computations at pre-specified levels of power, $1-\beta$.*

$100\beta\%$:	20%	15%	10%	5%	1%
$100(1-\beta)\%$:	80%	85%	90%	95%	99%
Critical Value, z_β:	0.842	1.036	1.282	1.645	2.326

Similar computations can be done to find sample sizes for 85%, 90%, 95% and 99% power. The results are summarized in the following table:

Sample sizes for 1% tests of $p = 0.22$ *versus* $p \neq 0.22$*:*

Power as a Percent:	80%	85%	90%	95%	99%
True Proportion is $p_1 = 0.12$*:*	180	198	221	257	333
True Proportion is $p_1 = 0.32$*:*	214	241	278	337	464

Thus, if resources are available, the researchers may choose to use a larger sample size in order to increase the test's ability detect clinically important differences.

Footnotes. There are three safety measures built into the design of the staph infection study in the example above:

1. Although the researchers believe that their ointment will reduce the rate of serious post-surgical infections, they recognize that it is also important to determine if the ointment will actually do harm. Thus, the study is designed using a two-sided alternative ($p \neq 0.22$) and not a one-sided alternative ($p < 0.22$).

2. Although many studies are designed to be conducted at the 5% significance level, the staph infection study will be conducted at the 1% level. This stringent criterion for statistical significance is more appropriate for drug studies than the 5% level.

3. Using the criterion of 10% difference in post-surgical infection rate as clinically important, the researchers checked the sample sizes for both 10% below and 10% above the current post-surgical infection rate, and then chose the larger number. This method of choosing the larger sample size is in keeping with the two-sided approach the researchers are taking in the study.

An even better study design would be a randomized comparative study using a total of $n = n_1 + n_2$ individuals, where n_1 individuals are randomly selected to use the nasal ointment developed by the researchers and the remaining n_2 individuals to use a placebo ointment. Methods for analyzing randomized comparative studies are discussed in the next chapter.

6.6 Brief Summary and Additional Examples

The term *statistical inference* refers to a collection of methods for generalizing from a sample to a population. Valid inferences can only be made when samples have been chosen using probability (or chance) methods. If a chance method has not been used in sampling, then no inference can be made from sample to population.

Simple random samples. This chapter assumes that simple random samples are used to make inferences about population means and population proportions. If a different type of probability sample is used, then different statistical methods are needed. For example, if a cluster random sample or a stratified random sample (Section 2.4.2, page 8) is used instead of a simple random sample, then new formulas are needed for standard errors.

Example 6.23 (Cluster Sampling and Educational Levels) To illustrate this point, suppose there are 4,000 individuals living in an adult community and that the following table gives the distribution of the exact number of years each attended school:

Distribution of Number of Years of Schooling:

8 yrs	12 yrs	16 yrs	20 yrs	Total
1400 (35%)	1800 (45%)	400 (10%)	400 (10%)	4,000 (100%)

In addition, suppose that the community consists of 160 neighborhoods (or *clusters*) of exactly 25 individuals each,

$$(160 \text{ neighborhoods}) \times (25 \text{ individuals/neighborhood}) = 4{,}000 \text{ individuals},$$

and that we would like to construct the sampling distribution of the sample mean number of years an individual attended school, $\overline{X}$, based on the following sampling scheme:

1. Randomly choose 2 neighborhoods from among the 160 neighborhoods and
2. Compute the sample mean for the 50 individuals in the chosen neighborhoods.

Using information in the table above, we know that an adult living in this community attended school an average of 11.8 years, with a standard deviation of 3.68 years. If a simple random sample of size 50 were to be used, the sampling distribution of $\overline{X}$ would have mean 11.8 yrs and standard deviation approximately $3.68/\sqrt{50} = 0.52$ yrs.

Under the cluster sampling scheme described above, the mean is still 11.8 but the standard deviation depends on the distributions of educational levels in the various neighborhoods. Since neighborhoods with lower house prices tend to attract individuals with lower incomes, neighborhoods with higher house prices tend to attract individuals with higher incomes, and income and educational level are positively associated, the distributions of educational levels could be quite different in the various neighborhoods.

The computer was used to simulate neighborhood structures assuming that, on average, 75% of neighbors attended school the same number of years. The following table shows the distribution of sample means for all $\binom{160}{2} = 12,720$ choices of two neighborhoods:

[8,9.5)	[9.5,11)	[11,12.5)	[12.5,14)	[14,15.5)	[15.5,17)	[17,18.5)	[18.5,20]	Total
1431	3929	3428	1772	1461	436	206	57	12,720

In this simulation, the standard error of the sample mean (the standard deviation of the sampling distribution of $\overline{X}$ under cluster sampling) was 2.17. This number is quite a bit larger than the standard error of the sample mean under simple random sampling.

In general, for a given sample size, standard errors computed under cluster sampling will be larger than standard errors computed under simple random sampling. Thus, for a given sample size, confidence intervals based on cluster samples will be wider than intervals based on simple random samples, and hypothesis tests based on cluster samples will have lower power than tests based on simple random samples. The message here is not that cluster random sampling should never be used; in some situations, cluster sampling may be the best choice. The message is that the statistical method you use to analyze data from a study should match the probability method you used to produce the data.

Central limit theorem. This chapter focuses on situations where sampling distributions are approximately normally distributed when the sample size is large. The key result that allows us to use the large sample normal approximation is known as the *central limit theorem*.

The French mathematician Abraham de Moivre (1667-1754) proved the first version of the theorem. His work was later generalized by the French mathematician Pierre-Simon Laplace (1749-1827), whose name is usually associated with the result. It wasn't until the early part of the twentieth century, however, that the word "central" was applied to the name of the theorem, perhaps because sampling distributions focus on the center of the original distribution when sample sizes are large.

Two-sided versus one-sided inference. This chapter introduces methods for constructing two-sided confidence intervals and for conducting two-sided hypothesis tests. The introductions to confidence intervals and hypothesis tests were written to emphasize the parallels between these techniques and to emphasize the importance of the sampling distribution of the statistic used to estimate the parameter of interest in each case.

Two-sided inference methods are preferred to one-sided methods in practice. In addition, a two-sided confidence interval can often take the place of a two-sided test. For example, instead of testing

$$\mu = \mu_0 \text{ versus } \mu \neq \mu_0 \text{ at the } 100\alpha\% \text{ significance level,}$$

we could construct a $100(1-\alpha)\%$ confidence interval for μ. If μ_0 is in the interval, our conclusion is that we believe the observed difference between the sample mean $\overline{x}$ and the hypothesized population mean μ_0 occurred by chance; if μ_0 is not in the interval, our conclusion is that we believe the observed difference did not occur by chance.

Confidence intervals versus hypothesis tests. The observed significance level (or p-value) is an important concept from hypothesis testing that does not have a parallel in confidence interval methodology. Observed significance levels are routinely included in computer-generated reports of hypothesis tests.

Other important concepts from hypothesis testing that don't have parallels in confidence interval methodology are the type II error and the power at a specific alternative value of the parameter of interest. Designing a test with good power at alternatives of clinical importance is an essential component of good study design.

Observed significance level. People often mistake the observed significance level (or p-

value) with the probability that the null hypothesis is true. In fact, the observed significance level is computed *assuming* the null hypothesis is true.

Example 6.24 (Computer Simulation, continued) Consider again the computer simulation used to illustrate the concept of significance level from Section 6.5.5 (page 137).

The computer was used to generate 500 simple random samples of size 120 from a large population, where 28% of individuals have a characteristic of interest. For each sample, a sample proportion $\widehat{p}$ was computed. The sample proportions ranged from 0.175 to 0.408 with the following distribution:

$[0, 0.20)$	$[0.20, 0.25)$	$[0.25, 0.30)$	$[0.30, 0.35)$	$[0.35, 0.40)$	$[0.40, 1.00)$	Total
12	91	224	144	28	1	500

For each sample, the z-statistic,

$$z = (\widehat{p} - 0.28) \,/\, \left(\sqrt{(0.28)(0.72)/120}\right),$$

and the observed significance level,

$$\text{Observed Significance Level} = P(|Z| \geq |z|),$$

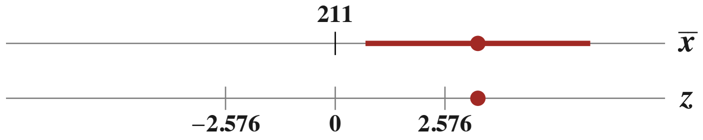
for a test of $p = 0.28$ versus $p \neq 0.28$ were computed. Observed significance levels (p-values) ranged from 0.0017 to 0.9352 with the following distribution:

$[0, 0.2)$	$[0.2, 0.4)$	$[0.4, 0.6)$	$[0.6, 0.8)$	$[0.8, 1.0]$	Total
116	94	102	123	65	500

In this simulation, the null hypothesis is true and the observed significance levels are distributed roughly uniformly over the interval $[0, 1]$.

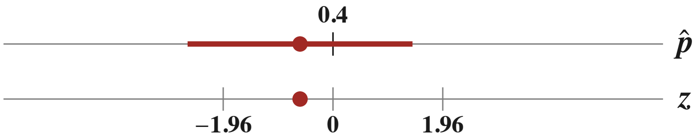
Review of the three main testing examples. It is instructive to use graphics to review the components of the three main testing examples in this chapter.

(1) *Mean Serum Cholesterol Level Example (beginning on page 134):*



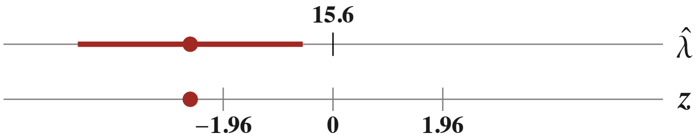
In this example, the hypothesized center of the distribution was $\mu_0 = 211$. The observed value of the test statistic was in the upper part of the rejection region for a 1% test, the observed significance level was less than 0.01, and the 99% confidence interval for the population mean (centered at $\overline{x} = 222.586$) was entirely above $\mu_0 = 211$.

(2) *Gender Bias Example (beginning on page 135):*



In this example, the hypothesized center of the distribution was $p_0 = 0.40$. The observed value of the test statistic was in the acceptance region for a 5% test, the observed significance level was greater than 0.05, and the 95% confidence interval for the population proportion (centered at $\widehat{p} = 0.30$) contained $p_0 = 0.40$.

(3) *Suicide Rates Example (beginning on page 136):*



In this example, the hypothesized center of the distribution was $\lambda_0 = 15.6$. The observed value of the test statistic was in the lower part of the rejection region for a 5% test, the observed significance level was less than 0.05, and the 95% confidence interval for the population mean rate (centered at $\widehat{\lambda} = 13.364$) was entirely below $\lambda_0 = 15.6$.

Significance and significant. The words "significance" and "significant" have been used in several places in this chapter:

1. The significance level α is the pre-specified probability of making a type I error.
2. The observed significance level is the probability that the test statistic equals the observed value or something even more extreme assuming the null hypothesis is true.
3. The results of a test are statistically significant at the $100\alpha\%$ level if the observed value of the test statistic falls in the rejection region for a test conducted at the $100\alpha\%$ level.

In addition, we often say that we are doing *significance testing* when the interpretation of results focuses on the observed significance level (or p-value) and there is no mention of a pre-specified value of α. If the p-value for a test was 0.00001, for example, then the results would be statistically significant at the $100\alpha\%$ level for *any* of the usual choices of α.

Many people mistakenly believe that

1. "statistically significant" is synonymous with "important" and
2. "not statistically significant" is synonymous with "unimportant"

when interpreting the results of hypothesis tests. An important result is one that has substantive or practical significance to researchers.

A finding can be statistically significant without being of practical importance. For example, suppose that we are interested in testing whether the mean serum cholesterol level for hypertensive male smokers is the same as the mean level for all men living in the United States (namely, 211 mg/dL) based on information from a simple random sample of size 2500 from the subpopulation of hypertensive male smokers, and that the following table summarizes the results:

Large sample test of $\mu = 211$ *versus* $\mu \neq 211$*:*

$\overline{x}$	n	s	se	z-Statistic	p-Value
215.1	2500	57.21	1.144	3.583	0.0003

The results are statistically significant using any of the usual choices for α. But the observed difference between the sample mean and the hypothesized population mean,

$$\overline{x} - \mu_0 = 215.1 - 211 = 4.1,$$

is of little or no practical importance. The very large sample size has made the test sensitive to detecting what appears to be a very small difference in means.

At the other extreme, when working with very small sample sizes, even large observed differences may not be statistically significant. We will revisit this situation on page 181, once we have developed methods for analyzing small samples.

The lessons here are the following:

(1) You should design studies so that the sample size is large enough to detect differences of clinical importance with high power, and

(2) You should report as "significant" only those results that demonstrate both statistical significance and practical importance.

Type II errors, power and sample size computations. Finding the probability of making a type II error and corresponding power at a specific alternative value, and finding the sample size needed to conduct a test with given power at a specific alternative value, involve many computations. For this reason, the illustrations in Section 6.5.7 (page 139), Section 6.5.8 (page 141) and Section 6.5.9 (page 143) were restricted to situations involving tests of population proportions (p). Similar methods could be developed for situations involving tests of population means (μ) and population mean rates (λ).

7 Large Sample Analysis of Two Samples

This chapter continues our introduction to statistical inference. Large sample methods for analyzing two samples drawn independently from two populations are introduced. Applications are stressed throughout the chapter.

Recall (from Section 6.2, page 119) that the sampling distributions of the sample mean, the sample proportion and the sample mean rate are well-approximated by normal distributions when sample sizes are large. In the two sample setting, and when both sample sizes are large, the sampling distributions of

- The difference in sample means,
- The difference in sample proportions, and
- The difference in sample mean rates

are also well-approximated by normal distributions, and critical values for the standard normal distribution (from Table 6.1, page 124) can be used to construct confidence intervals and to conduct hypothesis tests.

7.1 Large Sample Confidence Interval Methods

We begin with confidence interval procedures for our three main application areas. In each case, the subscript $i = 1$ corresponds to information for the first sample and the subscript $i = 2$ corresponds to information from the second sample. Further, each confidence interval formula will be of the form

$$\text{“}(\text{Estimate}_1 - \text{Estimate}_2) \pm z_{\alpha/2}\ \sqrt{(\text{ESE}_1)^2 + (\text{ESE}_2)^2}\text{,”}$$

where ESE_i is the estimated standard error of Estimate_i, for $i = 1, 2$, and $z_{\alpha/2}$ is the critical value for the standard normal distribution.

Note that, in the independent samples setting, the estimated standard error of the difference in estimates is the square root of the sum of the squares of the estimated standard errors of each estimate,

$$se(\text{Estimate}_1 - \text{Estimate}_2) = \sqrt{(\text{ESE}_1)^2 + (\text{ESE}_2)^2}.$$

7.1.1 Difference in Population Means

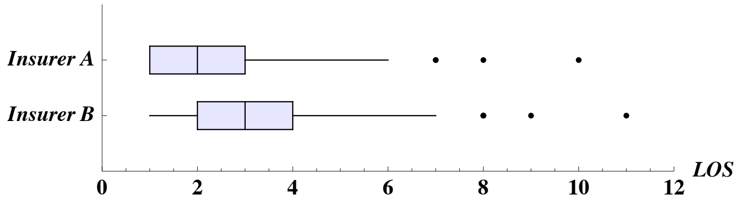
Let $\overline{x}_i$ be the sample mean and s_i be the sample standard deviation of a simple random sample of size n_i chosen from a population with mean μ_i, for $i = 1, 2$.

Then an approximate $100(1-\alpha)\%$ confidence interval for the difference in population means, $\mu_1 - \mu_2$, is the interval

$$(\overline{x}_1 - \overline{x}_2) \pm z_{\alpha/2}\sqrt{(s_1^2/n_1) + (s_2^2/n_2)},$$

This formula can be used when the samples have been chosen independently, and when both sample sizes are large, but small relative to their population sizes.

Figure 7.1: *Summaries of lengths of hospital stays for pediatric asthma patients insured by two different insurance companies.*



Insurer A	*Q1:* 1.0 days	*Median:* 2.0 days	*Mean:* 2.321 days
Group:	*Q2:* 2.0 days	*IQR:* 2.0 days	*SD:* 1.226 days
	Q3: 3.0 days	*Outliers:* 7, 7, 8, 10	*Sample Size:* 393
Insurer B	*Q1:* 2.0 days	*Median:* 3.0 days	*Mean:* 2.909 days
Group:	*Q2:* 3.0 days	*IQR:* 2.0 days	*SD:* 1.579 days
	Q3: 4.0 days	*Outliers:* 8, 8, 9, 11	*Sample Size:* 396

Example 7.1 (Difference in Mean Lengths of Hospital Stay) (Houchens & Schoeps, in Peck et al, 1998, pages 45–64) As part of a study examining the quality of health care for pediatric asthma patients, researchers compared the length of hospital stay for 393 young people whose families were insured through "Insurer A" and 396 young people whose families were insured through "Insurer B." The samples were chosen randomly from all asthma admissions at 29 metropolitan hospitals over a one year period. Children ranged in age from 2 to 17 years old.

Numerical and graphical summaries of the length of stay data are given in Figure 7.1 (page 152). For these data, the difference in mean lengths of stay is

$$\overline{x}_A - \overline{x}_B = 2.321 - 2.909 = -0.588 \text{ days}$$

and a 99% confidence interval for the difference in population means $(\mu_A - \mu_B)$ is

$$-0.588 \pm 2.576\sqrt{(1.226)^2 + (1.579)^2} \;\Rightarrow\; -0.588 \pm 0.259 \;\Rightarrow\; (-0.847, -0.329).$$

Thus, with 99% confidence, we believe that the mean length of hospital stay for those whose families were insured through Insurer A is between about 0.30 and 0.85 days <u>*less*</u> than the mean for those whose families were insured through Insurer B.

Footnote. The analysis reported in the example above served as a starting point for further studies of the possible differences between the two insurance companies. Researchers used further data to examine many questions, including those listed below:

Were there systematic differences in the health care interventions supported by the two companies? Was Insurer A doing a better job managing health care costs or was the company pressuring hospitals to release sick patients earlier than they should be released? Are patients whose families use Insurer B generally sicker than those whose families use Insurer A?

7.1.2 Difference in Population Proportions

Let $\widehat{p}_i$ be the sample proportion of a simple random sample of size n_i chosen from a population with proportion p_i, for $i = 1, 2$. Then an approximate $100(1-\alpha)\%$ confidence interval for the difference in population proportions, $p_1 - p_2$, is the interval

$$(\widehat{p}_1 - \widehat{p}_2) \pm z_{\alpha/2}\sqrt{(\widehat{p}_1(1-\widehat{p}_1)/n_1) + (\widehat{p}_2(1-\widehat{p}_2)/n_2)},$$

This formula can be used when the samples have been chosen independently, when both sample sizes are large, but small relative to their population sizes, and when $n_i p_i \geq 10$ and $n_i(1-p_i) \geq 10$ for $i = 1, 2$.

