Statistical Methods for the Social Sciences

Fifth Edition

Alan Agresti

University of Florida

Pearson

Chapter

Sampling and Measurement 2

CHAPTER OUTLINE

To analyze social phenomena with a statistical analysis, *descriptive* methods summarize the data and *inferential* methods use sample data to make predictions about populations. In gathering data, we must decide which subjects to sample. (Recall that the *subjects* of a population to be sampled could be individuals, families, schools, cities, hospitals, records of reported crimes, and so on.) Selecting a sample that is representative of the population is a primary topic of this chapter.

For our sample, we must convert our ideas about social phenomena into data by deciding what to measure and how to measure it. Developing ways to measure abstract concepts such as performance, achievement, intelligence, and prejudice is one of the most challenging aspects of social research. A measure should have *validity*, describing what it is intended to measure and accurately reflecting the concept. It should also have *reliability*, being consistent in the sense that a subject will give the same response when asked again. Invalid or unreliable data-gathering instruments render statistical analyses of the data meaningless and even possibly misleading.

The first section of this chapter introduces definitions pertaining to measurement, such as types of data. The other sections discuss ways, good and bad, of selecting the sample.

2.1 Variables and Their Measurement

Statistical methods help us determine the factors that explain *variability* among subjects. For instance, variation occurs from student to student in their college grade point average (GPA). What is responsible for that variability? The way those students vary in how much they study per week? How much they watch TV per day? Their IQ? Their college board scores? Their high school GPA?

VARIABLES

Any characteristic that we can measure for each subject is called a ***variable***. The name reflects that values of the characteristic *vary* among subjects.

Variable

> A ***variable*** is a characteristic that can vary in value among subjects in a sample or population.

Different subjects may have different values of a variable. Examples of variables are income last year, number of siblings, whether employed, and gender. The values the variable can take form the ***measurement scale***. For gender, for instance, the measurement scale consists of the two labels, (female, male). For number of siblings, it is (0, 1, 2, 3, 4, and so on).

The valid statistical methods for a variable depend on its measurement scale. We treat a numerical-valued variable such as annual income differently from a variable with a measurement scale consisting of categories, such as (yes, no) for whether employed. We next present ways to classify variables. The first type refers to whether the measurement scale consists of categories or numbers. Another type refers to the number of levels in that scale.

QUANTITATIVE VARIABLES AND CATEGORICAL VARIABLES

A variable is called ***quantitative*** when the measurement scale has numerical values that represent different magnitudes of the variable. Examples of quantitative variables are a subject's annual income, number of siblings, age, and number of years of education completed.

A variable is called ***categorical*** when the measurement scale is a set of categories. For example, marital status, with categories (single, married, divorced, widowed), is categorical. For Canadians, the province of residence is categorical, with the categories (Alberta, British Columbia, and so on). Other categorical variables are whether employed (yes, no), primary clothes shopping destination (local mall, local downtown, Internet, other), favorite type of music (classical, country, folk, jazz, rock), religious affiliation (Christianity, Islam, Hinduism, Buddhism, Jewish, other, none), and political party preference.

For categorical variables, distinct categories differ in quality, not in numerical magnitude. Categorical variables are often called ***qualitative***. We distinguish between categorical and quantitative variables because different statistical methods apply to each type. For example, the *average* is a statistical summary for quantitative variables, because it uses numerical values. It's possible to find the average for a quantitative variable such as income, but not for a categorical variable such as favorite type of music.

NOMINAL, ORDINAL, AND INTERVAL SCALES OF MEASUREMENT

For a quantitative variable, the possible numerical values are said to form an ***interval*** scale, because they have a numerical distance or *interval* between each pair of levels. For annual income, for instance, the interval between $40,000 and $30,000 equals $10,000. We can compare outcomes in terms of how much larger or how much smaller one is than the other.

Categorical variables have two types of scales. For the categorical variables mentioned in the previous subsection, such as religious affiliation, the categories are *unordered*. The scale does not have a "high" or "low" end. The categories are then said to form a ***nominal scale***. For another example, a variable measuring primary mode of transportation to work might use the nominal scale (automobile, bus, subway, bicycle, walking). For a nominal variable, no category is greater than or smaller than any other category. Labels such as "automobile" and "bus" for mode of transportation identify the categories but do not represent different magnitudes. By contrast, each possible value of a quantitative variable is *greater than* or *less than* any other possible value.

A third type of scale falls, in a sense, between nominal and interval. It consists of categorical scales having a natural *ordering* of values. The categories form an ***ordinal scale***. Examples are social class (upper, middle, lower), political philosophy (very liberal, slightly liberal, moderate, slightly conservative, very conservative), government spending on the environment (too little, about right, too much), and frequency of religious activity (never, less than once a month, about 1–3 times a month, every week, more than once a week). These scales are not nominal, because the categories are ordered. They are not interval, because there is no defined distance between levels. For example, a person categorized as very conservative is *more* conservative than a person categorized as slightly conservative, but there is no numerical value for *how much more* conservative that person is.

The scales refer to the actual measurement and not to the phenomena themselves. *Place of residence* may indicate a geographic place name such as a county (nominal), the distance of that place from a point on the globe (interval), the size of the place (interval or ordinal), or other kinds of variables.

QUANTITATIVE ASPECTS OF ORDINAL DATA

Levels of nominal scales are qualitative, varying in quality, not in quantity. Levels of interval scales are quantitative, varying in magnitude. The position of ordinal scales on the quantitative–qualitative classification is fuzzy. Because their scale is a set of categories, they are often analyzed using the same methods as nominal scales. But in many respects, ordinal scales more closely resemble interval scales. They possess an important quantitative feature: Each level has a *greater* or *smaller* magnitude than another level.

Some statistical methods apply specifically to ordinal variables. Often, though, it's helpful to analyze ordinal scales by assigning numerical scores to categories. By treating ordinal variables as interval scale rather than nominal scale, we can use the more powerful methods available for quantitative variables. For example, course grades (such as A, B, C, D, E) are ordinal. But, we treat them as interval when we assign numbers to the grades (such as 4, 3, 2, 1, 0) to compute a grade point average.

DISCRETE AND CONTINUOUS VARIABLES

One other way to classify a variable also helps determine which statistical methods are appropriate for it. This classification refers to the *number* of values in the measurement scale.