Example 7.2 (Difference in Proportions of Heart Attacks) (*New England Journal of Medicine* (1989) 321:129-135) As part of a study of possible health benefits of taking low-dose aspirin (325 milligrams every other day) for extended periods of time, more than 22 thousand healthy male physicians at least 40 years old were randomly assigned to either take aspirin or a pill with no active ingredients (a placebo) for a period of 5 years. One question of interest to the researchers was whether taking low-dose aspirin would reduce the incidence of fatal heart attacks. The results are summarized in the following table:

	Heart Attack	*No Heart Attack*	*Total*
1: Placebo Group	26	11008	11034
2: Aspirin Group	10	11027	11037

For these data, the difference in sample proportions of doctors experiencing a fatal heart attack within a 5-year time period is

$$\widehat{p}_1 - \widehat{p}_2 = (26/11034) - (10/11037) = 0.001450$$

(roughly 14.5 per ten thousand) with an estimated standard error of

$$\sqrt{\left(\frac{(26/11034)(11008/11034)}{11034}\right) + \left(\frac{(10/11037)(11027/11037)}{11037}\right)} = 0.0005432,$$

and a 95% confidence interval for the difference in population proportions $(p_1 - p_2)$ is

$$0.00145 \pm 1.960(0.0005432) \;\Rightarrow\; 0.001450 \pm 0.001065 \;\Rightarrow\; (0.0003856, 0.002515).$$

Thus, with 95% confidence, we believe there will be between about 3.9 and 25.2 more fatal heart attacks per ten thousand men within a 5-year period among those who do not take low-dose aspirin regularly than among those who do take the medication.

Footnotes. The aspirin and heart attack example above is based on the "aspirin arm" of the first Physicians' Health Study. Large sample sizes were needed to assess relatively small differences in proportions. Other outcomes, such as incidence of stroke, were also considered.

When proportions are small, researchers often report the ratio of proportions (known as the *relative risk*) rather than the difference in proportions. For the study above, the estimated relative risk is

$$RR = \widehat{p}_1/\widehat{p}_2 = 2.60.$$

That is, male physicians who do not take low-dose aspirin regularly are about 2.60 times more likely to die of a heart attack during a 5-year period than those who do take aspirin.

Finally, it should be noted that the first Nurses' Health Study, which enrolled thousands of female nurses, began in the late 1970's. This series of studies has yielded important results about women's health. See page 178 for an example based on the first study.

7.1.3 Difference in Population Mean Rates

Let $\widehat{\lambda}_i$ be the sample mean rate for a Poisson distribution with rate λ_i per unit time or space observed over n_i units, for $i = 1, 2$. Then an approximate $100(1-\alpha)\%$ confidence interval for the difference in Poisson mean rates, $\lambda_1 - \lambda_2$, is the interval

$$(\widehat{\lambda}_1 - \widehat{\lambda}_2) \pm z_{\alpha/2}\sqrt{(\widehat{\lambda}_1/n_1) + (\widehat{\lambda}_2/n_2)},$$

This formula can be used when the samples have been observed independently, and when $n_i\lambda_i > 100$, for $i = 1, 2$.

Example 7.3 (Difference in Mean Infection Rates) Hepatitis B is a serious liver infection caused by the hepatitis B virus. In its acute form, the infection lasts about six months. In the late 1970's, public health officials in the Sokolov district of Czechoslovakia (now the Czech Republic) determined that the occurrence of acute hepatitis B could be modeled using Poisson distributions.

In this example, we will compare the mean quarterly rates of infection for two groups of patients with acute hepatitis B (*HBsAg-positive* and *HBsAg-negative*) using information gathered over a 6-year period. Each patient was classified as either HBsAg-positive ("hepatitis B surface antigen positive") or HBsAg-negative ("hepatitis B surface antigen negative") on the first day of hospitalization. Patients classified as HBsAg-positive can easily pass the infection to others. The numbers of cases and mean infection rates were as follows:

Sample 1: HBsAg-Positive

Number of Cases	Number of Quarters	Rate per Quarter
120	24	5.0

Sample 2: HBsAg-Negative

Number of Cases	Number of Quarters	Rate per Quarter
321	24	13.375

For these data, the difference in sample mean rates is

$$\widehat{\lambda}_1 - \widehat{\lambda}_2 = 5.0 - 13.375 = -8.375 \text{ cases per quarter},$$

with an estimated standard error of

$$\sqrt{(5.0/24) + (13.375/24)} = 0.875,$$

and a 99% confidence interval for the difference in population parameters $(\lambda_1 - \lambda_2)$ is

$$-8.375 \pm (2.576)(0.875) \Rightarrow -8.375 \pm 2.254 \Rightarrow [-10.629, -6.121].$$

Thus, with 99% confidence, we believe that the mean infection rate for HBsAg-positive cases of acute viral hepatitis is between roughly 6.1 and 10.6 cases per quarter <u>*less*</u> than the mean rate for HBsAg-negative cases of acute viral hepatitis.

Footnote. The study scenario presented in the example above suggests the importance of the Poisson distribution as a model for the occurrence of rare events. Modeling the occurrence of rare events is especially important in the health sciences.

7.2 Large Sample Hypothesis Testing Methods

We next consider tests of equality of population parameters for our three main application areas. In each case, the subscript $i = 1$ corresponds to information for the first sample and the subscript $i = 2$ corresponds to information from the second sample.

Test statistics have the form

$$z = (\text{Estimate}_1 - \text{Estimate}_2)/\text{ESE},$$

where the numerator is the difference in parameter estimates and the denominator is the estimated standard error of this difference assuming the null hypothesis is true.

7.2.1 Test of Equality of Population Means

Suppose that we are interested in testing the null hypothesis that two population means are equal versus the alternative hypothesis that the means are not equal using information from independent simple random samples drawn from distinct populations. Equivalently, suppose that we are interested in testing

$$H_0\text{: } \mu_1 - \mu_2 = 0 \text{ versus } H_A\text{: } \mu_1 - \mu_2 \neq 0$$

based on information from independent random samples.

Let $\overline{x}_i$ be the sample mean and s_i be the sample standard deviation of a simple random sample of size n_i drawn from a population with mean μ_i, for $i = 1, 2$. The estimated standard error of the difference in sample means is

$$\text{ESE} = \sqrt{(s_1^2/n_1) + (s_2^2/n_2)}.$$

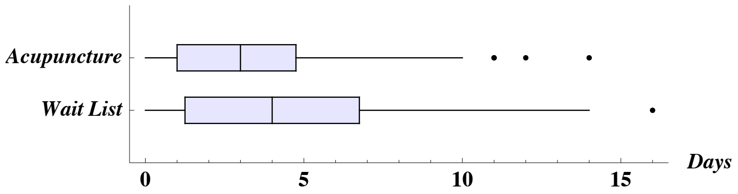
If the sample sizes are large, but small relative to their population sizes, then the test statistic

$$z = (\overline{x}_1 - \overline{x}_2)/\sqrt{(s_1^2/n_1) + (s_2^2/n_2)},$$

has an approximate standard normal distribution, and the rejection region for a two-sided test conducted at the $100\alpha\%$ significance level is $|z| \geq z_{\alpha/2}$.

Example 7.4 (Equality of Mean Numbers of Migraine Days) (based on Baldi & Moore, 2018, page 470) As part of a study examining the effectiveness of acupuncture for migraine headaches, more than 3 hundred individuals who experience migraine headaches on a regular basis were randomly assigned to three groups: those receiving the actual acupuncture treatment, those receiving a "sham" acupuncture treatment (the control group) and those who received no treatment at all (the "wait list" group). Each participant was asked to record the number of headache days they experienced over a 3-month period.

Figure 7.2: *Summaries of numbers of migraine days over a 3-month period for patients using acupuncture and patients using no treatment at all.*



Acupuncture Group:	*Q1:* 1.0 days	*Median:* 2.0 days	*Mean:* 3.348 days
	Q2: 3.0 days	*IQR:* 3.75 days	*SD:* 2.831 days
	Q3: 4.75 days	*Outliers:* 11, 11, 12, 14	*Sample Size:* 132
Wait List Group:	*Q1:* 1.25 days	*Median:* 4.0 days	*Mean:* 4.562 days
	Q2: 4.0 days	*IQR:* 5.5 days	*SD:* 3.703 days
	Q3: 6.75 days	*Outliers:* 16	*Sample Size:* 64

Graphical and numerical summaries for those who received acupuncture (*sample 1*) and for those who were in the wait list group (*sample 2*) are summarized in Figure 7.2 (page 156). These individuals experienced between 0 and 16 migraine days during the study period. Since the observed difference in sample means is negative ($\overline{x}_1 - \overline{x}_2 = 3.348 - 4.562 = -1.214$), there is a suggestion that the acupuncture treatment is effective.

Consider testing the null hypothesis that the mean numbers of migraine days are equal versus the alternative that the means are not equal at 5% significance level using the information in the figure. The rejection region for the test is $|z| \geq 1.960$.

Since the observed value of the test statistic,

$$z = (3.348 - 4.562)/\sqrt{(2.831)^2/132 + (3.703)^2/64} = -2.753,$$

is in the rejection region, the results are statistically significant at the 5% level. Further, we have reason to believe that the mean number of migraine days for those treated with acupuncture is less than the mean number of days for those who use no treatment.

Footnotes. Recall that we can use Tables 5.1-5.2 (pages 109-110) to find cumulative probabilities for normal distributions. Using these tables, the observed significance level for the test in the example above is

$$2P(Z \leq -2.753) = 2(0.0032) = 0.0064,$$

which confirms our conclusion that the results are statistically significant at the 5% level.

Note that it is common practice to include a $100(1-\alpha)\%$ confidence interval for the difference in population means with the results of hypothesis tests for the equality of means. For the example above, a 95% confidence interval is

$$-1.214 \pm 1.960\sqrt{(2.831)^2/132+(3.703)^2/64} \;\Rightarrow\; -1.214 \pm 0.864 \;\Rightarrow\; [-2.078, -0.350].$$

Thus, with 95% confidence, we believe that acupuncture treatments can reduce the mean number of migraine days by between about 0.35 and 2.01 days in a 3-month period.

7.2.2 Test of Equality of Population Proportions

Suppose that we are interested in testing the null hypothesis that two population proportions are equal versus the alternative hypothesis that the proportions are not equal using information from independent simple random samples drawn from distinct populations. Equivalently, suppose that we are interested in testing

$$H_0\colon\ p_1 - p_2 = 0 \text{ versus } H_A\colon\ p_1 - p_2 \neq 0$$

based on information from independent random samples.

Let $\widehat{p}_i = x_i/n_i$ be the sample proportion of individuals from a subpopulation of interest based on a sample of size n_i, for $i = 1, 2$, and let

$$\widehat{p} = (x_1 + x_2)/(n_1 + n_2)$$

be the estimated common proportion assuming the population proportions are equal. The estimated standard error of the difference in sample proportions is

$$\text{ESE} = \sqrt{\widehat{p}(1-\widehat{p})(1/n_1 + 1/n_2)}$$

under the null hypothesis of equality of population proportions.

If the sample sizes are large, but small relative to their population sizes, and $n_i p_i \geq 10$ and $n_i(1-p_i) \geq 10$ for $i = 1, 2$, then the test statistic,

$$z = (\widehat{p}_1 - \widehat{p}_2)/\sqrt{\widehat{p}(1-\widehat{p})(1/n_1 + 1/n_2)},$$

has an approximate standard normal distribution, and the rejection region for a two-sided test conducted at the $100\alpha\%$ significance level is $|z| \geq z_{\alpha/2}$.

Example 7.5 (Equality of Proportions Infected with HIV) (based on Baldi & Moore, 2018, page 504) As part of a double-blind, placebo-controlled study of a promising new vaccine for HIV (human immunodeficiency virus), a group of more than 16 thousand men and women living in Thailand between the ages of 18 and 30 were randomly assigned to either receive the drug or to receive a placebo (with no active ingredients). The participants were carefully monitored for a 3-year period.

By the end of the study period, 0.8% (125/16395) of study participants had become infected with the HIV virus. The results by treatment group are as follows:

	1: Placebo Group	2: Vaccination Group
HIV Infections	0.9% (74/8198)	0.6% (51/8197)

Consider testing the null hypothesis that the proportions infected with HIV in the two groups are equal versus the alternative that they are not equal at the 5% significance level using the information above. The rejection region for this test is $|z| \geq 1.960$.

Since the observed value of the test statistic,

$$z = (74/8198 - 51/8197)/\sqrt{(125/16395)(16270/16395)(1/8198 + 1/8197)} = 2.064,$$

is in the rejection region, the results are statistically significant at the 5% level. Further, we have reason to believe that individuals who receive the vaccination are *less likely* to become infected with the HIV virus than those who are not vaccinated.

Footnotes. There are several ways to add to the analysis above. First, the observed significance level for the test of equality of population proportions is

$$2(1 - P(Z \leq 2.064)) = 2(1 - 0.9803) = 0.0394,$$

which confirms our conclusion that the results are statistically significant at the 5% level.

Second, a 95% confidence interval for the difference in population proportions $(p_1 - p_2)$ can be developed using the methods from Section 7.1.2, (page 153).

Since the observed difference in sample proportions is

$$\widehat{p}_1 - \widehat{p}_2 = 0.0028048$$

(roughly 28 per ten thousand), and the estimated standard error is

$$\sqrt{\left(\frac{(74/8198)(8124/8198)}{8198}\right)\left(\frac{(51/8197)(8146/8197)}{8197}\right)} = 0.00135847,$$

the interval becomes

$$0.0028048 \pm 1.960(0.00135847) \;\Rightarrow\; 0.0028048 \pm 0.0026626 \;\Rightarrow\; [0.0001422, 0.00546741].$$

Thus, with 95% confidence, we believe that there will be between about 1.4 and 54.7 more HIV infections per ten thousand individuals within a 3-year period among those who are not vaccinated than among those who are vaccinated.

Third, we can report the relative risk of infection:

$$RR = \widehat{p}_1/\widehat{p}_2 = 1.4508.$$

That is, we believe that those who are not vaccinated will become infected with the HIV virus within a 3-year period about 1.45 times more often than those who are vaccinated.

7.2.3 Test of Equality of Population Mean Rates

Suppose that we are interested in testing the null hypothesis that two Poisson mean rates are equal versus the alternative hypothesis that the rates are not equal using information from independent samples. Equivalently, suppose that we are interested in testing

$$H_0\colon\ \lambda_1 - \lambda_2 = 0 \text{ versus } H_A\colon\ \lambda_1 - \lambda_2 \neq 0$$

based on information from independently chosen samples.

Let $\widehat{\lambda}_i = x_i/n_i$ be the sample mean rate (the ratio of the number of events to the number of units), for $i = 1, 2$, and let

$$\widehat{\lambda} = (x_1 + x_2)/(n_1 + n_2)$$

be the estimated common mean rate assuming the population rates are equal. The estimated standard error of the difference in sample mean rates is

$$\text{ESE} = \sqrt{\widehat{\lambda}(1/n_1 + 1/n_2)}$$

under the null hypothesis of equality of population parameters.

If occurrence of events follow the assumptions for the Poisson distribution (Section 5.3, page 101) and $n_i\lambda_i > 100$ for $i = 1, 2$, then the test statistic is

$$z = \left(\widehat{\lambda}_1 - \widehat{\lambda}_2\right) / \sqrt{\widehat{\lambda}(1/n_1 + 1/n_2)},$$

has an approximate standard normal distribution, and the rejection rejection for a two-sided test conducted at the $100\alpha\%$ significance level is $|z| \geq z_{\alpha/2}$.

Example 7.6 (Equality of Mean Rates of Childhood Leukemia) (Waller et al, in Lange et al, 1994, pages 3–23) The spatial distribution of the incidence of certain diseases may indicate possible causes. This example is based on an article by University of Minnesota Professor Lance Waller, Cornell University Professor Bruce Turnbull, University of Arizona Professor Larry Clark, and New York State Department of Health Official Philip Nasca.

After the 1986 publication of a research article demonstrating a statistically significant relationship between the incidence of childhood leukemia in children aged 0 to 19 years and exposure to wells contaminated with the volatile organic compound trichloroethylene (TCE) in Woburn, Massachusetts, officials at the New York State Department of Health initiated a series of meetings of statisticians, epidemiologists and public health officials to discuss policies for monitoring the geographic distribution of cancer cases in New York state. The Department of Health was interested in developing surveillance methods to be used in conjunction with the Cancer Registry to scan for possible disease *clusters* (unusually large numbers of cases in small geographic areas). If there was an objective way to detect, prioritize and monitor possible disease clusters, then the limited resources available to the Department of Health could be used to best advantage to save lives.

One of the challenges the researchers faced was the limitations of the data available to them. Because of privacy concerns, they could not obtain exact locations where cancer patients lived. They could only obtain population counts and leukemia incidence counts by *census tract* (small subregions of between 1000 and 4000 individuals).

By working with 1980 census data, and Cancer Registry data for the five years from the beginning of 1978 to the end of 1982, the researchers determined that 592 cases of leukemia were reported in the eight-county region of upstate New York (including Cayuga, Onondaga, Madison, Tompkins, Cortland, Chenango, Tioga, and Broome counties), where roughly one million people lived. This gives an overall approximate incidence rate of

59.2 cases of leukemia per hundred thousand population per 5 years.

They next separated the information into geographic areas: the southern (more industrial) part of upstate New York versus the northern (less industrial) part, with roughly 5 hundred thousand individuals in each part. The results were as follows:

	Southern Part	Northern Part
Rate per hundred thousand	67.8 (339/5)	50.6 (253/5)

Consider testing the null hypothesis that the rates per hundred thousand population in the southern and northern parts of upstate New York are equal versus the alternative that they are not equal using the information above. The rejection region for this test is $|z| \geq 1.960$.

Since the observed value of the test statistic,

$$z = (67.8 - 50.6)/\sqrt{59.2(1/5 + 1/5)} = 17.2/4.8662 = 3.535,$$

is in the rejection region, the results are statistically significant at the 5% level. Further, we have reason to believe that the rate of childhood leukemia is *greater* in the southern part of upstate New York than in the northern part.

Footnotes. There are several ways that we can add to the analysis above, including finding the observed significance level and finding a 95% confidence interval for the difference in rates parameter. The following table summarizes the computations for the rates of childhood leukemia example:

Large Sample Test of $\lambda_1 - \lambda_2 = 0$ *versus* $\lambda_1 - \lambda_2 \neq 0$*:*

$\widehat{\lambda_1} - \widehat{\lambda_2}$	n_1	n_2	95% CI	z-Statistic	p-Value
17.2	5	5	$[7.66, 26.74]$	3.535	0.0004

The small p-value confirms our conclusion that the results are significant at the 5% level, and the fact that the confidence interval lies entirely above 0 confirms our belief that the rate is higher in the southern part than in the northern part of upstate New York. Further, the interval gives us a range of values within which we believe the true rate lies.