Discrete and Continuous Variables

A variable is ***discrete*** if its possible values form a set of separate numbers, such as (0, 1, 2, 3, ...). It is ***continuous*** if it can take an infinite continuum of possible real number values.

An example of a discrete variable is the number of siblings. Any variable phrased as "the number of ..." is discrete, because it is possible to list its possible values {0, 1, 2, 3, 4, ...}.

Examples of continuous variables are height, weight, and the amount of time it takes to read a passage of a book. It is impossible to write down all the distinct potential values, since they form an interval of infinitely many values. The amount of time needed to read a book, for example, could take the value 8.62944 ... hours.

Discrete variables have a basic unit of measurement that cannot be subdivided. For example, 2 and 3 are possible values for the number of siblings, but 2.571 is not. For a continuous variable, by contrast, between any two possible values there is always another possible value. For example, age is continuous in the sense that an individual does not age in discrete jumps. At some well-defined point during the year in which you age from 21 to 22, you are 21.385 years old, and similarly for every other real number between 21 and 22. A continuous, infinite collection of age values occurs between 21 and 22 alone.

Any variable with a finite number of possible values is discrete. Categorical variables, nominal or ordinal, are discrete, having a finite set of categories. Quantitative variables can be discrete or continuous; age is continuous, and number of siblings is discrete.

For quantitative variables, the distinction between discrete and continuous variables can be blurry, because of how variables are actually measured. In practice, we round continuous variables when measuring them, so the measurement is actually discrete. We say that an individual is 21 years old whenever that person's age is somewhere between 21 and 22. On the other hand, some variables, although discrete, have a very large number of possible values. In measuring annual family income in dollars, the potential values are (0, 1, 2, 3, . . .), up to some very large value in many millions.

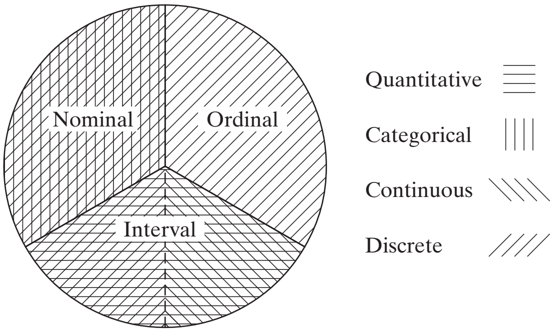
What's the implication of this? Statistical methods for discrete variables are mainly used for quantitative variables that take relatively few values, such as the number of times a person has been married. Statistical methods for continuous variables are used for quantitative variables that can take lots of values, regardless of whether they are theoretically continuous or discrete. For example, statisticians treat variables such as age, income, and IQ as continuous.

In summary,

- Variables are either *quantitative* (numerical-valued) or *categorical*. Quantitative variables are measured on an *interval* scale. Categorical variables with unordered categories have a *nominal* scale, and categorical variables with ordered categories have an *ordinal* scale.
- Categorical variables (nominal or ordinal) are *discrete*. Quantitative variables can be either discrete or continuous. In practice, quantitative variables that can take lots of values are treated as *continuous*.

Figure 2.1 summarizes the types of variables, in terms of the (quantitative, categorical), (nominal, ordinal, interval), and (continuous, discrete) classifications.

FIGURE 2.1: Summary of Quantitative–Categorical, Nominal–Ordinal–Interval, and Continuous–Discrete Classifications



Note: Ordinal data are treated sometimes as categorical and sometimes as quantitative

2.2 Randomization

Inferential statistical methods use sample statistics to make predictions about values of population parameters. The quality of the inferences depends on how well the sample represents the population. This section introduces ***randomization***, the mechanism for achieving good sample representation.

In this section and throughout the text, we let n denote the number of subjects in the sample. This is called the ***sample size***.

SIMPLE RANDOM SAMPLING

Simple random sampling is a method of sampling for which every possible sample of size n has equal chance of selection.

Simple Random Sample

A ***simple random sample*** of n subjects from a population is one in which each possible sample of that size has the same probability (chance) of being selected.

For instance, suppose you want to select a simple random sample of a student from a class of 60 students. For a simple random sample of $n = 1$ student, each of the 60 students is equally likely to be selected. You could select one by numbering the students from 01 to 60, placing the 60 numbers on 60 identical ballots, and selecting one blindly from a hat. For a simple random sample of $n = 2$ students from the class, each possible sample of size 2 is equally likely. The potential samples are (01, 02), (01, 03), (01, 04), ..., (59, 60). To select the sample, you blindly select two ballots from the hat. But this is unwieldy if the population size is large, and these days we can easily select the sample using a *random number generator* with software.

A simple random sample is often just called a ***random sample***. The *simple* adjective is used to distinguish this type of sampling from more complex sampling schemes presented in Section 2.4 that also have elements of randomization.

Why is it a good idea to use random sampling? Because everyone has the same chance of inclusion in the sample, so it provides fairness. This reduces the chance that the sample is seriously biased in some way, leading to inaccurate inferences about the population. Most inferential statistical methods assume randomization of the sort provided by random sampling.

HOW TO SELECT A SIMPLE RANDOM SAMPLE?

To select a random sample, we need a list of all subjects in the population. This list is called the ***sampling frame***. Suppose you plan to sample students at your school. The population is all students at the school. One possible sampling frame is the student directory.

The most common method for selecting a random sample is to (1) number the subjects in the sampling frame, (2) generate a set of these numbers randomly, and (3) sample the subjects whose numbers were generated. Using *random numbers* to select the sample ensures that each subject has an equal chance of selection.

Random Numbers

Random numbers are numbers that are computer generated according to a scheme whereby each digit is equally likely to be any of the integers 0, 1, 2, ..., 9 and does not depend on the other digits generated.

Table 2.1 shows a table containing random numbers, in sets of size 5. The numbers fluctuate according to no set pattern. Any particular number has the same chance of being a 0, 1, 2, ..., or 9. The numbers are chosen independently, so any one digit chosen has no influence on any other selection. If the first digit in a row of the table is a 9, for instance, the next digit is still just as likely to be a 9 as a 0 or 1 or any other number.