The simple analysis given in the leukemia and geographic location example suggests, but does *not* prove, a link between leukemia incidence in upstate New York and proximity to industrial waste sites. The researchers identified 11 hazardous sites containing TCE (trichloroethylene, the organic compound implicated in the famous Woburn, Massachusetts, case[1]); 10 of the 11 sites were in the southern geographic part of their study region. They used the information described in the example to develop several sophisticated strategies for analyzing risk of disease as a function of distance to a given site. When they applied their methods to the upstate New York data, however, they could only show weak associations. Part of the problem was that person-level information was not available to them, due to privacy issues. If greater access to some personal information could be allowed, while respecting the privacy of individuals, then stronger conclusions could be drawn.

[1] See, for example, http://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH/Woburn/ for information about the effects of environmental contamination on the community of Woburn. Another good source of information is the best-selling book *A Civil Action* by Jonathan Harr. The book chronicles the trial that attempted to hold those responsible for the pollution accountable, and was later made into a famous movie of the same name.

7.3 Brief Summary and Additional Examples

This chapter introduces large sample methods for working with independently chosen samples in the three areas of application introduced in Chapter 6.

Comparative studies. Many examples in this chapter illustrate study design principles introduced in Chapter 2. In particular, the study of possible health benefits of taking low-dose aspirin (beginning on page 153) and the study of a promising new vaccine for HIV (beginning on page 157) are examples of randomized comparative experiments.

These two studies are also notable because they involved very large samples. Very large sample sizes were needed because both the incidence of the events of interest (fatal heart attack or infection with the HIV virus) and the expected treatment effects (possible reduction in the proportions experiencing fatal attacks or in the proportions infected with HIV) were small.

Reporting scales. Another notable feature of the studies mentioned in the previous paragraph is that the confidence intervals for the difference in proportions were interpreted using a different scale: "fatal heart attacks per ten thousand men" in the low-dose aspirin study and "HIV infections per ten thousand individuals" in the HIV vaccination study.

The two studies involving differences in Poisson mean rates make direct use of different reporting scales: In the study on hepatitis B infections (beginning on page 154), the reporting scale is "quarterly rates of infection." In the study of incidence of childhood leukemia study (beginning on page 159), the reporting scale is "cases of leukemia per hundred thousand population" in a 5-year period.

Different reporting scales are useful when studying rare (but catastrophic) events.

More on women's health. Women are often underrepresented in clinical trials. The series of studies known collectively as the "Nurses Health Study," mentioned earlier in this chapter, has increased our understanding of women's health issues. Another series of studies, known as the Women's Health Initiative (WHI), also deserves mention. The initial study was launched in 1993 and enrolled more than 160 thousand women nationwide. Several "WHI Extension" studies have continued this work.

Statistics and public policy. Statisticians are often called upon to offer expert testimony in criminal and civil proceedings. The following example is based on an article co-written by the statisticians who served on opposite sides of a criminal case. The example puts the statistical concepts we have learned so far into proper perspective.

Example 7.7 (United States *v.* Kristen Gilbert) (Cobb & Gehlbach, in Peck et al, 2006, pages 3–18) In the early 1990's, Kristen Gilbert was a well-respected member of the nursing staff at a VA (Veterans Administration) hospital in Massachusetts. But other nursing staff members began to notice that the number of deaths on shifts in which Ms. Gilbert was present appeared to be much greater than on shifts in which she was absent. A formal investigation began and Professor Gehlbach was asked to assist the prosecution. Gehlbach examined records for 1641 shifts over 547 days,

$$547 \text{ days} \times 3 \text{ shifts/day} = 1641 \text{ shifts},$$

and discovered that there were hospital deaths during 4.5% (74/1641) of shifts. Gehlbach then separated the information according to when Ms. Gilbert was present or absent from the hospital. The results were as follows:

Gilbert Present:

Shifts	Shifts	
with Deaths	without Deaths	Total
40	217	257
(15.6%)	(84.4%)	(100%)

Gilbert Absent:

Shifts	Shifts	
with Deaths	without Deaths	Total
34	1350	1384
(2.5%)	(97.5%)	(100%)

If we let p_1 be the proportion of shifts in which there are deaths when Ms. Gilbert is present at the hospital and p_2 be the proportion of shifts in which there are deaths when Ms. Gilbert is absent from the hospital, then the difference in sample proportions is

$$\widehat{p}_1 - \widehat{p}_2 = 0.156 - 0.025 = 0.131.$$

Further, the observed value of z for a test of the null hypothesis that the population proportions are equal is

$$z = (0.156 - 0.025)/\sqrt{0.045(1-0.045)(1/257 + 1/1384)} = 0.131/0.0141 = 9.291,$$

and the observed significance level is $2(1 - P(Z \leq 9.291)) = 0$ to 4 decimal places of accuracy.

Since the results are statistically significant using any of the usual significance levels, the conclusion is that we believe the observed difference in proportions did not occur by chance.

Professor Gehlbach's results were presented to the grand jury examining the case, and the grand jury eventually indicted Ms. Gilbert. But, the story does not end here. During pre-trial proceedings, Professor Cobb was asked by the defense team to re-examine Professor Gehlbach's work, with the hope of excluding the analyses from the trial. After examining the report, Professor Cobb made three important points in a report to the judge:

1. Professor Cobb agreed with Professor Gehlbach's analysis of the information. Since the observed significance level was virtually 0, the only conclusion to be drawn was that the observed difference was likely due to something other than chance.

2. Since the information available for analysis was observational and not experimental, the conclusion that the observed difference was likely due to something other than chance does *not* necessarily imply that Ms. Gilbert caused the excess number of deaths during the shifts in which she was present in the hospital. It would be up to the prosecution to eliminate all other reasonable potential causes for the observed difference.

3. Most people do not understand that the observed significance level is calculated assuming that the null hypothesis that the two population proportions are equal is true, and are likely to misinterpret the results.

Because of the potential for misinterpretation, the judge excluded the formal statistical analysis from the criminal trial. Although the statistical analysis was an important first step, it could not be used to establish that Ms. Gilbert had actually killed many of her patients. Finally, it should be noted that Ms. Gilbert was eventually convicted on the strength of the other evidence presented by the prosecution.

8 Small Sample Analyses of Means

This chapter continues our introduction to statistical inference. Small sample methods for analyzing population means, and differences in population means, are introduced. Applications are stressed throughout the chapter.

Analyses of population means in small sample situations requires an additional assumption, namely that the distribution of the characteristic of interest is well approximated by a normal distribution, and a new sampling distribution. The first section introduces this distribution.

8.1 Student t Distribution

Let X represent the value of a characteristic of interest for an individual chosen at random from a study population and suppose that the mean and standard deviation of X are

$$E(X) = \mu \text{ and } SD(X) = \sigma.$$

Further, suppose that the distribution of X is well-approximated by a normal distribution.

t-score. If $\overline{x}$ is the sample mean and s is the sample standard deviation of a simple random sample of size n from the population, then the quantity

$$t = (\overline{x} - \mu)/(s/\sqrt{n})$$

is called the *t-score* for the sample. The t-score is an approximate standardized score, using the sample standard deviation s to approximate σ.

Student t distribution. The sampling distribution of t-scores depends on the sample size. For a sample of size n, the sampling distribution has an approximate

Student t distribution with $(n-1)$ *degrees of freedom.*

Example 8.1 (Women's Heights) (Agresti & Franklin, 2007, page 306) Distributions of heights for adult women are often well-approximated by normal distributions. For example, the distribution of women's heights in North America is well-approximated by a normal distribution with mean 65.0 inches and standard deviation 3.5 inches.

The computer was used to generate 10,000 random samples of size 5 from the normal model for women's heights and to compute the t-score,

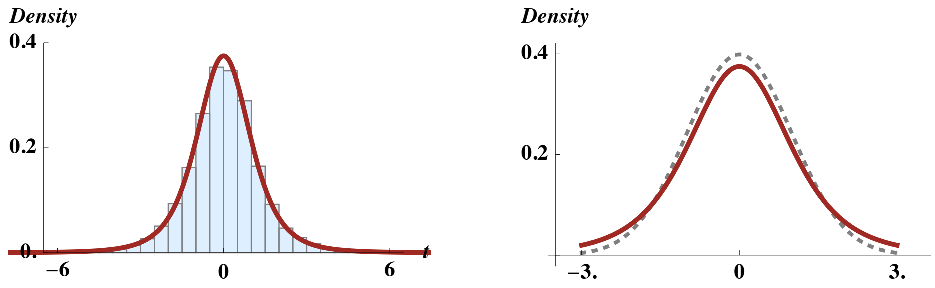
$$t = (\overline{x} - 65.0)/(s/\sqrt{5}),$$

for each sample. Values ranged from -9.66 to 11.69 with the following distribution:

$[-12,-9)$	$[-9,-6)$	$[-6,-3)$	$[-3,0)$	$[0,3)$	$[3,6)$	$[6,9)$	$[9,12]$	Total
2	12	166	4759	4848	196	14	3	10,000

Notice, in particular, that 96.07% (9607/10000) of sample t-scores were between -3 and 3, and that 3.93% (393/10000) of t-scores were outside this range.

Figure 8.1: *Student t distribution with 4 degrees of freedom. Part (a) is a histogram of 10,000 simulated t-scores based on samples of size 5 from the distribution of women's heights, with the Student t distribution superimposed. Part (b) is a plot of the standard normal curve (in gray), with the Student t distribution superimposed, for scores between −3 and 3.*



(a) *Sample t-scores when* $n = 5$ (b) *Comparison of Distributions*

Figure 8.1(a) (page 164) is a histogram of the simulated t-scores, with the Student t distribution with 4 degrees of freedom superimposed. The Student t curve is a close approximation to the histogram.

If the sample size n is large, then the standard normal curve can be used to approximate the sampling distribution of t-scores. But if n is small — as it is in this example — the standard normal curve is too narrow. This is illustrated in Figure 8.1(b), where the curve for the Student t distribution with 4 degrees of freedom (the solid curve) is superimposed on the standard normal curve (the dotted curve) for scores between −3 and 3.

A further comparison is given in the following tables:

Standard Normal Distribution:

Below −3	Between −3 and 3	Above 3	Total
0.0013	0.9973	0.0013	1.0000

Student t Distribution with 4 Degrees of Freedom:

Below −3	Between −3 and 3	Above 3	Total
0.01997	0.96006	0.01997	1.0000

Based on the Student t distribution, we would expect about 3.994% of scores to fall outside the −3 to 3 range; based on the standard normal distribution, we would expect only about 0.26% of scores to fall outside this range. The Student t distribution reflects the increased variability in scores in small samples.

Critical values. Critical values from Student t distributions will take the place of critical values from the standard normal distribution when constructing confidence intervals for population means and conducting two-sided tests about population means when n is small.

The notation $t_{\alpha/2}$ ("t-sub-$\alpha/2$") will be used to denote the critical value needed for a two-sided test conducted at the $100\alpha\%$ significance level and for a $100(1-\alpha)\%$ confidence interval.

Table 8.1: *Critical values for two-sided statistical inference when sampling distributions are well-approximated by Student t distributions, when df ≤ 40.*

Two-Sided Confidence Level as a Percent: $100(1-\alpha)\%$

50%	60%	70%	80%	90%	95%	99%	99.9%

Two-Sided Significance Level as a Percent: $100\alpha\%$

50%	40%	30%	20%	10%	5%	1%	0.1%

df	$t_{\alpha/2}$	$t_{\alpha/2}$	$t_{\alpha/2}$	$t_{\alpha/2}$	$t_{\alpha/2}$	$t_{\alpha/2}$	$t_{\alpha/2}$	$t_{\alpha/2}$
1	1.000	1.376	1.963	3.078	6.314	12.706	63.657	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	9.925	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	5.841	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.861	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.845	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.831	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.819	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.807	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.797	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.787	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.779	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.771	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.763	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.756	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.750	3.646
31	0.682	0.853	1.054	1.309	1.696	2.040	2.744	3.633
32	0.682	0.853	1.054	1.309	1.694	2.037	2.738	3.622
33	0.682	0.853	1.053	1.308	1.692	2.035	2.733	3.611
34	0.682	0.852	1.052	1.307	1.691	2.032	2.728	3.601
35	0.682	0.852	1.052	1.306	1.690	2.030	2.724	3.591
36	0.681	0.852	1.052	1.306	1.688	2.028	2.719	3.582
37	0.681	0.851	1.051	1.305	1.687	2.026	2.715	3.574
38	0.681	0.851	1.051	1.304	1.686	2.024	2.712	3.566
39	0.681	0.851	1.050	1.304	1.685	2.023	2.708	3.558
40	0.681	0.851	1.050	1.303	1.684	2.021	2.704	3.551

Table 8.2: *Critical values for two-sided statistical inference when sampling distributions are well-approximated by Student t distributions, when df* > 40.

Two-Sided Confidence Level as a Percent: $100(1-\alpha)\%$

50%	60%	70%	80%	90%	95%	99%	99.9%

Two-Sided Significance Level as a Percent: $100\alpha\%$

50%	40%	30%	20%	10%	5%	1%	0.1%

df	$t_{\alpha/2}$	$t_{\alpha/2}$	$t_{\alpha/2}$	$t_{\alpha/2}$	$t_{\alpha/2}$	$t_{\alpha/2}$	$t_{\alpha/2}$	$t_{\alpha/2}$
41	0.681	0.850	1.050	1.303	1.683	2.020	2.701	3.544
42	0.680	0.850	1.049	1.302	1.682	2.018	2.698	3.538
43	0.680	0.850	1.049	1.302	1.681	2.017	2.695	3.532
44	0.680	0.850	1.049	1.301	1.680	2.015	2.692	3.526
45	0.680	0.850	1.049	1.301	1.679	2.014	2.690	3.520
46	0.680	0.850	1.048	1.300	1.679	2.013	2.687	3.515
47	0.680	0.849	1.048	1.300	1.678	2.012	2.685	3.510
48	0.680	0.849	1.048	1.299	1.677	2.011	2.682	3.505
49	0.680	0.849	1.048	1.299	1.677	2.010	2.680	3.500
50	0.679	0.849	1.047	1.299	1.676	2.009	2.678	3.496
51	0.679	0.849	1.047	1.298	1.675	2.008	2.676	3.492
52	0.679	0.849	1.047	1.298	1.675	2.007	2.674	3.488
53	0.679	0.848	1.047	1.298	1.674	2.006	2.672	3.484
54	0.679	0.848	1.046	1.297	1.674	2.005	2.670	3.480
55	0.679	0.848	1.046	1.297	1.673	2.004	2.668	3.476
56	0.679	0.848	1.046	1.297	1.673	2.003	2.667	3.473
57	0.679	0.848	1.046	1.297	1.672	2.002	2.665	3.470
58	0.679	0.848	1.046	1.296	1.672	2.002	2.663	3.466
59	0.679	0.848	1.046	1.296	1.671	2.001	2.662	3.463
60	0.679	0.848	1.045	1.296	1.671	2.000	2.660	3.460
65	0.678	0.847	1.045	1.295	1.669	1.997	2.654	3.447
70	0.678	0.847	1.044	1.294	1.667	1.994	2.648	3.435
75	0.678	0.846	1.044	1.293	1.665	1.992	2.643	3.425
80	0.678	0.846	1.043	1.292	1.664	1.990	2.639	3.416
85	0.677	0.846	1.043	1.292	1.663	1.988	2.635	3.409
90	0.677	0.846	1.042	1.291	1.662	1.987	2.632	3.402
95	0.677	0.845	1.042	1.291	1.661	1.985	2.629	3.396
100	0.677	0.845	1.042	1.290	1.660	1.984	2.626	3.390
105	0.677	0.845	1.042	1.290	1.659	1.983	2.623	3.386
110	0.677	0.845	1.041	1.289	1.659	1.982	2.621	3.381
115	0.677	0.845	1.041	1.289	1.658	1.981	2.619	3.377
120	0.677	0.845	1.041	1.289	1.658	1.980	2.617	3.373
125	0.676	0.845	1.041	1.288	1.657	1.979	2.616	3.370
130	0.676	0.844	1.041	1.288	1.657	1.978	2.614	3.367
135	0.676	0.844	1.040	1.288	1.656	1.978	2.613	3.364
140	0.676	0.844	1.040	1.288	1.656	1.977	2.611	3.361
145	0.676	0.844	1.040	1.287	1.655	1.976	2.610	3.359
150	0.676	0.844	1.040	1.287	1.655	1.976	2.609	3.357
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.576	3.291

Tables 8.1-8.2 (pages 165-166) list critical values, where the rows correspond to the degrees of freedom (*df*) of the Student t distribution and the columns correspond to the commonly used confidence and significance levels.

For example,

(1) If we are interested in constructing a 90% confidence interval and the sample size is 18, then the critical value is 1.740 (the entry for the row corresponding to 17 degrees of freedom and the column corresponding to 90% confidence).

(2) If we are interested in conducting a two-sided test at the 5% significance level and the sample size is 12, then the critical value is 2.201 (the entry for the row corresponding to 11 degrees of freedom and the column corresponding to 5% significance).

Footnote: Development of the Student t distribution. The Student t distribution was developed by the British statistician and chemist William Sealy Gosset (1876-1937). Gosset realized that small sample methods were needed to analyze the results of small-scale experiments conducted at the Guinness Breweries in Dublin, Ireland. He published his methods in 1908 using the pseudonym "Student" in order to keep his identity private.

8.2 Analyses of Single Samples

Let $\overline{x}$ be the sample mean and s be the sample standard deviation of a simple random sample of size n from a large population.

8.2.1 Small Sample Confidence Interval for a Population Mean

If the distribution of the characteristic of interest is approximately normal, then an approximate $100(1-\alpha)\%$ confidence interval for the population mean is the interval

$$\overline{x} \pm t_{\alpha/2}\,(s/\sqrt{n})\,,$$

where $t_{\alpha/2}$ is the critical value for the Student t distribution with $n-1$ degrees of freedom.

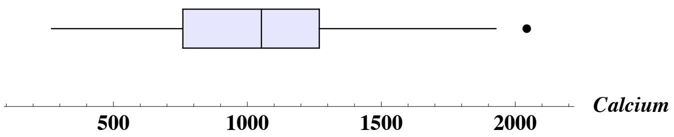
Example 8.2 (Mean Calcium Intake) As part of a study on women's bone health, researchers gathered information on a simple random sample of 24 women between the ages of 19 and 25 living in the greater Washington, DC, area. The following table lists the daily calcium intake in milligrams (mg) for these women, arranged in increasing order.