TABLE 2.1: A Table of Random Numbers

Line/Col.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	90826	68432	36255	32536	92103	76082	82293	78852
2	77714	33924	86688	94720	45943	83064	68007	10523
3	34371	53100	81078	34696	92393	92799	72281	62696

Source: Constructed using `sample` function in R.

Although random numbers are available in published tables, we can easily generate them with software and many statistical calculators. For example, suppose you want to randomly select $n = 4$ students out of a class of size 60. After assigning the numbers $(01, 02, \ldots, 60)$ to the class members, you can use software to generate four random numbers between 01 and 60. R is a software package that can do this. It is available to download for free at `www.r-project.org`. In R, the `sample` function performs simple random sampling from a numbered population list. Here is how to select a sample of size 4 from a population of size 60 (the > is the R system prompt, and you type in *sample(1:60, 4)* and press the *enter* key on your keyboard):

```
> sample(1:60, 4) # put comments on command line after the # symbol
[1] 22 47 38 44   # these are the four numbers randomly generated
```

The sample of size 4 selects the students numbered 22, 47, 38, and 44.

COLLECTING DATA WITH SAMPLE SURVEYS

Many studies select a sample of people from a population and interview them. This method of data collection is called a ***sample survey***. The interview could be a personal interview, telephone interview, or self-administered questionnaire.

The General Social Survey (GSS) is an example of a sample survey. It gathers information using personal interviews of a random sample of subjects from the U.S. adult population to provide a snapshot of that population. (They do not use *simple* random sampling but rather a method discussed later in the chapter that incorporates multiple stages and clustering but is designed to give each family the same chance of inclusion.) National polls such as the Gallup Poll are also sample surveys. They usually use telephone interviews. Since it is often difficult to obtain a sampling frame, especially since many people now have cell phones but not landline phones, many telephone interviews obtain the sample with *random digit dialing*.

COLLECTING DATA WITH AN EXPERIMENT

In some studies, data result from a planned ***experiment***. The purpose of most experiments is to compare responses of subjects on some outcome measure, under different conditions. Those conditions are levels of a variable that can influence the outcome. The scientist has the experimental control of being able to assign subjects to the conditions.

The conditions in an experiment are called ***treatments***. For instance, the treatments might be different drugs for treating some illness. To conduct the experiment, the researcher needs a plan for assigning subjects to the treatments. These plans are called *experimental designs*. Good experimental designs use randomization to determine which treatment a subject receives. This reduces bias and allows us to use statistical inference to make predictions.

In the late 1980s, the Physicians' Health Study Research Group at Harvard Medical School designed an experiment to analyze whether regular intake of aspirin reduces mortality from heart disease. Of about 22,000 male physicians, half were randomly chosen to take an aspirin every other day. The remaining half took a placebo, which had no active agent. After five years, rates of heart attack were compared. By using randomization to determine who received which treatment, the researchers knew the groups would roughly balance on factors that could affect heart attack rates, such as age and quality of health. If the physicians could decide on their own which treatment to take, the groups might have been out of balance on some important factor. Suppose, for instance, younger physicians were more likely to select aspirin. Then, a lower heart attack rate among the aspirin group could occur merely because younger subjects are less likely to suffer heart attacks.

In medicine, experiments using randomization (so-called *randomized clinical trials*) have been the gold standard for many years. But randomized experiments are also increasingly used in the social sciences. For example, researchers use randomized experiments to evaluate programs for addressing poverty in the developing world. For many examples, see the websites

`www.povertyactionlab.org/methodology` and `www.nature.com/news`,

at the latter site searching for the article "Can randomized trials eliminate global poverty?" (by J. Tollefson, August 12, 2015).

COLLECTING DATA WITH AN OBSERVATIONAL STUDY

In social research, it is often not feasible to conduct experiments. It's usually not possible to randomly assign subjects to the groups we want to compare, such as levels of gender or race or educational level or annual income. Many studies, such as sample surveys, merely *observe* the outcomes for available subjects on the variables without any experimental manipulation of the subjects. Such studies are called ***observational studies***. The researcher measures subjects' responses on the variables of interest but has no experimental control over the subjects.

With observational studies, comparing groups is difficult because the groups may be imbalanced on variables that affect the outcome. This is true even with random sampling. For instance, suppose we plan to compare black students, Hispanic students, and white students on some standardized exam. If white students have a higher average score, a variety of variables might account for that difference, such as parents' education or parents' income or quality of school attended. This makes it difficult to compare groups with observational studies, especially when some key variables may not have been measured in the study.

Establishing cause and effect is central to science. But it is not possible to establish cause and effect definitively with a nonexperimental study, whether it be an observational study with an available sample or a sample survey using random sampling. An observational study always has the possibility that some unmeasured variable could be responsible for patterns observed in the data. By contrast, with an experiment that randomly assigns subjects to treatments, those treatments should roughly balance on any unmeasured variables. For example, in the aspirin and heart attack study mentioned above, the doctors taking aspirin would not tend to be younger or of better health than the doctors taking the placebo. Because a randomized experiment balances the groups being compared on other factors, we can use it to study cause and effect.

2.3 Sampling Variability and Potential Bias

Even if a study wisely uses randomization, the results of the study still depend on which subjects are sampled. Two researchers who separately select random samples from some population may have little overlap, if any, between the two sample memberships. Therefore, the values of sample statistics will differ for the two samples, and the results of analyses based on these samples may differ.

SAMPLING ERROR

Suppose the Gallup, Harris, Ipsos-Reid, and Pew polling organizations each randomly sample 1000 adult Canadians, in order to estimate the percentage of Canadians who give the prime minister's performance in office a favorable rating. Based on the samples they select, perhaps Gallup reports an approval rating of 53%, Harris reports 58%, Ipsos-Reid 55%, and Pew 54%. These differences could reflect slightly different question wording. But even if the questions are worded exactly the same, the percentages would probably differ somewhat because the samples are different.

For conclusions based on statistical inference to be worthwhile, we should know the potential ***sampling error***—how much the statistic differs from the parameter it predicts because of the way results naturally exhibit variation from sample to sample.

Sampling Error

> The ***sampling error*** of a statistic is the error that occurs when we use a statistic based on a sample to predict the value of a population parameter.

Suppose that the percentage of the population of adult Canadians who give the prime minister a favorable rating is 56%. Then the Gallup organization, which predicted 53%, had a sampling error of 53% − 56% = −3%. The Harris organization, which predicted 58%, had a sampling error of 58% − 56% = 2%. In practice, the sampling error is unknown, because the values of population parameters are unknown.

Random sampling protects against bias, in the sense that the sampling error tends to fluctuate about 0, sometimes being positive (as in the Harris Poll) and sometimes being negative (as in the Gallup Poll). Random sampling also allows us to predict the likely size of the sampling error. For sample sizes of about 1000, we'll see that the sampling error for estimating percentages is usually no greater than plus or minus 3%. This bound is the *margin of error*. Variability also occurs in the values of sample statistics with nonrandom sampling, but the extent of the sampling error is not predictable as it is with random sampling.

SAMPLING BIAS: NONPROBABILITY SAMPLING

Other factors besides sampling error can cause results to vary from sample to sample. These factors can also possibly cause bias. We next discuss three types of bias. The first is called ***sampling bias***.

For simple random sampling, each possible sample of n subjects has the same probability of selection. This is a type of ***probability sampling*** method, meaning that the probability any particular sample will be selected is known. Inferential statistical methods assume probability sampling. ***Nonprobability sampling*** methods are ones for which it is not possible to determine the probabilities of the possible samples. Inferences using such samples have unknown reliability and result in ***sampling bias***.

The most common nonprobability sampling method is ***volunteer sampling***. As the name implies, subjects volunteer for the sample. But the sample may poorly represent the population and yield misleading conclusions. Examples of volunteer sampling are visible daily on Internet sites and television news programs. Viewers register their opinions on an issue by voting over the Internet. The viewers who respond are unlikely to be a representative cross section, but will be those who can easily access the Internet and who feel strongly enough to respond. Individuals having a particular opinion might be much more likely to respond than individuals having a different opinion. For example, one night the ABC TV program *Nightline* asked viewers whether the United Nations should continue to be located in the United

States. Of more than 186,000 respondents, 67% wanted the United Nations out of the United States. At the same time, a poll using a random sample of about 500 respondents estimated the population percentage to be about 28%. Even though the random sample had a much smaller size, it is far more trustworthy.

A large sample does not help with volunteer sampling—the bias remains. In 1936, the newsweekly *Literary Digest* sent over 10 million questionnaires in the mail to predict the outcome of the presidential election. The questionnaires went to a relatively wealthy segment of society (those having autos or telephones), and fewer than 25% were returned. The journal used these to predict an overwhelming victory by Alfred Landon over Franklin Roosevelt. The opposite result was predicted by George Gallup with a much smaller sample in the first scientific poll taken for this purpose. In fact, Roosevelt won in a landslide.

The sampling bias inherent in volunteer sampling is also called ***selection bias***. It is problematic to evaluate policies and programs when individuals can choose whether or not to participate in them. For example, if we were evaluating a program such as Head Start in which participation is partly based on a parental decision, we would need to consider how family background variables (such as mother's educational level) could play a role in that decision and in the outcome evaluated.

Unfortunately, volunteer sampling is sometimes unavoidable, especially in medical studies. Suppose a study plans to investigate how well a new drug performs compared to a standard drug, for subjects who suffer from high blood pressure. The researchers are not going to be able to find a sampling frame of all who suffer from high blood pressure and take a simple random sample of them. They may, however, be able to sample such subjects at certain medical centers or using volunteers. Even then, randomization should be used wherever possible. For the study patients, the researchers can randomly select who receives the new drug and who receives the standard one.

Even with random sampling, sampling bias can occur. One case is when the sampling frame suffers from ***undercoverage***: It lacks representation from some groups in the population. A telephone survey will not reach prison inmates or homeless people, whereas families that have many phones will tend to be over-represented. Responses by those not having a telephone might tend to be quite different from those actually sampled, leading to biased results. About 21% of adults are under age 30, yet only 5% of the population having a landline phone are under age 30, so substantial bias could occur if we sampled only landlines.[1] Likewise there would be bias if we sampled only cell phones, because adults who have only a cell phone tend to be younger, poorer, more likely to be renters, to live with unrelated adults, and to be Hispanic than those who also have a landline phone.

RESPONSE BIAS

In a survey, the way a question is worded or asked can have a large impact on the results. For example, when a New York Times/CBS News poll asked whether the interviewee would be in favor of a new gasoline tax, only 12% said yes. When the tax was presented as reducing U.S. dependence on foreign oil, 55% said yes, and when asked about a gas tax that would help reduce global warming, 59% said yes.[2]

Poorly worded or confusing questions result in ***response bias***. Even the order in which questions are asked can influence the results dramatically. During the Cold War, a study asked, "Do you think the U.S. should let Russian newspaper reporters come here and send back whatever they want?" and "Do you think Russia should let American newspaper reporters come in and send back whatever they want?" The

[1] See `http://magazine.amstat.org/blog/2014/10/01/prescolumnoct14.`

[2] Column by T. Friedman, *New York Times*, March 2, 2006.

percentage of yes responses to the first question was 36% when it was asked first and 73% when it was asked second.[3]

In an interview, characteristics of the interviewer may result in response bias. Respondents might lie if they think their belief is socially unacceptable. They may be more likely to give the answer that they think the interviewer prefers. In a study on the effect of the interviewer's race, following a phone interview, respondents were asked whether they thought the interviewer was black or white (all were actually black). Perceiving a white interviewer resulted in more conservative opinions. For example, 14% agreed that "American society is fair to everyone" when they thought the interviewer was black, but 31% agreed to this when they thought the interviewer was white.[4]

NONRESPONSE BIAS: MISSING DATA

Some subjects who are selected for the sample may refuse to participate, or it may not be possible to reach them. This results in ***nonresponse bias***. If only half the intended sample was actually observed, we should worry about whether the half not observed differ from those observed in a way that causes biased results. Even if we select the sample randomly, the results are questionable if there is substantial nonresponse, say, over 20%.

For her book *Women in Love*, author Shere Hite surveyed women in the United States. One of her conclusions was that 70% of women who had been married at least five years have extramarital affairs. She based this conclusion on responses to questionnaires returned by 4500 women. This sounds like an impressively large sample. However, the questionnaire was mailed to about 100,000 women. We cannot know whether the 4.5% of the women who responded were representative of the 100,000 who received the questionnaire, much less the entire population of American women. This makes it dangerous to make an inference to the larger population.