Daily Calcium Intake ($n = 24$)*:*

269	283	436	613	706	747	796	877	927	1004	1009	1019
1086	1106	1182	1196	1199	1258	1272	1281	1506	1540	1928	2043

Numerical and graphical summaries of these data are given in Figure 8.2 (page 168). Note that, since the box plot of observed calcium intake values is approximately symmetric, and the single outlier is close to the endpoint of the upper whisker, there is no reason to doubt that the population distribution of daily calcium intake values is approximately normally distributed.

Figure 8.2: *Summaries of daily calcium intake for 24 young women.*



Q1: 759.25 mg	*Median:* 1052.50 mg	*Mean:* 1053.46 mg
Q2: 1052.50 mg	*IQR:* 509.25 mg	*SD:* 442.653 mg
Q3: 1268.50 mg	*Outliers:* 2043	*Sample Size:* 24

Consider estimating the population mean with 95% confidence. Since $n = 24$, the critical value is 2.069 (the entry for the row corresponding to 23 degrees of freedom and the column corresponding to 95% confidence in Table 8.1, page 165).

The confidence interval is computed as follows:

$$1053.46 \pm (2.069)(442.653/\sqrt{24}) \;\Rightarrow\; 1053.46 \pm 186.947 \;\Rightarrow\; [866.513, 1240.410].$$

Thus, with 95% confidence, we believe that the mean daily calcium intake for young women aged 19–25 living in the greater DC area is between about 866.5 and 1240.4 mg.

8.2.2 Small Sample Test of a Population Mean

Suppose that we are interested in testing the null hypothesis that a population mean equals a fixed known constant μ_0 ("mu-naught") versus the alternative hypothesis that the mean is not equal to μ_0,

$$H_0\text{: } \mu = \mu_0 \text{ versus } H_A\text{: } \mu \neq \mu_0,$$

using information from a simple random sample of size n from the population.

If the population size is large and the distribution of the characteristic of interest is approximately normal, then the test statistic is

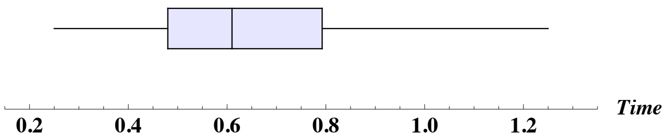
$$t = (\overline{x} - \mu_0)/(s/\sqrt{n}),$$

where $\overline{x}$ is the sample mean and s is the sample standard deviation. The rejection region for a two-sided test conducted at the $100\alpha\%$ significance level is

$$|t| \geq t_{\alpha/2},$$

where $t_{\alpha/2}$ is the critical value for the Student t distribution with $n - 1$ degrees of freedom.

Figure 8.3: *Summaries of cataract surgery times for 24 patients.*



Q1: 0.48 hrs	*Median:* 0.61 hrs	*Mean:* 0.666 hrs
Q2: 0.61 hrs	*IQR:* 0.3125 hrs	*SD:* 0.246 hrs
Q3: 0.7925 hrs	*Outliers:* none	*Sample Size:* 24

Example 8.3 (Mean Cataract Surgery Time) (Kokoska, 2015, page 435) A cataract is a clouding of the normally clear lens of the eye. Many adults with cataracts elect to have cataract surgery, a generally safe elective procedure in which the natural lens is removed and an artificial lens implant is inserted. Past records indicate that the mean time for this type of surgery is 0.75 hours.

As part of a study on elective surgery practices, a simple random sample of 24 adults who had the procedure within the last year was chosen and the surgery times of each individual was recorded. The following table lists the times in hours, arranged in increasing order.

Cataract Surgery Times ($n = 24$)*:*

0.25	0.38	0.42	0.43	0.45	0.48	0.48	0.50	0.55	0.57	0.60	0.60
0.62	0.68	0.75	0.75	0.77	0.77	0.80	0.80	0.88	1.03	1.17	1.25

Numerical and graphical summaries of the surgery times are given in Figure 8.3 (page 169). Note that since the box plot of observed times is approximately symmetric with no outliers, there is no reason to doubt that the population distribution of cataract surgery times is approximately normally distributed.

Consider testing the null hypothesis that the mean surgery time is 0.75 hours versus the alternative that the mean is different from 0.75 hours at the 5% significance level. Since $n = 24$, the rejection region is $|t| \geq 2.069$ (using the entry for the row corresponding to 23 degrees of freedom and the column corresponding to 5% significance in Table 8.1, page 165).

Since the observed value of the test statistic,

$$t = (0.666 - 0.75)/(0.246/\sqrt{24}) = -1.673,$$

is in the acceptance region, the results are not statistically significant at the 5% level. Thus, we believe that the observed difference occurred by chance.

Footnotes. There are several ways to add to the analysis above, including finding an observed significance level (p-value) and constructing a confidence interval for the population mean. The following table summarizes the cataract surgery example:

Small sample test of $\mu = 0.75$ versus $\mu \neq 0.75$:

$\overline{x}$	n	s	se	95% CI	t-Statistic	p-Value
0.666	24	0.246	0.050	[0.562,0.770]	–1.673	0.1079

In this table, $se = s/\sqrt{24} = 0.050$ is the estimated standard error of the mean, the confidence interval was computed using the methods of the last section, and the computer was used to find the observed significance level.

The fact that the observed significance level is greater than 0.05 confirms our conclusion that the test result is not statistically significant at the 5% level. The fact that the hypothesized mean of 0.75 is in the confidence interval tells us that we cannot rule out the possibility that the popultion mean cataract surgery time is 0.75 hours.

8.3 Analyses of Two Samples

This section considers methods for estimating the difference in population means or testing the equality of population means using information from two samples. The two samples can either be *paired* or independently chosen, and the methods depend on the sampling scheme.

Paired samples analysis. In paired samples analysis, for each observation in the first sample there is a corresponding observation in the second sample. We think of the observations as a list of pairs,

$$(x_{1i}, x_{2i}) \text{ for } i = 1, 2, \ldots, n,$$

and we examine differences of the form

$$d_i = x_{1i} - x_{2i} \text{ for } i = 1, 2, \ldots, n.$$

For example, suppose that we are interested in determining if two different methods for measuring an individual's body fat are equivalent. For individual i,

x_{1i} is the body fat measurement using the first method, and

x_{2i} is the body fat measurement using the second method.

The mean of the first measurements, $\overline{x}_1$, can be used to answer questions about the population mean of measurements using the first method (μ_1), the mean of the second measurements, $\overline{x}_2$, can be sued to answer questions about the population mean of measurements using the second method (μ_2), and the mean of the differences, $\overline{d}$, can be used to answer questions about the difference in population means ($\mu_1 - \mu_2$).

Independent samples analysis. In independent samples analysis, measurements are made on n_1 individuals from the first population,

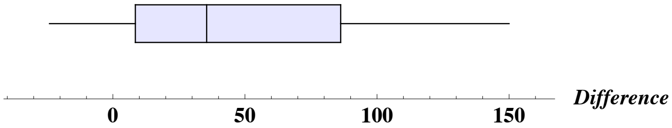
$$x_{1i} \text{ for } i = 1, 2, \ldots, n_1,$$

and on n_2 individuals independently chosen from the second population,

$$x_{2i} \text{ for } i = 1, 2, \ldots, n_2.$$

The difference in sample means ($\overline{x}_1 - \overline{x}_2$) can be used to answer questions about the difference in population means ($\mu_1 - \mu_2$).

Figure 8.4: *Summaries of average reaction time differences for 32 students.*



Q1: 8.5 msecs	*Median:* 35.5 msecs	*Mean:* 50.625 msecs
Q2: 35.5 msecs	*IQR:* 77.75 msecs	*SD:* 52.486 msecs
Q3: 86.25 msecs	*Outliers:* none	*Sample Size:* 32

8.3.1 Methods for Paired Samples

Let $d_i = x_{1i} - x_{2i}$, for $i = 1, 2, \ldots, n$, be the list of differences,

$$\overline{d} = \overline{x}_1 - \overline{x}_2 \text{ be the sample mean of differences,}$$

and s be the sample standard deviation of differences, based on a simple random sample of size n individuals from a large population.

Confidence interval procedure. If the distribution of differences is approximately normally distributed with mean $\mu_1 - \mu_2$, then a $100(1-\alpha)\%$ confidence interval for the difference in population means is

$$\overline{d} \pm t_{\alpha/2}(s/\sqrt{n}),$$

where $t_{\alpha/2}$ is the critical value for the Student t distribution with $n-1$ degrees of freedom.

Hypothesis test procedure. If the distribution of differences is approximately normally distributed with mean $\mu_1 - \mu_2$, then a test of the null hypothesis that $\mu_1 = \mu_2$ versus the alternative that the means are not equal, equivalently, a test of

$$H_0: \ \mu_1 - \mu_2 = 0 \ \text{ versus } \ H_A: \ \mu_1 - \mu_2 \neq 0,$$

can be conducted using the statistic

$$t = \overline{d}/(s/\sqrt{n}) = (\overline{x}_1 - \overline{x}_2)/(s/\sqrt{n}).$$

The rejection region for a two-sided test conducted at the $100\alpha\%$ level is

$$|t| \geq t_{\alpha/2},$$

where $t_{\alpha/2}$ is the critical value for the Student t distribution with $n-1$ degrees of freedom.

The following example will be used to illustrate both procedures.

Example 8.4 (Cell Phones and Driving) (Agresti & Franklin, 2007, page 459) As part of a study on whether the use of cell phones impairs reaction times in a driving skills test, researchers at the University of Utah gathered information on a sample of 32 students. Each student was asked to operate a machine that simulated different driving situations.

At irregular periods a target flashed red or green, and the student was instructed to press a "break button" as soon as possible after seeing a red light. During a warmup period, students were allowed to become familiar with the task without using a cell phone. During the trial period, students were required to perform the task while carrying out a conversation about a political issue on the cell phone with someone in an adjoining room.

The following table gives the average reaction time while using a cell phone, the average reaction time while not using a cell phone, and the difference in times, defined as

$$\text{Difference} = \text{Average Time with Phone} - \text{Average Time w/o Phone},$$

for each student, where times are measured in milliseconds (msecs).

Average Reaction Times and Differences ($n = 32$)*:*

With Cell Phone	595	543	520	501	536	468	527	456	525	609	573
Without Cell Phone	619	556	531	508	543	470	521	448	515	599	554
Difference	−24	−13	−11	−7	−7	−2	6	8	10	10	19

With Cell Phone	482	559	542	565	636	558	574	600	578	623	560
Without Cell Phone	462	537	513	536	604	519	529	544	512	556	487
Difference	20	22	29	29	32	39	45	56	66	67	73

With Cell Phone	615	554	554	679	626	688	601	960	647	672	
Without Cell Phone	540	470	467	589	525	558	459	814	499	522	
Difference	75	84	87	90	101	130	142	146	148	150	

Each column in the table corresponds to the information for one student. Students are ordered by increasing differences.

The mean reaction time during the test period was $\overline{x}_1 = 585.187$ milliseconds, the mean reaction time during the warmup period was $\overline{x}_2 = 534.562$ milliseconds, and the difference in mean reaction times was

$$\overline{d} = \overline{x}_1 - \overline{x}_2 = 585.187 - 534.562 = 50.625.$$

Additional summaries of the differences data are given in Figure 8.4 (page 171). Since the box plot of differences is approximately symmetric with no outliers, we have no reason to doubt that the population distribution of differences in reaction times is approximately normally distributed.

Consider estimating the difference in population means with 99% confidence. Since $n = 32$, the critical value is 2.744 (the entry for the row corresponding to 31 degrees of freedom and the column corresponding to 99% confidence in Table 8.1, page 165).

The confidence interval is computed as follows:

$$50.625 \pm 2.744(52.486/\sqrt{32}) \ \Rightarrow\ 50.625 \pm 25.459 \ \Rightarrow\ [25.166, 76.084].$$

Thus, with 99% confidence we believe that mean reaction times increase by between about 25.2 and 76.1 milliseconds when cell phones are used.

Next, consider testing

$$\mu_1 = \mu_2 \text{ versus } \mu_1 \neq \mu_2 \text{ at the 1\% significance level,}$$

where μ_1 is the mean reaction time for students engaged in political discussions on a cell phone while taking the driving skills test, and μ_2 is the mean reaction time for students who do not use a cell phone while taking the driving skills test. Equivalently, consider testing whether the difference in means, $\mu_1 - \mu_2$, is equal to 0 or not. The rejection region for the test is $|t| \geq 2.744$ (since the cutoff is the same for both 99% confidence and 1% significance).

Since the value of the test statistic,

$$t = 50.625/(52.486/\sqrt{32}) = 5.456,$$

is in the rejection region, the results are statistically significant at the 1% level. Further, since the value of t is positive, we have reason to believe that the mean reaction time when using cell phones is greater than when not using cell phones.

The following table summarizes the test and confidence interval results, and includes the value of the observed significance level (p-Value):

Paired small samples test of $\mu_1 - \mu_2 = 0$ versus $\mu_1 - \mu_2 \neq 0$:

$\overline{x}_1 - \overline{x}_2$	n	s	se	99% CI	t-Statistic	p-Value
50.625	32	52.486	9.278	[25.166,76.084]	5.456	0.000

In this table, $se = s/\sqrt{32} = 9.278$ is the estimated standard error of the sample mean difference, and the computer was used to find the observed significance level. Since the entire 99% confidence interval for the difference in means is above 0, it is not surprising that the observed significance level was below 0.01 (in this case, the observed significance level was 0 to three decimal places of accuracy).

Footnote: Paired t methods. Small sample methods based on the Student t distribution applied to differences data are often called “paired t methods.”

8.3.2 Methods for Independent Samples

Let $\overline{x}_i$ be the sample mean and s_i be the sample standard deviation of a simple random sample of size n_i chosen from a large population with mean μ_i, for $i = 1, 2$.

Degrees of freedom in the independent samples setting. The degrees of freedom will be calculated using the following formula:

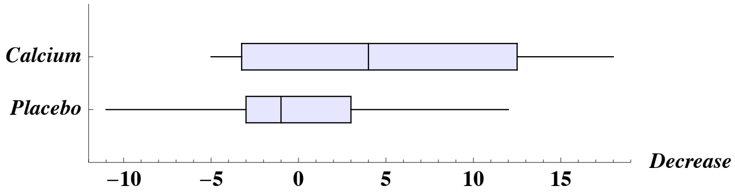
$$df = \frac{\left((s_1^2/n_1) + (s_2^2/n_2)\right)^2}{((s_1^2/n_1)^2/(n_1-1)) + ((s_2^2/n_2)^2/(n_2-1))},$$

where the value is rounded to the closest whole number.

Confidence interval procedure. If population distributions are approximately normally distributed, then an approximate $100(1-\alpha)\%$ confidence interval for the difference in population means, $\mu_1 - \mu_2$, can be computed using the following formula,

$$\overline{x}_1 - \overline{x}_2 \pm t_{\alpha/2}\sqrt{(s_1^2/n_1) + (s_2^2/n_2)},$$

Figure 8.5: *Summaries of decreases in blood pressure for men in 2 treatment groups.*



Calcium	*Q1:* −3.25 mmHg	*Median:* 4.0 mmHg	*Mean:* 5.0 mmHg
Group:	*Q2:* 4.0 mmHg	*IQR:* 15.75 mmHg	*SD:* 8.743 mmHg
	Q3: 12.5 mmHg	*Outliers:* none	*Sample Size:* 10
Placebo	*Q1:* −3.0 mmHg	*Median:* −1.0 mmHg	*Mean:* −0.636 mmHg
Group:	*Q2:* −1.0 mmHg	*IQR:* 6.0 mmHg	*SD:* 5.870 mmHg
	Q3: 3.0 mmHg	*Outliers:* none	*Sample Size:* 11

where $t_{\alpha/2}$ is the critical value for the Student t distribution with *df* degrees of freedom, and *df* is computed using the formula above.

Hypothesis test procedure. If the population distributions are approximately normally distributed, then a test of the null hypothesis that $\mu_1 = \mu_2$ versus the alternative that the means are not equal, equivalently a test of

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{versus} \quad H_A : \mu_1 - \mu_2 \neq 0,$$

can be conducted using the statistic

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}.$$

The rejection region for a two-sided test conducted at the $100\alpha\%$ level is

$$|t| \geq t_{\alpha/2},$$

where $t_{\alpha/2}$ is the critical value for the Student t distribution with *df* degrees of freedom, and *df* is computed using the formula above.

The following example will be used to illustrate both procedures.

Example 8.5 (Calcium and Blood Pressure) (Moore & McCabe, 1999, page 551) After examining the results of a large observational study — which seemed to suggest that calcium supplements could reduce blood pressure in adult men — researchers designed and carried out a double-blind, placebo-controlled study of calcium supplements using a sample of 21 male subjects.

The men were randomly assigned to two treatment groups: 10 men used the calcium supplement each day for twelve weeks, and the remaining 11 men used a placebo (which looked identical to the calcium pill) each day for twelve weeks. Each man's seated systolic blood pressure in millimeters of mercury (mmHg) was measured at the beginning and again at the end of the study period, and the decrease in blood pressure, defined as

$$\text{Decrease} = \text{Beginning Blood Pressure} - \text{Ending Blood Pressure},$$

was used in analyses. The results are summarized in the following tables, where a negative decrease is equivalent to an *increase* in seated SBP over the study period:

Calcium Group ($n = 10$)*:*

Begin	111	110	112	102	107	107	112	136	129	123
End	116	114	115	104	106	100	102	125	112	105
Decrease	−5	−4	−3	−2	1	7	10	11	17	18

Placebo Group ($n = 11$)*:*

Begin	110	114	102	130	112	112	117	123	98	119	109
End	121	119	105	133	114	113	118	124	95	114	97
Decrease	−11	−5	−3	−3	−2	−1	−1	−1	3	5	12

Each column in the first table corresponds to the information for one patient in the calcium group, and each column in the second table corresponds to the information for one patient in the placebo group.

The mean decrease for men in the calcium group was $\overline{x}_1 = 5$, the mean decrease for men in the placebo group was $\overline{x}_2 = -0.636$, and the difference in sample means was

$$\overline{x}_1 - \overline{x}_2 = 5 - (-0.636) = 5.636 \ \text{ mmHg.}$$

Additional summaries of the two samples are given in Figure 8.5 (page 174). Since the box plots are approximately symmetric with no outliers, we have no reason to doubt that the population distributions are approximately normally distributed.

Consider estimating the difference in population means with 95% confidence. Since the degrees of freedom is

$$df = \frac{\left(\left(8.743^2/10\right) + \left(5.870^2/11\right)\right)^2}{((8.743^2/10)^2/(10-1)) + ((5.870^2/11)^2/(11-1))} = 15.5391 \ \rightarrow \ 16,$$

the critical value is 2.120 (the entry corresponding to 16 degrees of freedom and the column corresponding to 95% confidence in Table 8.1, page 166), and the estimated standard error of the difference in sample means is

$$\sqrt{8.743^2/10 + 5.870^2/11} = 3.283.$$

The confidence interval is computed as follows:

$$5.636 \pm (2.120)(3.283) \ \Rightarrow \ 5.636 \pm 6.960 \ \Rightarrow \ [-1.324, 12.596].$$

Notice that the lower end of the confidence interval is below 0 and the upper end is above 0. Since 0 is in the interval, we cannot rule out the possibility that the means are actually equal.

Next, consider testing

$$\mu_1 = \mu_2 \text{ versus } \mu_1 \neq \mu_2 \text{ at the 5\% significance level,}$$

where μ_1 is the population mean decrease in blood pressure over a 12-week period for men who use calcium supplements and μ_2 is the population mean decrease in blood pressure over a 12-week period for men who do not take calcium supplements. Equivalently, consider testing whether the difference in means, $\mu_1 - \mu_2$, is equal to 0 or not. The rejection region for the test is $|t| \geq 2.120$ (since the cutoff is the same for both 95% confidence and 5% significance).

Since the value of the test statistic,

$$t = 5.636/3.283 = 1.717,$$

is in the acceptance region, the results are not statistically significant at the 5% level. We believe that the observed difference could have occurred by chance.

The following table summarizes the test and confidence interval results, and includes the value of the observed significance level (p-Value):

Independent small samples test of $\mu_1 - \mu_2 = 0$ versus $\mu_1 - \mu_2 \neq 0$:

$\overline{x}_1 - \overline{x}_2$	df	se	95% CI	t-Statistic	p-Value
5.636	17	3.283	$[-1.324, 12.596]$	1.717	0.1053

The computer was used to find the observed significance level.

Footnote: Welch t methods. The degrees of freedom formula for analyzing independently chosen samples from distributions that are approximately normally distributed is due to the British statistician and educator Bernard Lewis Welch (1911-1989), and the methods themselves are often referred to as the "Welch t methods."

8.3.3 Paired Samples versus Independent Samples

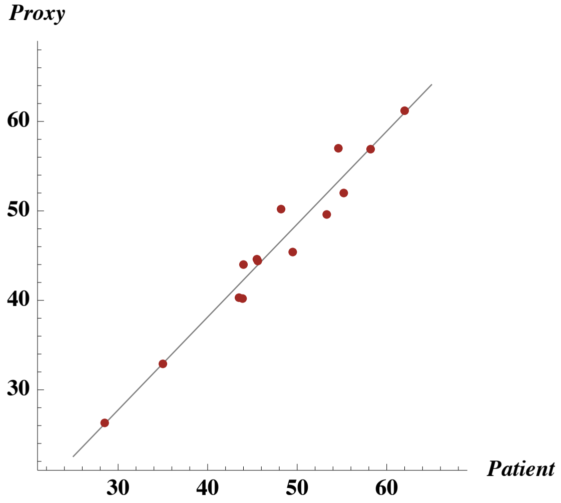
Researchers use paired samples designs in order to reduce the variability (and increase the power of detecting a difference if there truly is one) when answering questions about the difference in population means. The following example illustrates this idea.

Example 8.6 (Difference in Mean AMPAC Scores) The Activity Measure for Post-Acute Care (AMPAC) is a patient reported health survey measuring 3 physical domains. Scale scores are generated based on a patient's own responses to the survey items. However, when patients cannot provide their own data due to cognitive limitations, proxy reports from family members or clinicians are used. An interesting question is to compare the proxy reported scores to the patient recorded scores.

In one study, researchers examined the agreement between patient reported scores and proxy reported scores in the *AMPAC-Movement* domain from a simple random sample of 14 patients. Each patient was scored twice: the first score was based on the patient's own report and the second was based on the proxy report.

Numerical and graphical summaries are given in Figure 8.6, page 177. Notice, in particular, that there is a strong positive association between the patient scores and the proxy scores.

Figure 8.6: *Summaries of AMPAC-Movement scores for 14 patients.*



Patient Scores:
$\overline{x}_1 = 47.643$, $s_1 = 8.937$

Proxy Scores:
$\overline{x}_2 = 46.01$, $s_2 = 9.495$

Correlation between Scores:
$r = 0.9775$

Difference in Scores ($d_i = x_{1i} - x_{2i}$)*:*
$\overline{d} = 1.571$, $s = 2.032$

Consider testing the null hypothesis that the population mean scores are equal versus the alternative that they are not equal at the 5% significance level.

- Using paired samples methods (based on the differences data), the estimated standard error is $se = 0.543$, the value of the test statistic is $t = 2.893$ and the observed significance level (based on 13 degrees of freedom) is 0.0126.
- If the data had been generated from independent samples of patients and proxies, the estimated standard error would be $se = 3.485$, the value of the test statistic would be $t = 0.451$ and the observed significance level (based on 26 degrees of freedom) is 0.6557.

By using the paired samples design in a setting where there is a strong positive association between the scores (since patient scores and proxy scores are expected to be closely linked), the researchers were able to show a significant difference between mean scores reported by patients and their proxies.

Footnote. The scatter plot of paired scores in the figure above includes a least squares regression line. Least squares regression lines were introduced in Section 3.7 (page 45).

8.4 Brief Summary and Additional Examples

This chapter introduces small sample methods for answering questions about population means (and about differences in population means) when the population distributions of the characteristics of interest are approximately normally distributed.

Population distribution versus sampling distribution. It is worth reviewing the difference between population and sampling distributions, especially in the context of the work we have done in the last few chapters.

Population distributions allow us to model values for *individuals* drawn from a population, while sampling distributions allow us to model summaries of values for *samples* drawn from a population. The normal distribution can be used in both roles:

- In Chapter 5, the normal distribution was used as a population distribution to model the heights of adult women in the United States, and to model the heights of adult men in the United States. Tables 5.1-5.2 (pages 109-110) were used to find probabilities associated with each model.
- In Chapters 6 and 7, the normal distribution was used as an approximate sampling for summaries of single samples ($\overline{x}$, $\widehat{p}$, $\widehat{\lambda}$) and for differences of summaries of independently chosen samples ($\overline{x}_1 - \overline{x}_2$, $\widehat{p}_1 - \widehat{p}_2$, $\widehat{\lambda}_1 - \widehat{\lambda}_2$) when sample sizes are large. Table 6.1 (page 124) was used to find the critical values needed to answer questions about the corresponding population parameters or differences in population parameters.

In each example in this chapter, there is an implicit assumption that the normal distribution can be used to model values for individuals drawn from a population (or populations). That is, an implicit assumption that the normal distribution is the population distribution for the characteristic of interest. The Student t distribution is the correct sampling distribution in these situations. The Student t distribution is actually a family of distributions (indexed by the degrees of freedom *df*), and the information in Tables 8.1-8.2 (pages 165-166) allows us to find the critical values needed to answer questions about population means.

Symmetric versus asymmetric box plots. A quick check of the assumption that the normal distribution can be used to model values for individuals drawn from a population is to examine box plots of sample values. If the box plots are reasonably symmetric with few (or no) outliers, then you can proceed to use the methods of this chapter to answer questions about population means.

Note that if a box plot is extremely asymmetric with extreme outliers, then the median may be a better measure of the center of a population distribution than the mean. An introduction to methods focusing on the median instead of the mean is given in the next chapter.

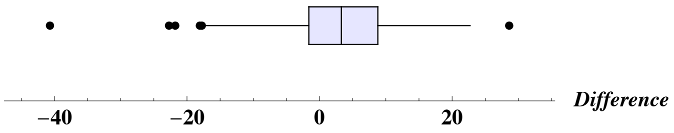
More on paired samples. The paired samples design for answering questions about the difference in population means was introduced in this chapter in the small sample setting. The following example illustrates the paired samples design in a large sample setting.

Example 8.7 (Difference in Mean Daily Intakes of Saturated Fats) (American Journal of Epidemiology (1985) 122:51-65) As part of the *Nurses' Health Study*,[1] a simple random sample of 173 nurses participated in a study designed to determine if dietary information gathered using food frequency questionnaires (FFQs) was as accurate as information gathered using careful dietary records (DRs) over a 4-month period.

At the end of the study period, the researchers used the FFQs and DRs to determine daily intake of several variables, including total saturated fats, total fat, total alcohol consumption and total calories.

[1] In the late 1970's and early 1980's, more than a hundred thousand female nurses participated in various aspects of the first Nurses' Health Study. The study gathered useful information on many health variables.

Figure 8.7: *Summaries of differences in fat intake measurements for 173 nurses.*



Q1: −1.615 gms	*Median:* 3.290 gms	*Mean:* 3.016 gms
Q2: 3.290 gms	*IQR:* 10.410 gms	*SD:* 9.012 gms
Q3: 8.795 gms	*Outliers:* $-40.67, -22.71, \ldots, 28.61$	*Sample Size:* 173

One question of interest was whether the mean level of daily intake of saturated fats using the accurate DR method (μ_1) was equal to the mean level using the FFQ method (μ_2).

The following table summarizes the information of daily intake of saturated fats in grams, and Figure 8.7 (page 179) gives additional information about the list of differences.

	Sample 1: *(DR Method)*	*Sample 2:* *(FFQ Method)*	*Differences:* *(DR–FFQ)*
Mean:	24.932	21.916	3.016
SD:	6.773	9.275	9.012

Consider estimating the difference in population means with 95% confidence. Since the sample size is large, Table 6.1 (page 124) is used to find the cutoff for the analysis. The confidence interval becomes

$$3.016 \pm (1.960)(9.012/\sqrt{173}) \;\Rightarrow\; 3.016 \pm 1.343 \;\Rightarrow\; [1.673, 4.359].$$

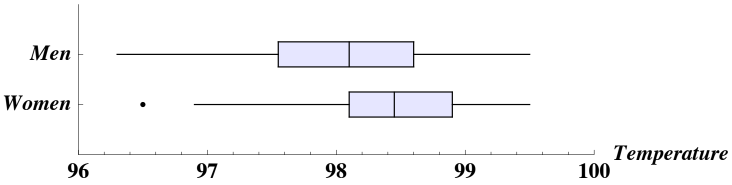
Thus, with 95% confidence, we believe that there is a difference of between about 1.7 and 4.4 grams between mean levels reported using the accurate DR method and the FFQ method.

Finally, note that the box plot of the differences data has several outliers, possibly indicating the FFQ questionnaires were not filled in properly by some study participants. In the large sample setting, the fact that the distribution of differences may not be well approximated by a normal distribution does not affect the conclusions.

Small sample versus large sample methods. There are no strict rules for when sample sizes are considered to be large enough to use large sample methods, although practitioners often use small sample methods when $n \leq 30$, and large sample methods otherwise, when doing calculations by hand. It is interesting to note that for samples of moderate sizes, inferences based on small sample methods are virtually identical to inferences based on large sample methods. The following example illustrates this point.

Example 8.8 (Equality of Mean Body Temperatures) (Shoemaker, *Journal of Statistics Education*, Volume 4, Number 2 (1996), `www.amstat.org/publications/jse/v4n2`) As

Figure 8.8: *Summaries of body temperatures for 2 groups of healthy adults.*



Men:	*Q1:* 97.55 deg	*Median:* 98.10 deg	*Mean:* 98.109 deg
	Q2: 98.10 deg	*IQR:* 1.05 deg	*SD:* 0.748 deg
	Q3: 98.6 deg	*Outliers:* none	*Sample Size:* 45
Women:	*Q1:* 98.10 deg	*Median:* 98.45 deg	*Mean:* 98.428 deg
	Q2: 98.45 deg	*IQR:* 0.80 deg	*SD:* 0.632 deg
	Q3: 98.90 deg	*Outliers:* 96.5	*Sample Size:* 40

part of a study of factors affecting body temperatures in healthy adults, researchers collected information from more than one hundred individuals.

Graphical and numerical summaries for 45 men (*Sample 1*) and 40 women (*Sample 2*) who participated in the study are given in Figure 8.8 (page 180).

Consider testing the null hypothesis that the mean body temperatures of healthy men and women are equal versus the alternative that they are not equal at the 5% significance level.

The following tables summarize the results using both small and large sample methods.

Independent small sample test of $\mu_1 - \mu_2 = 0$ *versus* $\mu_1 - \mu_2 \neq 0$*:*

$\overline{x}_1 - \overline{x}_2$	df	*se*	95% CI	t-Statistic	p-Value
-0.319	83	0.150	$[-0.617, -0.020]$	-2.129	0.036

Independent large sample test of $\mu_1 - \mu_2 = 0$ *versus* $\mu_1 - \mu_2 \neq 0$*:*

$\overline{x}_1 - \overline{x}_2$	*se*	95% CI	z-Statistic	p-Value
-0.319	0.150	$[-0.612, -0.025]$	-2.129	0.033

The difference in sample means, the estimated standard errors and the values of the test statistics are the same in both tables. Since the cutoff for small sample analysis when $df = 83$ is slightly larger than the cutoff for large sample analysis ($1.989 > 1.960$), the first confidence interval is slightly wider than the second and the first p-value is slightly larger than the second.

In both cases, the evidence suggests that the mean body temperature for healthy adult men is slightly smaller than the mean for healthy adult women.

Statistical significance versus practical importance, revisited. As mentioned previously, many people mistakenly believe that

1. "statistically significant" is synonymous with "important" and
2. "not statistically significant" is synonymous with "unimportant"

when interpreting the results of hypothesis tests. In fact, an important result is one that has substantive or practical significance to researchers.

The example on page 150 illustrated a large sample test of $\mu = 211$ versus $\mu \neq 211$ that was statistically significant at all the usual significance levels even though the observed difference,

$$\overline{x} - \mu_0 = 215.1 - 211 = 4.1,$$

was of no practical importance. (The very large sample size made the large sample test sensitive to a very small difference in means.)

We now have the tools to examine a situation at the other extreme. Namely, a situation where a difference of practical importance was not statistically significant due to a very small sample size. For example, suppose that the researchers from the example on page 150 decided to sample 10 men (instead of the original sample size of 2500 men), with results as follows:

Small sample test of $\mu = 211$ versus $\mu \neq 211$:

$\overline{x}$	n	s	se	t-Statistic	p-Value
241.1	10	57.21	18.091	1.664	0.130

In this case, the results are not statistically significant when $\alpha < 0.130$, but the observed difference between the sample mean and the hypothesized population mean,

$$\overline{x} - \mu_0 = 241.1 - 211 = 30.1$$

is large enough to be of interest to researchers.

When designing studies, researchers need to be mindful that the sample sizes they choose are large enough to detect important differences with high power. In some situations, like the aspirin and heart attacks example (beginning on page 153) from the last chapter, very large samples were needed because the small expected difference was of practical importance. In other situations, small to moderate sized samples are sufficient, and methods like the ones presented in this chapter can be used to analyze the data.

9 Introduction to Nonparametric Analysis

This chapter introduces additional methods for two samples that do not require the assumption that population distributions are approximately normally distributed. The methods are part of a broad field within statistical inference known as *nonparametric analysis*. References for this chapter include the texts by Agresti & Franklin (2007, Chapter 14), Baldi & Moore (2009, Chapter 27), Freedman et al (1991, Parts VII–VIII), Moore & McCabe (1999, Chapter 14) and Pagano & Gauvreau (2000, Chapter 13).

9.1 Signed Rank Test

This section introduces a method for analyzing differences data when the distribution of differences is symmetric and the goal is to determine if the difference in population *medians* is zero. The method can be applied to quantitative data from continuous distributions, quantitative data from discrete distributions, and ordinal data.

Signed rank statistics. The key to this approach is to replace the absolute value of each difference by its *rank* (or position) in an ordered list, and to work with either

$$S_+ = \text{Sum of Ranks for Positive Differences, or}$$

$$S_- = \text{Sum of Ranks for Negative Differences.}$$

Example 9.1 ($n = 5$) For example, if $n = 5$ and the differences are $-12, -1, 4, 5, 17$, then the following table shows the ordered absolute differences and corresponding ranks,

	−	+	+	−	+
Absolute Differences:	1	4	5	12	17
Ranks of Absolute Differences:	1	2	3	4	5

and the values of the statistics are: $S_+ = 2 + 3 + 5 = 10$ and $S_- = 1 + 4 = 5$.

Example 9.2 ($n = 9$) Similarly, if $n = 9$ and the differences are

$$-39, -29, -23, -17, -13, 1, 2, 10, 36,$$

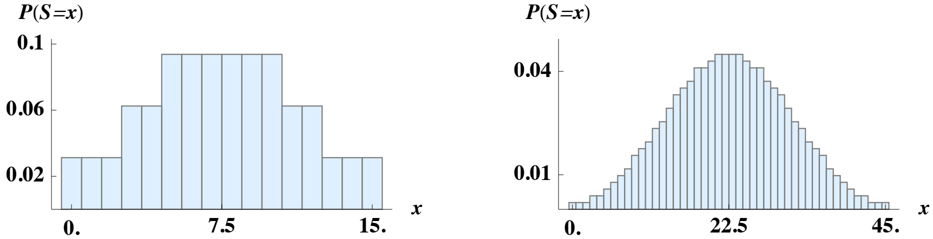
then the following table shows the ordered absolute differences and corresponding ranks,

	+	+	+	−	−	−	−	+	−
Absolute Differences:	1	2	10	13	17	23	29	36	39
Ranks of Absolute Differences:	1	2	3	4	5	6	7	8	9

and the values of the statistics are: $S_+ = 1 + 2 + 3 + 8 = 14$ and $S_- = 4 + 5 + 6 + 7 + 9 = 31$.

Zeros and tied absolute differences. If an observed difference is exactly zero, then that observation drops out of the analysis and the sample size is reduced by 1. If two or more absolute differences are equal, then the average of their positions in the ordered list is used as the rank for each observation.

Figure 9.1: *Probability histograms for signed rank statistics.*



(a) *Distribution when* $n = 5$

(b) *Distribution when* $n = 9$

Example 9.3 (Zeros and Ties) For example, if the differences are

$$-13, -3, -2, 0, 3, 6, 7, 7,$$

then we would work with the $n = 7$ nonzero differences only. The following table shows the absolute differences and corresponding ranks, where the two 3's are each assigned the rank of $(2 + 3)/2 = 2.5$ and the two 7's are each assigned the rank $(5 + 6)/2 = 5.5$:

	$-$	$-$	$+$	$+$	$+$	$+$	$-$
Absolute Differences:	2	3	3	6	7	7	13
Ranks of Absolute Differences:	1	2.5	2.5	4	5.5	5.5	7

The statistic values are

$$S_+ = 2.5 + 4 + 5.5 + 5.5 = 17.5 \text{ and } S_- = 1 + 2.5 + 7 = 10.5.$$

Sampling distributions. Let n be the number of nonzero differences, and assume that there are no tied values. Then the sampling distributions of S_+ and S_- are obtained by assigning "±" signs to the whole numbers between 1 and n in all possible ways, and computing the values of the statistics in each case. Each assignment of signs to ranks is assumed to be equally likely.