A problem in many studies is ***missing data***: Some subjects do not provide responses for some of the variables measured. This problem is especially common in studies that observe people over time (called *longitudinal studies*), as some people may drop out of the study for various reasons. Even in censuses, which are designed to observe everyone in a country, some people are not observed or fail to cooperate. A statistical analysis that ignores cases for which some observations are missing wastes information and has possible bias.

SUMMARY OF TYPES OF BIAS

In summary, sample surveys have potential sources of bias:

- ***Sampling bias*** occurs from using nonprobability samples, such as the *selection bias* inherent in volunteer samples.
- ***Response bias*** occurs when the subject gives an incorrect response (perhaps lying), or the question wording or the way the interviewer asks the questions is confusing or misleading.
- ***Nonresponse bias*** occurs when some sampled subjects cannot be reached or refuse to participate or fail to answer some questions.

These sources of bias can also occur in observational studies other than sample surveys and even in experiments. In any study, carefully assess the scope of conclusions. Evaluate critically the conclusions by noting the makeup of the sample. How

[3] See Crossen (1994).

[4] *Washington Post*, June 26, 1995.

was the sample selected? How large was it? How were the questions worded or the variables measured? Who sponsored and conducted the research? The less information that is available, the less you should trust it.

Finally, be wary of any study that makes inferences to a broader population than is justified by the sample chosen. Suppose a psychologist performs an experiment using a random sample of students from an introductory psychology course. With statistical inference, the sample results generalize to the population of all students in the class. For the results to be of wider interest, the psychologist might claim that the conclusions extend to *all* college students, to all young adults, or even to all adults. These generalizations may well be wrong, because the sample may differ from those populations in fundamental ways, such as in average age or socioeconomic status.

2.4 Other Probability Sampling Methods*

Section 2.2 introduced ***simple random sampling*** and explained its importance to statistical inference. In practice, other probability sampling methods that utilize randomization can be simpler to obtain.

SYSTEMATIC RANDOM SAMPLING

Systematic random sampling selects a subject near the beginning of the sampling frame list, skips names and selects another subject, skips names and selects the next subject, and so forth. The number of names skipped at each stage depends on the chosen sample size. Here's how it is done:

Systematic Random Sample

> Denote the sample size by n and the population size by N. Let $k = N/n$, the population size divided by the sample size. A ***systematic random sample*** (1) selects a subject at random from the first k names in the sampling frame, and (2) selects every kth subject listed after that one. The number k is called the *skip number*.

Suppose you want a systematic random sample of 100 students from a population of 30,000 students listed in a campus directory. Then, $n = 100$ and $N = 30{,}000$, so $k = 30{,}000/100 = 300$. The population size is 300 times the sample size, so you need to select one of every 300 students. You select one student at random, using random numbers, from the first 300 students in the directory. Then you select every 300th student after the one selected randomly. This produces a sample of size 100. For example, suppose the random number you choose between 001 and 300 is 104. Then, the numbers of the students selected are 104, $104 + 300 = 404$, $404 + 300 = 704$, $704 + 300 = 1004$, $1004 + 300 = 1304$, and so on. The 100th student selected is listed in the last 300 names in the directory.

Systematic random sampling typically provides as good a representation of the population as simple random sampling, because for alphabetic listings such as directories of names, values of most variables fluctuate randomly through the list. With this method, statistical formulas based on simple random sampling are usually valid.

A systematic random sample is not a simple random sample, because all samples of size n are not equally likely. For instance, unlike in a simple random sample, two subjects listed next to each other on the sampling frame list cannot both appear in the sample.

STRATIFIED RANDOM SAMPLING

Another probability sampling method, useful in social science research for studies comparing groups, is ***stratified sampling***.

Stratified Random Sample

A ***stratified random sample*** divides the population into separate groups, called ***strata***, and then selects a simple random sample from each stratum.

Suppose a study in Cambridge, Massachusetts, plans to compare the opinions of registered Democrats and registered Republicans about whether government should guarantee health care to all citizens. Stratifying according to political party registration, the study selects a random sample of Democrats and another random sample of Republicans.

Stratified random sampling is called ***proportional*** if the sampled strata proportions are the same as those in the entire population. For example, in the study of opinions about health care, if 90% of registered voters in Cambridge are Democrats and 10% are Republicans, then the sampling is proportional if the sample size for Democrats is nine times the sample size for Republicans.

Stratified random sampling is called ***disproportional*** if the sampled strata proportions differ from the population proportions. This is useful when the population size for a stratum is relatively small. A group that comprises a small part of the population may not have enough representation in a simple random sample to allow precise inferences. It is not possible to compare accurately Republicans to Democrats, for example, if only 10 people in a sample of size 100 are Republican. By contrast, a disproportional stratified sample of size 100 might randomly sample 50 of each party.

To implement stratification, we must know the stratum into which each subject in the sampling frame belongs. This usually restricts the variables that can be used for forming the strata. The variables must have strata that are easily identifiable. For example, it would be easy to select a stratified sample of a school population using grade level as the stratification variable, but it would be difficult to prepare an adequate sampling frame of city households stratified by household income.

CLUSTER SAMPLING

Simple, systematic, and stratified random sampling are often difficult to implement, because they require a complete sampling frame. Such lists are easy to obtain for sampling cities or hospitals or schools, but more difficult for sampling individuals or families. ***Cluster sampling*** is useful when a complete listing of the population is not available.

Cluster Random Sample

Divide the population into a large number of ***clusters***, such as city blocks. Select a simple random sample of the clusters. Use the subjects in those clusters as the sample.

For example, a study might sample about 1% of the families in a city, using city blocks as clusters. Using a map to identify city blocks, it could select a simple random sample of 1% of the blocks and then sample every family on each block. A study of patient care in mental hospitals in Ontario could first sample mental hospitals (the clusters) in that province and then collect data for patients within those hospitals.

What's the difference between a stratified sample and a cluster sample? A stratified sample uses *every* stratum. The strata are usually groups we want to compare. By contrast, a cluster sample uses a *sample* of the clusters, rather than all of them. In cluster sampling, clusters are merely ways of easily identifying groups of subjects. The goal is not to compare the clusters but rather to use them to obtain a sample. Most clusters are not represented in the eventual sample.