Formulas for summary measures. Under this scheme, both S_+ and S_- have the same sampling distribution, and either statistic can be used for two-sided inference. If S is either S_+ or S_-, then

$$E(S) = n(n + 1)/4 \text{ and } SD(S) = \sqrt{n(n + 1)(2n + 1)/24},$$

and S takes whole number values between 0 and $n(n + 1)/2$.

Example 9.4 (Sampling Distributions) Figure 9.1(a) (page 184) is a probability histogram for S when $n = 5$. Values of S range from 0 to 15, with the following summary measures:

$$E(S) = 5(6)/4 = 7.5 \text{ and } SD(S) = \sqrt{5(6)(11)/24} = 3.71.$$

Table 9.1: *Critical values for two-sided signed rank tests when there are no (or few) ties in the absolute differences. In the table, n is the number of nonzero differences.*

	Two-Sided Significance Level as a Percent: $100\alpha\%$				
	20%	10%	5%	1%	0.1%
n	$s^*_{\alpha/2}$	$s^*_{\alpha/2}$	$s^*_{\alpha/2}$	$s^*_{\alpha/2}$	$s^*_{\alpha/2}$
4	1.461	1.826			
5	1.483	1.753	2.023		
6	1.363	1.782	1.992	2.201	
7	1.352	1.690	2.028	2.366	
8	1.400	1.680	1.960	2.521	
9	1.362	1.718	1.955	2.429	2.666
10	1.376	1.682	1.988	2.497	2.803
11	1.334	1.689	1.956	2.490	2.934
12	1.334	1.726	1.961	2.510	2.981
13	1.363	1.712	1.992	2.481	2.970
14	1.350	1.664	1.977	2.480	3.045
15	1.306	1.704	1.988	2.499	3.067
16	1.344	1.655	1.965	2.534	3.051
17	1.302	1.681	1.965	2.533	3.053
18	1.328	1.677	1.982	2.504	3.070
19	1.328	1.650	1.972	2.535	3.099
20	1.307	1.680	1.979	2.539	3.099
21	1.303	1.651	1.964	2.520	3.111
22	1.315	1.672	1.964	2.516	3.133
23	1.308	1.673	1.977	2.524	3.133
24	1.314	1.657	1.971	2.543	3.143
25	1.305	1.655	1.951	2.543	3.135
26	1.308	1.664	1.968	2.527	3.137
27	1.297	1.658	1.970	2.523	3.147
28	1.298	1.662	1.958	2.528	3.165
29	1.308	1.654	1.957	2.541	3.168
30	1.306	1.656	1.964	2.540	3.157
31	1.293	1.666	1.960	2.548	3.175
32	1.290	1.664	1.963	2.543	3.179
33	1.295	1.653	1.957	2.546	3.172
34	1.291	1.650	1.958	2.539	3.189
35	1.294	1.654	1.965	2.539	3.178
36	1.304	1.650	1.964	2.545	3.189
37	1.290	1.652	1.954	2.542	3.191
38	1.298	1.661	1.965	2.545	3.198
39	1.298	1.661	1.954	2.540	3.196
40	1.290	1.653	1.962	2.540	3.199
∞	1.282	1.645	1.960	2.576	3.291

Figure 9.1(b) is a probability histogram for S when $n = 9$. In this case, values of S range from 0 to 45, with the following summary measures:

$$E(S) = 9(10)/4 = 22.5 \text{ and } SD(S) = \sqrt{9(10)(19)/24} = 8.44.$$

Notice that each distribution is symmetric around its center, $E(S)$.

Two-sided tests. Suppose that we are interested in testing the null hypothesis that the difference in population medians equals 0 versus the alternative hypothesis that the difference is not equal to 0 using information from a simple random sample of pairs.

If the distribution of differences is symmetric, then the test statistic is

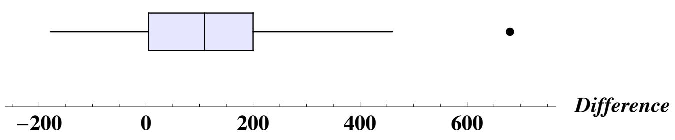
$$Z_S = (S - E(S))/SD(S),$$

where S is either S_+ or S_-, and the summary measures are

$$E(S) = n(n+1)/4 \text{ and } SD(S) = \sqrt{n(n+1)(2n+1)/24}$$

when n is the number of nonzero differences and there are no ties (or few ties) among the absolute differences.

Figure 9.2: *Differences in reduced lung functions for 14 cystic fibrosis patients.*



Q1: 4.5 ml	*Median:* 109.5 ml	*Mean:* 136.5 ml
Q2: 109.5 ml	*IQR:* 195.5 ml	*SD:* 223.175 ml
Q3: 200.0 ml	*Outliers:* 680	*Sample Size:* 14

The rejection region for a two-sided test conducted at the 100α% significance level is

$$|Z_S| \geq s^*_{\alpha/2}.$$

Critical values for the test ($s^*_{\alpha/2}$) are given in Table 9.1 (page 185).

For example,

(1) If $n = 12$ and we are interested in conducting a test at the 10% significance level, then the critical value is 1.726.

(2) If $n = 26$ and we are interested in conducting the test at the 5% significance level, then the critical value is 1.968.

For sample sizes greater than 40, critical values on the last row, where n equals "∞" ("infinity"), should be used. Critical values on the last row are the critical values of the standard normal distribution.

Example 9.5 (Amiloride and Cystic Fibrosis) (Pagano & Gauvreau, 2000, page 305) Cystic fibrosis is a genetic disease of the mucus and sweat glands. It affects many organs, including the lungs. In patients with cystic fibrosis, mucus (which is normally watery) becomes thick and sticky. The mucus builds up in the lungs and blocks the airways. It is believed that the drug amiloride may help to improve air flow in the lungs and thereby delay the loss of pulmonary function often associated with the disease.

As part of a study on the use of amiloride as a therapy for patients with cystic fibrosis, information was gathered on a sample of 14 patients. Researchers measured the loss of lung function over a 25-week period using amiloride and the corresponding loss over a 25-week period using a placebo, and considered the differences in these measures. In order to reduce potential biases, the researchers used a double-blind crossover design, where the order of treatment (placebo or amiloride) was determined randomly for each patient and neither the patient nor the physician knew the order of treatment at the time of the study.

To measure lung function, the researchers used forced vital capacity (FVC), which is the volume of air in milliliters (ml) a person can expel from the lungs in 6 seconds. To measure the loss of lung function over a fixed period of time, the researchers used the reduction in forced vital capacity, defined as

$$\text{Reduction in FVC} = \text{Initial FVC} - \text{Final FVC},$$

where "Initial FVC" is the value of FVC at the beginning of the study period and "Final FVC" is the value of FVC at the end of the study period. Finally, to compare the two treatments, the researchers considered differences of the form:

$$\text{Difference} = \text{Reduction on Placebo} - \text{Reduction on Amiloride}.$$

The following table summarizes the results, where each column corresponds to the information for one patient, and patients have been ordered by increasing differences:

Reductions in FVC and Differences ($n = 14$)*:*

Reduction on Placebo	−38	293	80	224	75	541	74	85	−23	525	508	255	525	1023
Reduction on Amiloride	140	445	95	213	33	440	−32	−28	−178	367	323	10	65	343
Difference	−178	−152	−15	11	42	101	106	113	155	158	185	245	460	680

Notice that 3 patients had negative reductions in FVC. Numerical and graphical summaries of the differences data are in Figure 9.2 (page 186).

To determine if amiloride is effective, the researchers tested the null hypothesis that the difference in population medians is 0 versus the alternative hypothesis that the difference in population medians is not 0 at the 5% significance level. Using Table 9.1 (page 185), the rejection region for this test is $|Z_S| \geq 1.977$.

The following table shows absolute differences ($|D|$) and corresponding ranks:

	+	−	+	+	+	+	−	+	+	−	+	+	+	+
$\vert D\vert$:	11	15	42	101	106	113	152	155	158	178	185	245	460	680
Rank:	1	2	3	4	5	6	7	8	9	10	11	12	13	14

The values of the statistics are $S_+ = 86$ and $S_- = 19$. Since $n = 14$,

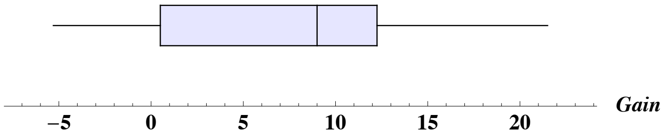
$$E(S) = (14)(15)/4 = 52.5 \text{ and } SD(S) = \sqrt{(14)(15)(29)/24} = 15.93.$$

Using $S = S_+$, the observed value of the test statistic is

$$z_S = (86 - 52.5)/15.93 = 2.103.$$

Since the observed value of the test statistic is in the rejection region for the two-sided test, the results are statistically significant at the 5% level. Further, since the observed median difference (109.5 ml) is positive, there is reason to believe that median lung function loss on placebo is *greater than* median lung function loss on amiloride. (That is, there is reason to believe that the drug is effective in diminishing the loss of pulmonary function.)

Figure 9.3: *Summaries of weight gains for 17 teenage girls with anorexia.*



Q1: 0.5 lbs	*Median:* 9.0 lbs	*Mean:* 7.088 lbs
Q2: 9.0 lbs	*IQR:* 11.75 lbs	*SD:* 7.384 lbs
Q3: 12.25 lbs	*Outliers:* none	*Sample Size:* 17

Footnotes. The analysis in the example above uses differences at two levels: first, to measure the loss in lung function over a 25-week study period; and second, to compare losses under different treatments (either placebo or amiloride). We would expect most differences at the first level to be positive, since patients are expected to lose some lung function over a 25-week period. If amiloride is effective in diminishing the loss of lung function, then we would expect most differences at the second level to be positive.

Signed rank tests can be based on either S_+ or S_-. For example, using $S = S_-$ in the amiloride and cystic fibrosis study, we get

$$z_S = (19 - 52.5)/15.93 = -2.103,$$

which is the negative of the value computed above; our conclusion (that there is reason to believe that observed differences did not occur by chance) remains the same.

Example 9.6 (Anorexia and Family Therapy) (Hand et al, 1994, page 229) Anorexia is an eating disorder characterized by abnormally low body weight, low self-esteem, a distorted perception of one's physical appearance, fear of gaining weight, and, in many cases, clinical depression. About 90% of anorexic individuals are girls or women. There is no single known cause of anorexia. But, individuals with anorexia can get better with the help of health care teams that include doctors, nutritionists and therapists.

As part of a study on treatments for anorexia, researchers at a large medical center gathered information on a sample of 17 teenage girls who received a family therapy treatment. Each girl's weight was measured at the beginning of the study period and again at the end of the period. The variable of interest was weight gain, defined as

$$\text{Weight Gain} = \text{Final Weight} - \text{Initial Weight},$$

where weights are measured in pounds (lbs).

The following table gives the initial and final weights, and weight gains, for each girl in the family therapy treatment group. Each column in the table corresponds to the information for one patient. A negative weight gain corresponds to a *loss* of weight over the study period.

Weights and Weight Gains ($n = 17$)*:*

Final Weight	75.2	77.8	76.8	76.7	93.8	91.5	91.7	101.6	92.5	91.9	98.0	94.3	95.2	90.7	95.5	100.3	94.9
Initial Weight	80.5	81.6	79.9	79.6	89.9	86.0	86.0	94.2	83.5	82.5	87.3	83.3	83.8	77.6	82.1	86.7	73.4
Weight Gain	−5.3	−3.8	−3.1	−2.9	3.9	5.5	5.7	7.4	9.0	9.4	10.7	11.0	11.4	13.1	13.4	13.6	21.5

Notice that 4 patients lost weight during the study period. Numerical and graphical summaries of weight gains are given in Figure 9.3 (page 188).

To determine if family therapy is effective, the researchers tested the null hypothesis that the difference in population median weight gains is 0 versus the alternative hypothesis that the difference is not 0 at 5% significance level. Using Table 9.1 (page 185), the rejection region for this test is $|Z_S| \geq 1.965$.

The following table shows absolute differences ($|D|$) and corresponding ranks:

	−	−	−	+	−	+	+	+	+	+	+	+	+	+	+	+	+
$\|D\|$:	2.9	3.1	3.8	3.9	5.3	5.5	5.7	7.4	9.0	9.4	10.7	11.0	11.4	13.1	13.4	13.6	21.5
Rank:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17

The values of the statistics are $S_+ = 142$ and $S_- = 11$. Since $n = 17$,

$$E(S) = (17)(18)/4 = 76.5 \text{ and } SD(S) = \sqrt{(17)(18)(35)/24} = 21.125.$$

Using $S = S_+$, the observed value of the test statistic is

$$z_S = (142 - 76.5)/21.125 = 3.101.$$

Since the observed value of the test statistic is in the rejection region for the two-sided test, the results are statistically significant at the 5% level. Further, since the observed median difference (9.0 lbs) is positive, there is reason to believe that median weight gain for young women who undergo family therapy is positive. (That is, there is reason to believe that family therapy is an effective treatment for anorexia.)

Footnotes. The anorexia and family therapy example gives the results of one of three treatment arms of a randomized comparative experiment on treatments for anorexia. Each girl who participated in the study was randomly assigned to receive either the standard treatment, the family therapy treatment or the cognitive behavioral therapy treatment.

The study had mixed results. Family therapy appeared to be best overall, while hardly distinguishable from cognitive behavioral therapy, as a treatment for anorexia.

9.2 Rank Sum Test

This section introduces a method for independent samples when the goal is to determine if the difference in population *medians* is zero. The method can be applied to quantitative data from continuous distributions, quantitative data from discrete distributions, and ordinal data.

Rank sum statistics. The key to this approach is to replace each observation by its *rank* (or position) in the ordered list of all $n_1 + n_2$ observations, and to work with either

$$R_1 = \text{Sum of Ranks for Observations in Sample 1, or}$$

$$R_2 = \text{Sum of Ranks for Observations in Sample 2.}$$

Example 9.7 ($n_1 = 3$, $n_2 = 4$) For example, if $n_1 = 3$, $n_2 = 4$ and the observations are

$$-2.0,\ 0.9,\ 6.9 \text{ and } -3.9,\ -1.6,\ 1.4,\ 5.7,$$

then the following table shows the ordered list of 7 observations and corresponding ranks,

	2	1	2	1	2	2	1
Obstn:	−3.9	−2.0	−1.6	0.9	1.4	5.7	6.9
Rank:	1	2	3	4	5	6	7

and the values of the statistics are $R_1 = 2 + 4 + 7 = 13$ and $R_2 = 1 + 3 + 5 + 6 = 15$.

Example 9.8 ($n_1 = 7$, $n_2 = 5$) Similarly, if $n_1 = 7$, $n_2 = 5$ and the observations are

$$-3.8,\ 0,\ 0.8,\ 1.4,\ 7.7,\ 9,\ 9.3 \text{ and } -2.8,\ -1.8,\ -1.5,\ 5.2,\ 5.8,$$

then the following table shows the ordered list of 12 observations and corresponding ranks,

	1	2	2	2	1	1	1	2	2	1	1	1
Obstn:	−3.8	−2.8	−1.8	−1.5	0	0.8	1.4	5.2	5.8	7.7	9.0	9.3
Rank:	1	2	3	4	5	6	7	8	9	10	11	12

and the values of the statistics are:

$$R_1 = 1 + 5 + 6 + 7 + 10 + 11 + 12 = 52 \text{ and } R_2 = 2 + 3 + 4 + 8 + 9 = 26.$$

Tied observations. If two or more observations are equal, then the average of their positions in the ordered list is used as the rank for each observation.

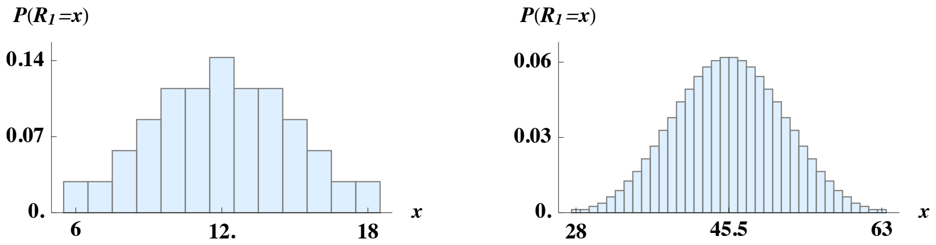
Example 9.9 (Ties) For example, if $n_1 = 6$, $n_2 = 8$ and the observations are

$$-1.5,\ -0.7,\ 1.3,\ 1.3,\ 2.0,\ 3.7 \text{ and } -2.0,\ -1.5,\ -0.9,\ 1.3,\ 2.1,\ 2.5,\ 3.9,\ 4.8,$$

then the following table shows the ordered list of 14 observations and corresponding ranks, where the two -1.5's are each assigned the rank of $(2 + 3)/2 = 2.5$, and the three 1.3's are each assigned the rank of $(6 + 7 + 8)/3 = 7$:

	2	1	2	2	1	1	1	2	1	2	2	1	2	2
Obstn:	−2.0	−1.5	−1.5	−0.9	−0.7	1.3	1.3	1.3	2.0	2.1	2.5	3.7	3.9	4.8
Rank:	1	2.5	2.5	4	5	7	7	7	9	10	11	12	13	14

Figure 9.4: *Probability histograms for R_1 statistics.*



(a) *Distribution when* $n_1 = 3$, $n_2 = 4$ (b) *Distribution when* $n_1 = 7$, $n_2 = 5$

The values of the statistics are

$$R_1 = 2.5 + 5 + 7 + 7 + 9 + 12 = 42.5 \text{ and } R_2 = 1 + 2.5 + 4 + 7 + 10 + 11 + 13 + 14 = 62.5.$$

Sampling distributions. Let n_1 be the size of Sample 1 and n_2 be the size of Sample 2, and assume that there are no ties in the n_1+n_2 observations. Then the sampling distributions of R_1 and R_2 are obtained by randomly partitioning the whole numbers between 1 and n_1+n_2 into complementary subsets of sizes n_1 and n_2, respectively, and computing the values of the statistics in each case. Each partition is assumed to be equally likely.

Formulas for summary measures. Under this scheme, the sampling distribution of R_1 has summary measures

$$E(R_1) = n_1(n_1 + n_2 + 1)/2 \text{ and } SD(R_1) = \sqrt{n_1 n_2(n_1 + n_2 + 1)/12},$$

and takes whole number values between $n_1(n_1 + 1)/2$ and $n_1 n_2 + n_1(n_1 + 1)/2$.

Similarly, the sampling distribution of R_2 has summary measures

$$E(R_2) = n_2(n_1 + n_2 + 1)/2 \text{ and } SD(R_2) = \sqrt{n_1 n_2(n_1 + n_2 + 1)/12},$$

and takes whole number values between $n_2(n_2 + 1)/2$ and $n_1 n_2 + n_2(n_2 + 1)/2$.

Example 9.10 (Sampling Distributions) Figure 9.4(a) (page 191) is a probability histogram for R_1 when $n_1 = 3$ and $n_2 = 4$. Values of R_1 range from 6 to 18, with the following summary measures:

$$E(R_1) = 3(8)/2 = 12 \text{ and } SD(R_1) = \sqrt{3(4)(8)/12} = 2.828.$$

Figure 9.4(b) is a probability histogram for R_1 when $n_1 = 7$ and $n_2 = 5$. In this case, values of R_1 range from 28 to 63, with the following summary measures:

$$E(R_1) = 7(13)/2 = 45.5 \text{ and } SD(R_1) = \sqrt{7(5)(13)/12} = 6.158.$$

Notice that each distribution is symmetric around its center, $E(R_1)$.

Two-sided tests. Suppose that we are interested in testing the null hypothesis that the difference in population medians equals 0 versus the alternative that the difference is not 0 using information from independent simple random samples. Then the test statistic is

$$Z_R = (R - E(R))/SD(R),$$

where R is either R_1 or R_2. If there are no (or few) ties among the $n_1 + n_2$ observations and we let $R = R_1$, then the summary measures are

$$E(R) = n_1(n_1 + n_2 + 1)/2 \text{ and } SD(R) = \sqrt{n_1 n_2(n_1 + n_2 + 1)/12}.$$

If there are no (or few) ties among the $n_1 + n_2$ observations and we let $R = R_2$, then the summary measures are

$$E(R) = n_2(n_1 + n_2 + 1)/2 \text{ and } SD(R) = \sqrt{n_1 n_2(n_1 + n_2 + 1)/12}.$$

The rejection region for a two-sided test conducted at the $100\alpha\%$ significance level is

$$|Z_R| \geq r^*_{\alpha/2}.$$

Critical values for the test ($r^*_{\alpha/2}$) are given in Tables 9.2–9.3 (pages 193–194).

For example,

(1) If the smaller sample size is 8, the larger sample size is 14 and we are interested in conducting a test at the 10% level, then the critical value is 1.706 (from Table 9.2).

(2) If the smaller sample size is 11, the larger sample size is 18 and we are interested in conducting a test at the 1% level, then the critical value is 2.562 (from Table 9.3).

Critical values on the last row of Table 9.3, where sample sizes equal "∞" ("infinity"), can be used when the smaller sample size is more than 20. Critical values on the last row are the critical values of the standard normal distribution.

Example 9.11 (Low Birthweight And SIDS) (Pagano & Gauvreau, 2000, page 319) As part of a study examining characteristics of low birthweight infants dying of sudden infant death syndrome (SIDS), data were gathered on 16 boys and 11 girls.

The following tables give the ages in days at time of death for each child.

Boys ($n = 16$)*:*

46	52	58	59	77	78	80	81	84	103	114	115	133	134	167	175

Girls ($n = 11$)*:*

53	56	60	60	78	87	102	117	134	160	277

Numerical and graphical summaries of these data are in Figure 9.5 (page 195).

To determine if there are gender differences, the researchers tested the null hypothesis that the difference in median ages at time of death for low birthweight infants dying of SIDS is 0 versus the alternative that the difference is not 0 at the 5% significance level. Using Table 9.3 (page 194), the rejection region for this test is $|Z_R| \geq 1.974$.

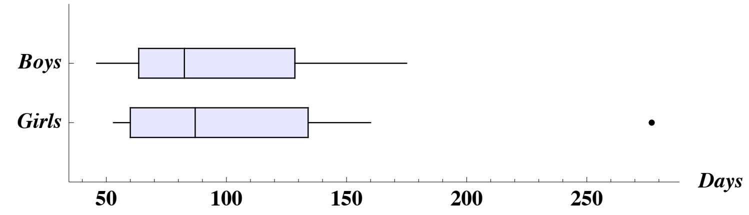
Table 9.2: *Critical values for two-sided rank sum tests when the smaller sample size is between 4 and 8, the larger sample size is at most 20, and there are no (or few) ties in the observations. In the table, n_s is the smaller sample size and n_ℓ is the larger sample size.*

		Two-Sided Significance Level as a Percent: 100α%			
		10%	5%	1%	0.1%
n_s	n_ℓ	$r^*_{\alpha/2}$	$r^*_{\alpha/2}$	$r^*_{\alpha/2}$	$r^*_{\alpha/2}$
4	4	1.732	2.021	2.309	
	5	1.715	1.960	2.449	
	6	1.706	2.132	2.558	
	7	1.701	2.079	2.457	2.646
	8	1.698	2.038	2.548	2.717
	9	1.697	2.006	2.469	2.777
	10	1.697	1.980	2.546	2.828
	11	1.697	1.958	2.481	2.872
	12	1.698	1.940	2.547	2.910
	13	1.698	2.038	2.491	2.944
	14	1.699	2.018	2.443	2.867
	15	1.700	2.000	2.500	2.900
	16	1.701	1.984	2.457	2.929
	17	1.702	1.970	2.508	2.956
	18	1.703	1.958	2.469	2.894
	19	1.703	1.947	2.514	2.920
	20	1.704	2.014	2.479	2.943
5	5	1.776	1.984	2.611	2.611
	6	1.826	2.008	2.556	2.739
	7	1.705	2.030	2.517	2.842
	8	1.757	2.049	2.489	2.928
	9	1.667	2.067	2.467	3.000
	10	1.715	1.960	2.572	2.939
	11	1.756	1.983	2.549	3.002
	12	1.687	2.003	2.530	2.951
	13	1.725	2.021	2.514	3.006
	14	1.666	1.944	2.500	2.963
	15	1.702	1.964	2.488	3.011
	16	1.734	1.982	2.477	2.973
	17	1.684	1.998	2.468	3.016
	18	1.714	2.012	2.534	2.981
	19	1.670	1.955	2.523	3.021
	20	1.698	1.970	2.514	2.989
6	6	1.761	2.082	2.562	2.882
	7	1.714	2.000	2.571	3.000
	8	1.678	2.066	2.582	2.969
	9	1.768	2.003	2.475	3.064
	10	1.735	1.952	2.495	3.037
	11	1.709	2.010	2.513	3.015
	12	1.686	1.967	2.529	3.091
	13	1.666	2.017	2.543	3.070
	14	1.732	1.979	2.557	3.052
	15	1.713	1.946	2.491	3.036
	16	1.696	1.990	2.507	3.023
	17	1.680	1.960	2.521	3.081
	18	1.667	2.000	2.533	3.067
	19	1.654	1.972	2.545	3.054
	20	1.704	1.947	2.495	3.043
7	7	1.725	1.981	2.492	3.003
	8	1.736	1.967	2.546	3.009
	9	1.747	1.958	2.488	3.017
	10	1.659	1.952	2.537	3.025
	11	1.675	1.947	2.491	3.034
	12	1.690	2.028	2.535	3.043
	13	1.704	2.021	2.496	3.051
	14	1.716	2.014	2.537	3.059
	15	1.657	2.009	2.502	3.066
	16	1.670	2.004	2.539	3.074
	17	1.683	2.001	2.509	3.080
	18	1.695	1.997	2.542	3.087
	19	1.705	1.994	2.515	3.093
	20	1.660	1.992	2.545	3.098
8	8	1.680	1.995	2.521	3.046
	9	1.732	2.021	2.502	3.079
	10	1.688	1.955	2.577	3.110
	11	1.651	1.982	2.560	3.055
	12	1.697	2.006	2.546	3.086
	13	1.666	1.955	2.535	3.114
	14	1.706	1.979	2.525	3.140
	15	1.678	2.001	2.517	3.098
	16	1.653	1.960	2.511	3.123
	17	1.689	1.981	2.505	3.088
	18	1.667	2.000	2.556	3.111
	19	1.699	1.965	2.549	3.133
	20	1.678	1.983	2.543	3.102

Table 9.3: *Critical values for two-sided rank sum tests when the smaller sample size is 9 or more, and there are no (or few) ties in the observations. In the table, n_s is the smaller sample size and n_ℓ is the larger sample size.*

		Two-Sided Significance Level as a Percent: $100\alpha\%$			
		10%	5%	1%	0.1%
n_s	n_ℓ	$r^*_{\alpha/2}$	$r^*_{\alpha/2}$	$r^*_{\alpha/2}$	$r^*_{\alpha/2}$
9	9	1.722	1.987	2.517	3.135
	10	1.715	1.960	2.531	3.103
	11	1.709	2.013	2.545	3.077
	12	1.706	1.990	2.558	3.127
	13	1.703	1.970	2.571	3.105
	14	1.701	1.953	2.520	3.150
	15	1.699	1.998	2.534	3.130
	16	1.698	1.981	2.548	3.114
	17	1.698	1.967	2.560	3.153
	18	1.697	1.955	2.520	3.138
	19	1.697	1.992	2.533	3.124
	20	1.650	1.980	2.546	3.158
10	10	1.663	1.965	2.570	3.099
	11	1.690	1.972	2.535	3.098
	12	1.714	1.978	2.572	3.099
	13	1.674	1.985	2.543	3.163
	14	1.698	1.991	2.518	3.162
	15	1.664	1.997	2.552	3.162
	16	1.687	1.950	2.530	3.162
	17	1.657	1.958	2.561	3.163
	18	1.678	1.966	2.541	3.164
	19	1.652	1.973	2.524	3.166
	20	1.672	1.980	2.552	3.168
11	11	1.674	2.003	2.528	3.119
	12	1.662	1.969	2.523	3.139
	13	1.651	1.999	2.520	3.158
	14	1.697	1.971	2.518	3.120
	15	1.687	1.998	2.517	3.140
	16	1.678	1.974	2.566	3.158
	17	1.670	1.952	2.564	3.175
	18	1.663	1.978	2.562	3.146
	19	1.657	1.958	2.561	3.163
	20	1.651	1.982	2.560	3.179
12	12	1.674	1.963	2.540	3.118
	13	1.686	1.958	2.556	3.155
	14	1.697	1.955	2.520	3.138
	15	1.659	1.952	2.537	3.172
	16	1.671	1.996	2.553	3.157
	17	1.683	1.993	2.524	3.144
	18	1.651	1.990	2.540	3.175
	19	1.663	1.987	2.555	3.163
	20	1.674	1.985	2.530	3.153
13	13	1.667	1.974	2.538	3.154
	14	1.650	1.990	2.523	3.154
	15	1.681	1.958	2.557	3.155
	16	1.666	1.973	2.543	3.157
	17	1.653	1.988	2.532	3.160
	18	1.681	1.962	2.562	3.163
	19	1.669	1.976	2.552	3.165
	20	1.658	1.953	2.542	3.169
14	14	1.654	1.976	2.527	3.170
	15	1.658	1.964	2.531	3.186
	16	1.663	1.954	2.536	3.159
	17	1.667	1.985	2.540	3.176
	18	1.671	1.975	2.545	3.191
	19	1.676	1.967	2.550	3.169
	20	1.680	1.960	2.554	3.184
15	15	1.680	1.970	2.551	3.173
	16	1.660	1.976	2.530	3.162
	17	1.680	1.983	2.549	3.191
	18	1.663	1.952	2.567	3.182
	19	1.648	1.960	2.549	3.174
	20	1.667	1.967	2.533	3.200
16	16	1.658	1.960	2.563	3.166
	17	1.657	1.981	2.558	3.170
	18	1.656	1.967	2.553	3.174
	19	1.656	1.954	2.550	3.179
	20	1.655	1.974	2.547	3.184
17	17	1.671	1.981	2.566	3.186
	18	1.650	1.980	2.541	3.201
	19	1.664	1.980	2.551	3.185
	20	1.676	1.981	2.560	3.200
18	18	1.677	1.962	2.563	3.195
	19	1.671	1.975	2.552	3.191
	20	1.666	1.959	2.543	3.187
19	19	1.650	1.971	2.555	3.197
	20	1.658	1.967	2.557	3.203
20	20	1.650	1.975	2.543	3.192
∞	∞	1.645	1.960	2.576	3.291

Figure 9.5: *Summaries of age at death for 2 groups of low birthweight SIDS babies.*



Boys:	*Q1:* 63.5 days	*Median:* 82.5 days	*Mean:* 97.25 days
	Q2: 82.5 days	*IQR:* 65 days	*SD:* 39.466 days
	Q3: 128.5 days	*Outliers:* none	*Sample Size:* 16
Girls:	*Q1:* 60 days	*Median:* 87 days	*Mean:* 107.636 days
	Q2: 87 days	*IQR:* 74 days	*SD:* 66.132 days
	Q3: 134 days	*Outliers:* 277	*Sample Size:* 11

The following table shows ordered observations and corresponding ranks:

	1	1	2	2	1	1	2	2	1	1	2	1	1	1
Obstn:	46	52	53	56	58	59	60	60	77	78	78	80	81	84
Rank:	1	2	3	4	5	6	7.5	7.5	9	10.5	10.5	12	13	14

	2	2	1	1	1	2	1	1	2	2	1	1	2
Obstn:	87	102	103	114	115	117	133	134	134	160	167	175	277
Rank:	15	16	17	18	19	20	21	22.5	22.5	24	25	26	27

The values of the statistics are $R_1 = 221$ and $R_2 = 157$. Letting $R = R_1$,

$$E(R) = 16(28)/2 = 224, \;\; SD(R) = \sqrt{11(16)(28)/12} = 20.265$$

and the observed value of the test statistic is

$$z_R = (221 - 224)/20.265 = -0.148.$$

Since the observed value of the test statistic is in the acceptance region for the test, the results are not statistically significant at the 5% level. Although the sample median age at death for boys (82.5 days) was less than that for girls (87 days), we believe that a difference of this size could have occurred by chance.

Example 9.12 (Iron and Dietary Supplements) (Rice, 1995, page 396) As part of an experiment to determine whether one of two forms of iron (either Fe2+ or Fe3+) would make a better dietary supplement, 18 mice were randomly assigned to receive the first form of iron (Fe2+), and 18 to receive the second form of iron (Fe3+), at a fixed concentration. The mice were given the iron orally. The iron was radioactively labelled so that measurements of the initial amount ingested and the amount remaining in the system after a fixed period of time could be made. The percentage of iron retained in the system was then calculated.

The following table gives the percentage retained by each mouse:

Fe2+ Group ($n = 18$)*:*

4.04	4.16	4.42	4.93	5.49	5.77	5.86	6.28	6.97
7.06	7.78	9.23	9.34	9.91	13.46	18.40	23.89	26.39

Fe3+ Group ($n = 18$)*:*

2.20	2.93	3.08	3.49	4.11	4.95	5.16	5.54	5.68
6.25	7.25	7.90	8.85	11.96	15.54	15.89	18.30	18.59

Numerical and graphical summaries of these data are given in Figure 9.6 (page 197).

To determine if the two forms of iron were retained differently, the researchers tested the null hypothesis that the difference in median percentage retention for the Fe2+ and Fe3+ groups is 0 versus the alternative that the difference is not 0 at the 5% significance level. Using Table 9.3 (page 194), the rejection region for this test is $|Z_R| \geq 1.962$.

The following table shows ordered observations and corresponding ranks:

	2	2	2	2	1	2	1	1	1	2	2	1
Obstn:	2.20	2.93	3.08	3.49	4.04	4.11	4.16	4.42	4.93	4.95	5.16	5.49
Rank:	1	2	3	4	5	6	7	8	9	10	11	12

	2	2	1	1	2	1	1	1	2	1	2	2
Obstn:	5.54	5.68	5.77	5.86	6.25	6.28	6.97	7.06	7.25	7.78	7.90	8.85
Rank:	13	14	15	16	17	18	19	20	21	22	23	24

	1	1	1	2	1	2	2	2	1	2	1	1
Obstn:	9.23	9.34	9.91	11.96	13.46	15.54	15.89	18.30	18.40	18.59	23.89	26.39
Rank:	25	26	27	28	29	30	31	32	33	34	35	36

The values of the statistics are $R_1 = 362$ and $R_2 = 304$. Letting $R = R_1$,

$$E(R) = 18(37)/2 = 333, \ SD(R) = \sqrt{18(18)(37)/12} = 31.607$$

and the observed value of the test statistic is

$$z_R = (362 - 333)/31.607 = 0.918.$$

Since the observed value of the test statistic is in the acceptance region for the test, the results are not statistically significant at the 5% level. Although the observed median percentage retained for animals given the first form of iron (7.018 percent) is greater than the observed median percentage retained for animals given the second form of iron (5.965 percent), a difference of this size could have occurred by chance.

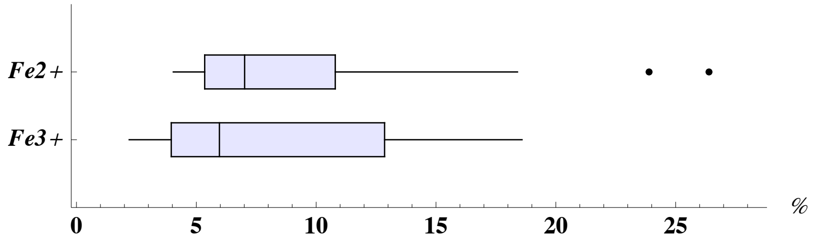
Footnote. Rank sum tests can be based on either R_1 or R_2. For example, using $R = R_2$ to compute the test statistic in the low birthweight and SIDS study, we get

$$E(R) = 11(28)/2 = 154, \ SD(R) = \sqrt{11(16)(28)/12} = 20.265$$

and

$$z_R = (157 - 154)/20.265 = 0.148.$$

Figure 9.6: *Summaries of percentage of iron retained using 2 forms of iron.*



Fe2+	*Q1:* 5.35 percent	*Median:* 7.015 percent	*Mean:* 9.632 percent
Group:	*Q2:* 7.015 percent	*IQR:* 5.4475 percent	*SD:* 6.691 percent
	Q3: 10.7975 percent	*Outliers:* 23.89, 26.39	*Sample Size:* 18
Fe3+	*Q1:* 3.955 percent	*Median:* 5.965 percent	*Mean:* 8.204 percent
Group:	*Q2:* 5.965 percent	*IQR:* 8.9 percent	*SD:* 5.447 percent
	Q3: 12.855 percent	*Outliers:* none	*Sample Size:* 18

The new value of z_R is the negative of the value computed earlier; our conclusion (that observed differences could have occurred by chance) remains the same.