Figure 2.2 illustrates the distinction among sampling subjects (simple random sample), sampling clusters of subjects (cluster random sample), and sampling subjects from within strata (stratified random sample). The figure depicts ways to survey 40 students at a school, to make comparisons among Freshmen, Sophomores, Juniors, and Seniors.

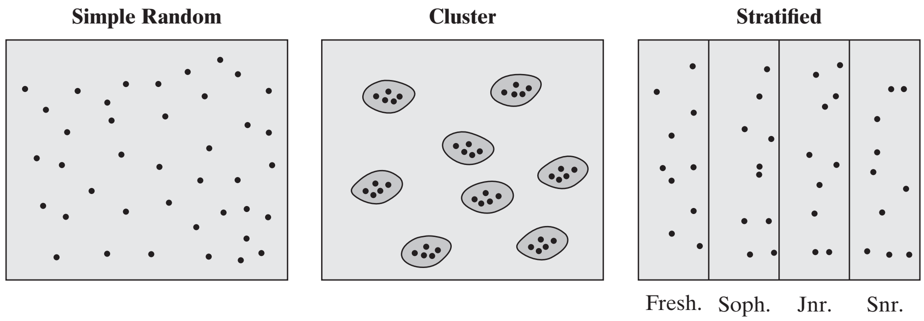


FIGURE 2.2: Ways of Randomly Sampling 40 Students. The figure is a schematic for a simple random sample, a cluster random sample of 8 clusters of students who live together, and a stratified random sample of 10 students from each class (Fresh., Soph., Jnr., Snr.).

MULTISTAGE SAMPLING

When conducting a survey for predicting elections, the Gallup organization often identifies election districts as clusters and takes a simple random sample of them. But then it also takes a simple random sample of households within each selected election district. This is more feasible than sampling *every* household in the chosen districts. This is an example of ***multistage sampling***, which uses combinations of sampling methods.

Here is an example of a multistage sample:

- Treat counties (or census tracts) as clusters and select a random sample of a certain number of them.
- Within each county selected, take a cluster random sample of square-block regions.
- Within each region selected, take a systematic random sample of every 10th house.
- Within each house selected, select one adult at random for the sample.

Multistage samples are common in social science research. They are simpler to implement than simple random sampling but provide a broader sampling of the population than a single method such as cluster sampling.

For statistical inference, stratified samples, cluster samples, and multistage samples use different formulas from the ones in this book. Cluster sampling requires a larger sample to achieve as much inferential precision as simple random sampling. Observations within clusters tend to be similar, because of the tendency of subjects living near one another to have similar values on opinion issues and on economic and demographic variables such as age, income, race, and occupation. So, we need more data to obtain a representative cross section. By contrast, the results for stratified sampling may be more precise than those stated in this textbook for simple random sampling. Books specializing in sampling methodology provide further details (Lohr, 2009; Scheaffer et al., 2011; Thompson, 2012).

2.5 Chapter Summary

Statistical methods analyze data on ***variables***, which are characteristics that vary among subjects. The statistical methods we can use depend on the type of variable:

- Numerically measured variables, such as family income and number of children in a family, are ***quantitative***. They are measured on an *interval scale*.
- Variables taking in a set of categories are ***categorical***. Those measured with unordered categories, such as religious affiliation and province of residence, have a *nominal scale*. Those measured with ordered categories, such as social class and political ideology, have an *ordinal scale* of measurement.
- Variables are also classified as ***discrete***, having possible values that are a set of separate numbers (such as 0, 1, 2, ...), or ***continuous***, having a continuous, infinite set of possible values. Categorical variables, whether nominal or ordinal, are discrete. Quantitative variables can be of either type, but in practice are treated as continuous if they can take a large number of values.

Inferential statistical methods require ***probability samples***, which incorporate randomization in some way. Random sampling allows control over the amount of ***sampling error***, which describes how results can vary from sample to sample. Random samples are much more likely to be representative of the population than are nonprobability samples such as volunteer samples.

- For a ***simple random sample***, every possible sample has the same chance of selection.
- Here are other types of probability sampling: ***Systematic*** random sampling takes every *k*th subject in the sampling frame list. ***Stratified*** random sampling divides the population into groups (strata) and takes a random sample from each stratum. ***Cluster*** random sampling takes a random sample of clusters of subjects (such as city blocks) and uses subjects in those clusters as the sample. ***Multistage*** sampling uses combinations of these methods.

Some social science research studies are ***experimental***, with subjects randomly assigned to different treatments that we want to compare. Most studies, such as ***sample surveys***, are ***observational***. They use available subjects in a sample to observe variables of interest, without any experimental control for randomly assigning subjects to groups we want to compare. We need to be very cautious in making causal conclusions based on inferential analyses with data from observational studies.

Chapter 3 introduces statistics for describing samples and corresponding parameters for describing populations. Hence, its focus is on *descriptive statistics*.

Exercises

Practicing the Basics

2.1. Explain the difference between

(a) Discrete and continuous variables.

(b) Categorical and quantitative variables.

(c) Nominal and ordinal variables.

Why do these distinctions matter for statistical analysis?

2.2. Identify each variable as categorical or quantitative:

(a) Number of pets in family.

(b) County of residence.

(c) Choice of auto (domestic or import).

(d) Distance (in miles) commuted to work.

(e) Choice of diet (vegetarian, nonvegetarian).

(f) Time spent in previous month browsing the World Wide Web.

(g) Ownership of personal computer (yes, no).

(h) Number of people you have known with AIDS (0, 1, 2, 3, 4 or more).

(i) Marriage form of a society (monogamy, polygyny, polyandry).

2.3. Which scale of measurement (nominal, ordinal, or interval) is most appropriate for

(a) Attitude toward legalization of marijuana (favor, neutral, oppose)?

(b) Gender (male, female)?

Chapter

PROBABILITY DISTRIBUTIONS 4

CHAPTER OUTLINE

Compared to most mathematical sciences, statistical science is young. Methods of statistical inference were developed within the past century. By contrast, ***probability***, the subject of this chapter, has a long history. For instance, mathematicians used probability in France in the seventeenth century to evaluate various gambling strategies. Probability is a highly developed subject, but this chapter limits attention to the basics that we'll need for statistical inference.

Following an introduction to probability, we introduce ***probability distributions***, which provide probabilities for all the possible outcomes of a variable. The ***normal distribution***, described by a bell-shaped curve, is the most important probability distribution for statistical inference. The ***sampling distribution*** is a fundamentally important type of probability distribution that we need to conduct statistical inference. It enables us to predict how close a sample mean falls to the population mean. The main reason for the importance of the normal distribution is the remarkable result that sampling distributions are usually bell shaped.

4.1 Introduction to Probability

In Chapter 2, we learned that randomness is a key component of good ways to gather data. For each observation in a random sample or randomized experiment, the possible outcomes are known, but it's uncertain which will occur.

PROBABILITY AS A LONG-RUN RELATIVE FREQUENCY

For a particular possible outcome for a random phenomenon, the ***probability*** of that outcome is the proportion of times that the outcome would occur in a very long sequence of observations.

Probability

> With a random sample or randomized experiment, the ***probability*** that an observation has a particular outcome is the proportion of times that outcome would occur in a very long sequence of like observations.

Later in this chapter, we'll analyze data for the 2014 California gubernatorial election, for which the winner was the Democratic party candidate, Jerry Brown. We'll use an exit poll that interviewed a random sample of voters in that election and asked whom they voted for. Suppose that the population proportion who voted for Brown is 0.60. Then, the probability that a randomly selected person voted for Brown is 0.60.

Why does probability refer to the *long run*? Because when you do not already know or assume some value for a probability, you need a large number of observations to accurately assess it. If you sample only 10 people and they are all right-handed, you can't conclude that the probability of being right-handed equals 1.0.

This book defines a probability as a proportion, so it is a number between 0 and 1. In practice, probabilities are often expressed also as percentages, then falling between 0 and 100. For example, if a weather forecaster says that the probability of rain today is 70%, this means that in a long series of days with atmospheric conditions like those today, rain occurs on 70% of the days.

This *long-run* approach is the standard way to define probability. This definition is not always applicable, however. It is not meaningful, for instance, for the probability that human beings have a life after death, or the probability that intelligent life exists elsewhere in the universe. If you start a new business, you will not have a long run of trials with which to estimate the probability that the business is successful. You must then rely on *subjective* information rather than solely on *objective* data. In the subjective approach, the probability of an outcome is defined to be your degree of belief that the outcome will occur, based on the available information, such as data that may be available from experiences of others. A branch of statistical science uses subjective probability as its foundation. It is called *Bayesian statistics*, in honor of an eighteenth-century British clergyman (Thomas Bayes) who discovered a probability rule on which it is based. We introduce this alternative approach in Section 16.8.

BASIC PROBABILITY RULES

Next, we'll present four rules for finding probabilities. We won't try to explain them with precise, mathematical reasoning, because for our purposes it suffices to have an intuitive feel for what each rule says.

Let $P(A)$ denote the probability of a particular possible outcome denoted by the letter A. Then,

- **$P(\text{not } A) = 1 - P(A)$.**

  If you know the probability a particular outcome occurs, then the probability it does *not* occur is 1 minus that probability. Suppose A represents the outcome that a randomly selected person favors legalization of same-sex marriage. If $P(A) = 0.66$, then $1 - 0.66 = 0.34$ is the probability that a randomly selected person does *not* favor legalization of same-sex marriage.

- **If A and B are distinct possible outcomes (with no overlap), then $P(A \text{ or } B) = P(A) + P(B)$.**

  In a survey to estimate the population proportion of people who favor legalization of marijuana, let A represent the sample proportion estimate being much too low, say more than 0.10 *below* the population proportion. Let B represent the sample proportion estimate being much too high—at least 0.10 *above* the population proportion. These are two distinct possible outcomes. From methods in this chapter, perhaps $P(A) = P(B) = 0.03$. Then, the overall probability the sample proportion is in error by more than 0.10 (without specifying the direction of error) is

$$P(A \text{ or } B) = P(A) + P(B) = 0.03 + 0.03 = 0.06.$$

- **If A and B are possible outcomes, then $P(A \text{ and } B) = P(A) \times P(B \text{ given } A)$.**

  From U.S. Census data, the probability that a randomly selected American adult is married equals 0.56. Of those who are married, General Social Surveys estimate that the probability a person reports being *very happy* when asked to choose among (very happy, pretty happy, not too happy) is 0.40; that is, given you are married, the probability of being very happy is 0.40. So,

$$P(\text{married and very happy}) =$$

$$P(\text{married}) \times P(\text{very happy given married}) = 0.56 \times 0.40 = 0.22.$$

About 22% of the adult population is both married *and* very happy. The probability $P(B \text{ given } A)$ is called a ***conditional probability*** and is often denoted by $P(B \mid A)$.

In some cases, A and B are "independent," in the sense that whether one occurs does not depend on whether the other does. That is, $P(B \text{ given } A) = P(B)$, so the previous rule simplifies:

- **If A and B are independent, then $P(A \text{ and } B) = P(A) \times P(B)$.**

For example, suppose that 60% of a population supports a carbon tax to diminish impacts of carbon dioxide levels on global warming. In random sampling from that population, let A denote the probability that the first person sampled supports the carbon tax and let B denote the probability that the second person sampled supports it. Then $P(A) = 0.60$ and $P(B) = 0.60$. With random sampling, successive observations are independent, so the probability that *both* people support a carbon tax is

$$P(A \text{ and } B) = P(A) \times P(B) = 0.60 \times 0.60 = 0.36.$$

This extends to multiple independent events. For 10 randomly sampled people, the probability that all 10 support a carbon tax is $0.60 \times 0.60 \times \cdots \times 0.60 = (0.60)^{10} = 0.006$.

4.2 Probability Distributions for Discrete and Continuous Variables

A variable can take at least two different values. For a random sample or randomized experiment, each possible outcome has a probability that it occurs. The variable itself is sometimes then referred to as a ***random variable***. This terminology emphasizes that the outcome varies from observation to observation according to random variation that can be summarized by probabilities. For simplicity, we'll continue to use the "variable" terminology regardless of whether the variation has a random aspect.