Similarly, using $R = R_2$ to compute the test statistic in the iron and dietary supplements experiment, we get

$$E(R) = 18(37)/2 = 333, \;\; SD(R) = \sqrt{18(18)(37)/12} = 31.607$$

and

$$z_R = (304 - 333)/31.607 = -0.918.$$

Once again, the new value of z_R is the negative of the value computed in the example, and the conclusion does not change.

9.3 Models for Inference

In the *population model* for inference, one or more samples are chosen from a population using a chance method and statistical methods are used to generalize from the sample to the population. An alternative model for inference, called the *randomization model*, can sometimes be used to analyze convenience samples (samples that have not been chosen from larger populations using chance methods).

In the paired samples setting, where the data are ordered pairs of the form

(Response on Treatment 1, Response on Treatment 2),

the randomization model can be applied when we are interested in testing the null hypothesis of no treatment difference versus the alternative hypothesis that the treatments differ in some way. If the null hypothesis is true, then the labelling of responses as

"Response on Treatment 1" and "Response on Treatment 2"

can be considered to be completely arbitrary, and the list of n ordered pairs can be thought of as one of 2^n equally likely lists. For example,

(1) In the cell phones and driving example (beginning on page 171):

If using a cell phone while taking a driving skills test does not change reaction times, then for each student the labelling of one time as "With Cell Phone" and the other as "Without Cell Phone" could be considered to be completely arbitrary, and the data could be analyzed under the randomization model for inference.

(2) In the amiloride and cystic fibrosis example (beginning on page 186):

If the drug amiloride is ineffective in diminishing the loss of pulmonary function in patients with cystic fibrosis, then the labelling of one reduction in FVC as "Reduction on Placebo" and the other as "Reduction on Amiloride" could be considered to be completely arbitrary, and the data could be analyzed under the randomization model for inference.

(3) In the anorexia and family therapy example (beginning on page 188):

If the family therapy treatment for anorexia is ineffective, then for each girl the labelling of one weight as her "Initial Weight" and the other weight as her "Final Weight" could be considered to be completely arbitrary, and the data could be analyzed under the randomization model for inference.

When applying the randomization model in the paired samples setting,

- The sampling distribution of the test statistic is computed for all 2^n possible lists obtained by relabelling the responses in each pair, and
- An observed significance level is computed based on this distribution.

The results provide a valid comparison of treatments among the subjects in the study. The signed rank test is based on this approach and is valid under both the population model for inference and the randomization model for inference.

In the independent samples setting, where the data are two lists of the form

n_1 Responses for Treatment 1, and n_2 Responses for Treatment 2,

the randomization model can be used when we are interested in testing the null hypothesis of no treatment difference versus the alternative hypothesis that the treatments differ in some way. If the null hypothesis is true, then the labelling of n_1 observations as

"Responses for Treatment 1" and the remaining as "Responses for Treatment 2"

can be thought of as completely arbitrary, and the partition of the $n_1 + n_2$ observations into complementary lists of sizes n_1 and n_2, respectively, can be thought of as one of $\binom{n_1+n_2}{n_1}$ equally likely choices. For example,

(1) In the calcium and blood pressure example (beginning on page 174):

If the use of calcium supplements is ineffective in reducing blood pressure in adult men, then the labelling of n_1 observations as "Decreases for Men in Calcium Group" and n_2 observations as "Decreases for Men in Placebo Group" could be considered to be completely arbitrary, and the data could be analyzed under the randomization model for inference.

(2) In the low birthweight and SIDS example (beginning on page 192):

If there are no gender differences in ages at death, then the labelling of n_1 observed ages as those for boys and n_2 observed ages as those for girls could be considered as completely arbitrary, and the data could be analyzed under the randomization model.

(3) In the iron and dietary supplements example (beginning on page 195):

If there are no differences in the amount of iron retained by the two forms of iron, then the labeling of n_1 percentages as those for animals given the Fe2+ supplement and those for animals given the Fe3+ supplement could be considered to be completely arbitrary, and the data could be analyzed under the randomization model.

When applying the randomization model in the independent samples setting,

- The sampling distribution of the test statistic is computed for all $\binom{n_1+n_2}{n_1}$ possible partitions obtained by randomly choosing n_1 observations as the responses for the first treatment and the remaining n_2 observations as the responses for the second treatment, and
- An observed significance level is computed based on this distribution.

The results provide a valid comparison of treatments among the subjects in the study. The rank sum test is based on this approach and is valid under both the population model for inference and the randomization model for inference.

Student t methods, revisited. Any test statistic could be used in a randomization analysis, including the t-scores used in paired samples analyses and in independent samples analyses from the last chapter.

Interestingly, the Student t distribution can be used as an approximate sampling distribution under both models of inference.

Example 9.13 (Sampling Distributions of t-Scores) To illustrate this point, the computer was used to construct the sampling distributions of t-scores based on the randomization model for inference using the information from two examples from the last chapter:

(a) The Cell Phones & Driving Example (beginning on page 171):

Figure 9.7(a) (page 201) shows the sampling distribution of t-scores for all

$$2^{32} = 4294967296$$

relabellings of average reaction times for each student driver as "With Cell Phone" and "Without Cell Phone." The Student t curve (with $df = 31$) is superimposed on the distribution, and closely approximates the histogram.

In the original analysis, the observed t-score was 5.456 and the observed significance level was

$$p\text{-Value} = 2P(T \geq 5.456) = 0.00000581,$$

where T has a Student t distribution with 31 degrees of freedom.

Values of t-scores based on the randomization model ranged from -6.415 to $+6.415$ with the following distribution:

$t \leq -5.456$	$-5.456 < t < 5.456$	$t \geq 5.456$	Total
3359	4294960578	3359	4294967296

The p-value based on this sampling distribution is the proportion of t-scores whose absolute values are 5.456 or more,

$$p\text{-Value} = (3359{+}3359)/4294967296 = 0.00000156.$$

The two p-values are very close.

(b) The Calcium & Blood Pressure Example (beginning on page 174):

Figure 9.7(b) (page 201) shows the sampling distribution of t-scores for all

$$\binom{21}{10} = 352716$$

partitions of the 21 observed changes in blood pressure into subsets of sizes 10 (for the "calcium group") and 11 (for the "placebo group"). The Student t curve (with $df = 16$) is superimposed on the distribution, and closely approximates the histogram.

In the original analysis, the observed t-score was 1.717 and the observed significance level was

$$p\text{-Value} = 2P(T \geq 1.717) = 0.1053,$$

where T has a Student t distribution with 16 degrees of freedom.

Values of t-scores based on the randomization model ranged from -5.128 to $+5.408$ with the following distribution:

$t \leq -1.717$	$-1.717 < t < 1.717$	$t \geq 1.717$	Total
18669	316381	17666	352716

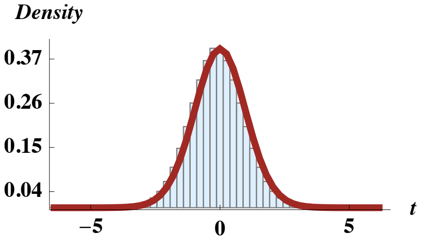
The p-value based on this sampling distribution is the proportion of t-scores whose absolute values are 3.961 or more,

$$p\text{-Value} = (18669{+}17666)/352716 = 0.103015.$$

Again, the two p-values are close.

Footnote. When using the t-score as test statistic under the population model for inference, the Student t distribution is an "idealized distribution" for simple random samples drawn from large populations. When using the t-score as test statistic under the randomization model for inference, the individuals in the study are fixed and the Student t-distribution is the "idealized distribution" for random relabellings of values.

Figure 9.7: *Sampling distributions of t-scores with Student t distributions superimposed.*



(a) *Cell Phones & Driving Example* (b) *Calcium & Blood Pressure Example*

9.4 Brief Summary and Additional Examples

This chapter introduces nonparametric methods for comparing medians in paired samples and independent samples settings, and methods for analyzing information under both the population and randomization models for inference.

Signed rank test. The distribution of the standardized signed rank statistic, Z_S, is constructed by finding the value of the statistic for each possible assignment of "±" signs to the list of absolute differences (equivalently, by finding the value of the statistic for each possible relabeling of responses on each treatment).

Under the population model for inference, we can generalize the results of the signed rank test to the larger population from which the individuals were drawn. Under the randomization model for inference, we cannot generalize the results beyond the individuals in the study.

The critical values for conducting two-sided signed rank tests given in Table 9.1 (page 185) are valid when there are no (or few) ties among the absolute differences. When there are many ties, the formula for the standard deviation, $SD(S)$, changes to accommodate the reduced variability of the signed rank statistics, and different critical values are needed when conducting signed rank tests.

Rank sum test. The distribution of the standardized rank sum statistic, Z_R, is constructed by finding the value of the statistic for each possible choice of n_1 observations as responses for the first treatment (with the remaining as responses for the second treatment).

Under the population model for inference, we can generalize the results of the rank sum test to the populations from which the data were drawn. Under the randomization model for inference, we cannot generalize the results beyond the individuals in the study.

The critical values for conducting rank sum tests given in Tables 9.2–9.3 (pages 193–194) are valid when there are no (or few) ties among the observations. When there are many ties, the formulas for the standard deviations of R_1 and R_2 change to accommodate the reduced variability of rank sums, and different critical values are needed when conducting two-sided rank sum tests.

Statistical methods for two or more samples. The following table summarizes the two sample methods discussed in the last two chapters, and includes methods for 3 or more samples that are commonly used in analyzing comparative studies:

	Is Population Distribution Approximately Normal?	
	Yes	*No*
Paired Samples:	Paired t Methods	Signed Rank Test (symmetric distributions)
2 Independent Samples:	Welch t Methods	Rank Sum Test
3+ Independent Samples:	Analysis of Variance Methods (equal variances assumed)	Kruskal-Wallis Test

As mentioned in the last chapter: (1) William Sealy Gosset (writing under the pseudonym "Student") developed his Student t methods to handle small sample analyses of means (and differences in means) arising from industrial experiments; (2) small sample methods based on the Student t distribution applied to differences data are often called *paired t methods*; and (3) small sample methods for independently chosen samples based on the Student t distribution are often called *Welch t methods* in honor of the statistician who developed the degrees of freedom formula used in the analyses.

Signed rank and rank sum methods were developed by Frank Wilcoxon (1892-1965) in the 1940's to handle comparisons in situations where populations distributions are not approximately normally distributed. Several different notations, and several equivalent statistics, are used in practice, so you need to check the form used in a particular program. For example, some statistical programs call the difference in statistics, $S_+ - S_-$, the signed rank statistic. In addition, many programs do not report the standardized forms of the statistics (Z_S or Z_R).

R.A. Fisher, whose name has been mentioned several times in this text, is credited with developing *analysis of variance* methods for analyzing 3 or more independent samples under the strict assumptions listed in the table. William Kruskal (1919-2005) and W.Allen Wallis (1912-1998) are credited with generalizing the rank sum test to 3 or more samples.

Finally, it should be noted that inference methods based on the *randomization model for inference* began with work published by R.A. Fisher and by E.J.G. Pitman (1897-1933) in the 1930's. Methods based on their work were of theoretical interest only before the modern era of high speed computers, but are now important practical tools in statistical analysis. In the words of John Tukey (whose name has also been mentioned several times in this text), methods based on the randomization model for inference have become

"The wave of the past in the future."

(See page 58 for information about many of Tukey's accomplishments.)

10 Introduction to Association Analysis

This chapter introduces a variety of methods for analyzing the association between variables. References for this chapter include Agresti & Franklin (2007, Chapters 10-11), Freedman et al (1991, Part VIII), Moore & McCabe (1999, Chapters 9-10), Moore & Notz (2006, Part IV) and Pagano & Gauvreau (2000, Chapters 15-18).

Analyses of association in categorical data settings rely on methods based on the chi-square distribution, with critical values from Table 10.1 (page 204). Analyses of association in least squares regression settings rely on methods based on the Student t distribution, with critical values from Tables 8.1-8.2 (pages 165-166).

Table 10.1: *Critical values for chi-square tests when df $\leq$ 40.*

	Significance Level as a Percent: 100α%				
	20%	10%	5%	1%	0.1%
df	χ^2_α	χ^2_α	χ^2_α	χ^2_α	χ^2_α
1	1.642	2.706	3.841	6.635	10.828
2	3.219	4.605	5.991	9.210	13.816
3	4.642	6.251	7.815	11.345	16.266
4	5.989	7.779	9.488	13.277	18.467
5	7.289	9.236	11.070	15.086	20.515
6	8.558	10.645	12.592	16.812	22.458
7	9.803	12.017	14.067	18.475	24.322
8	11.030	13.362	15.507	20.090	26.124
9	12.242	14.684	16.919	21.666	27.877
10	13.442	15.987	18.307	23.209	29.588
11	14.631	17.275	19.675	24.725	31.264
12	15.812	18.549	21.026	26.217	32.909
13	16.985	19.812	22.362	27.688	34.528
14	18.151	21.064	23.685	29.141	36.123
15	19.311	22.307	24.996	30.578	37.697
16	20.465	23.542	26.296	31.100	39.252
17	21.615	24.769	27.587	33.409	40.790
18	22.760	25.989	28.869	34.805	42.312
19	23.900	27.204	30.144	36.191	43.820
20	25.038	28.412	31.410	37.566	45.315
21	26.171	29.615	32.671	38.932	46.797
22	27.301	30.813	33.924	40.289	48.268
23	28.429	32.007	35.172	41.638	49.728
24	29.553	33.196	36.415	42.980	51.179
25	30.675	34.382	37.652	44.314	52.620
26	31.795	35.563	38.885	45.642	54.052
27	32.912	36.741	40.113	46.963	55.476
28	34.027	37.916	41.337	48.278	56.892
29	35.139	39.087	42.557	49.588	58.301
30	36.250	40.256	43.773	50.892	59.703
31	37.359	41.422	44.985	52.191	61.098
32	38.466	42.585	46.194	53.486	62.487
33	39.572	43.745	47.400	54.776	63.870
34	40.676	44.903	48.602	56.061	65.247
35	41.778	46.059	49.802	57.342	66.619
36	42.879	47.212	50.998	58.619	67.985
37	43.978	48.363	52.192	59.893	69.346
38	45.076	49.513	53.384	61.162	70.703
39	46.173	50.660	54.572	62.428	72.055
40	47.269	51.805	55.758	63.691	73.402

11 References

This chapter includes textbook references for the entire project (the main text plus the four workbooks), plus references for the initials used to identify sources of material for certain problems and examples. References for websites and research articles are given in the examples where they are used and are not repeated here.

Additional problems were kindly provided by Tom Crawford **[TC]**, Rob Gross **[RG]**, Charlie Landraitis **[CKL]**, Spencer Leslie **[SL]**, Ned Rosen **[NR]**, and Wei Tao **[WT]**.

Text references:

Agresti, A. (1990), *Categorical Data Analysis*, New York: John Wiley & Sons.

Agresti, A. and C. Franklin (2007), *Statistics: The Art and Science of Learning from Data*, New Jersey: Pearson Prentice Hall. **[AF]**

Baldi, B. and D.S. Moore (2009), *The Practice of Statistics in the Life Sciences*, New York: W.H. Freeman & Company. **[BM]**

Cook, R.D. and S. Weisberg (1994), *Introduction to Regression Graphics*, New York: John Wiley & Sons, Inc. **[CW]**

D'Agostino, R., Sullivan, L. and A. Beiser (2004), *Introductory Applied Biostatistics*, KY: Cengage Learning. **[DSB]**

Daniel, W.W. and C.L. Cross (2013), *Biostatistics: A Foundation for Analysis in the Health Sciences*, tenth edition, NJ: John Wiley & Sons Inc. **[DC]**

DeGroot, M.H. and M.J. Schervish, (2002), *Probability and Statistics*, third edition, New York: Addison Wesley. **[DS]**

Freedman, D.A. (2005), *Statistical Models: Theory and Practice*, New York: Cambridge University Press.

Freedman, D., R. Pisani, R. Purves and A. Adhikari (1991), *Statistics*, second edition, New York: W.W. Norton & Company, Inc. **[FPP]**

Hand, D. J., F. Daly, A. D. Lunn, K. J. McConway and E. Ostrowski (1994). A handbook of small data sets. New York: Chapman & Hall.

Ingelfinger, J.A., F. Mosteller, L.A. Thibodeau and J.H. Ware (1983), *Biostatistics in Clinical Medicine*, New York: Macmillan Publishing Company.

Kokoska, S. (2009), *Introduction to Statistics: A Problem Solving Approach*, NY: W.H. Freeman & Company. **[SK]**

Lange, N., L. Ryan, L, Billard, D. Brillinger, L. Conquest and J. Greenhouse (1994), *Case Studies in Biometry*, New York: John Wiley & Sons.

Larsen, R.J. and M.L. Marx (1986), *An Introduction to Mathematical Statistics and its Applications*, second edition, New Jersey: Prentice-Hall. **[LM]**

Mendenhall, W., Scheaffer, R.L., and D.D. Wackerly (1986), *Mathematical Statistics with Applications*, third edition, Boston: Duxbury Press. **[MSW]**

Moore, D.S. and G.P. McCabe (1999), *Introduction to the Practice of Statistics*, third edition, New York: W.H. Freeman and Company. **[MM]**

Moore, D.S. and W.I. Notz (2006), *Statistics: Concepts and Controversies*, sixth edition, New York: W.H. Freeman and Company. **[MN]**

Olkin, I., Gleser L.J. and C. Derman (1994), *Probability Models and Applications*, second edition, New York: Macmillan College Publishing Company. **[OGD]**

Pagano, M. and K. Gauvreau (2000), *Principles of Biostatistics*, second edition, Belmont, CA: Duxbury Press. **[PG]**

Peck, R., L.D. Haugh, A. Goodman (1998), *Statistical Case Studies: A Collaboration Between Academe and Industry*, Philadelphia, PA: the American Statistical Association and the Society for Industrial and Applied Mathematics.

Peck, R., G. Casella, G. Cobb, R. Hoerl, D. Nolan, R. Starbuck and H. Stern (2006), *Statistics: A Guide to the Unknown*, fourth edition, Belmont, CA: Duxbury Press.

Pitman, J. (1993), *Probability*, New York: Springer-Verlag, Inc.

Rice, J.A. (1995), *Mathematical Statistics and Data Analysis*, second edition, Belmont, CA: Duxbury Press.

Rosner, B. (2011), *Fundamentals of Biostatistics*, seventh edition, MA: Brooks/Cole. **[BR]**

Weinstein, M.C. and H.V. Fienberg (1980), *Clinical Decision Analysis*, Philadelphia, PA: W.B. Saunders Company. **[WF]**

Weiss, N.A. (2006), *A Course in Probability*, New York: Pearson Education, Inc. **[NW]**