Recall (from Section 2.1) that a variable is *discrete* if the possible outcomes are a set of separate values, such as a variable expressed as "the number of …" with possible values 0, 1, 2, …. It is *continuous* if the possible outcomes are an infinite continuum, such as all the real numbers between 0 and 1. A ***probability distribution*** lists the possible outcomes and their probabilities.

PROBABILITY DISTRIBUTIONS FOR DISCRETE VARIABLES

The probability distribution of a *discrete* variable assigns a probability to each possible value of the variable. Each probability is a number between 0 and 1. The sum of the probabilities of all possible values equals 1.

Let $P(y)$ denote the probability of a possible outcome for a variable y. Then,

$$0 \le P(y) \le 1 \quad \text{and} \quad \sum_{\text{all } y} P(y) = 1,$$

where the sum is over all the possible values of the variable.

Example 4.1

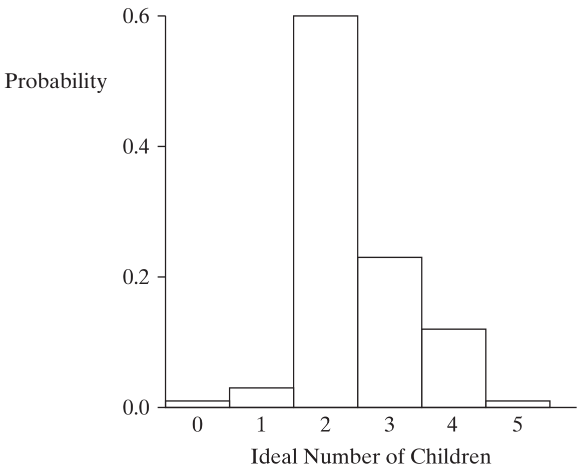
Ideal Number of Children for a Family Let y denote the response to the question "What do you think is the ideal number of children for a family to have?" This is a discrete variable, taking the possible values 0, 1, 2, 3, and so forth. According to recent General Social Surveys, for a randomly chosen person in the United States the probability distribution of y is approximately as Table 4.1 shows. The table displays the recorded y-values and their probabilities. For instance, $P(4)$, the probability that $y = 4$ children is regarded as ideal, equals 0.12. Each probability in Table 4.1 is between 0 and 1, and the sum of the probabilities equals 1. ■

TABLE 4.1: Probability Distribution of y = Ideal Number of Children for a Family

y	$P(y)$
0	0.01
1	0.03
2	0.60
3	0.23
4	0.12
5	0.01
Total	1.00

A *histogram* can portray the probability distribution. The rectangular bar over a possible value of the variable has height equal to the probability of that value. Figure 4.1 is a histogram for the probability distribution of the ideal number of children, from Table 4.1. The bar over the value 4 has height 0.12, the probability of the outcome 4.

FIGURE 4.1: Histogram for the Probability Distribution of the Ideal Number of Children for a Family



PROBABILITY DISTRIBUTIONS FOR CONTINUOUS VARIABLES

Continuous variables have an infinite continuum of possible values. Probability distributions of continuous variables assign probabilities to *intervals* of numbers. The probability that a variable falls in any particular interval is between 0 and 1, and the probability of the interval containing all the possible values equals 1.

A graph of the probability distribution of a continuous variable is a smooth, continuous curve. The *area* under the curve[1] for an interval of values represents the probability that the variable takes a value in that interval.

Example 4.2

Commuting Time to Work A recent U.S. Census Bureau study about commuting time for workers in the United States who commute to work[2] measured y = travel time, in minutes. The probability distribution of y provides probabilities such as $P(y < 15)$, the probability that travel time is less than 15 minutes, or $P(30 < y < 60)$, the probability that travel time is between 30 and 60 minutes.

Figure 4.2 portrays the probability distribution of y. The shaded area in the figure refers to the region of values higher than 45. This area equals 15% of the total area under the curve, representing the probability of 0.15 that commuting time is more than 45 minutes. Those regions in which the curve has relatively high height have the values most likely to be observed. ■

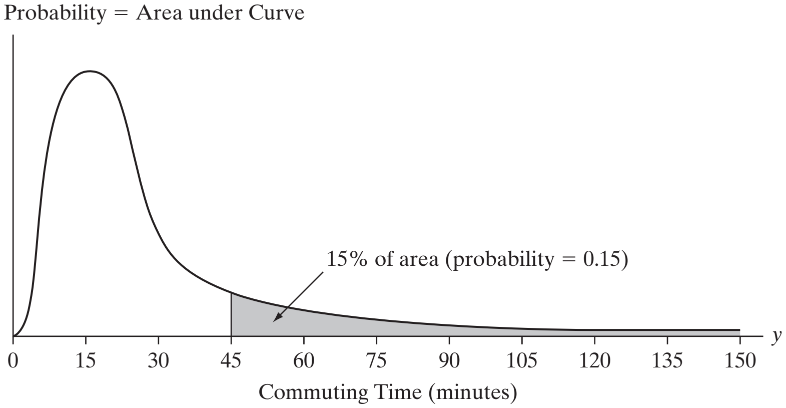


FIGURE 4.2: Probability Distribution of Commuting Time to Work. The area under the curve between two points represents the probability of that interval of values.

PARAMETERS DESCRIBE PROBABILITY DISTRIBUTIONS

Some probability distributions have formulas for calculating probabilities. For others, tables or software provide the probabilities. Section 4.3 shows how to find probabilities for the most important probability distribution.

Section 3.1 introduced the *population distribution* of a variable. This is, equivalently, the probability distribution of the variable for a subject selected randomly from the population. For example, if 0.12 is the population proportion of adults who believe the ideal number of children is 4, then the probability that an adult selected randomly from that population believes this is also 0.12.

Like a population distribution, a probability distribution has *parameters* describing center and variability. The *mean* describes center and the *standard deviation* describes variability. The parameter values are the values these measures would assume, *in the long run*, if the randomized experiment or random sample repeatedly took observations on the variable y having that probability distribution.

For example, suppose we take observations from the distribution in Table 4.1. Over the long run, we expect $y = 0$ to occur 1% of the time, $y = 1$ to occur 3% of the time, and so forth. In 100 observations, for instance, we expect about

$$\text{one } 0,\ 3\ 1's,\ 60\ 2's,\ 23\ 3's,\ 12\ 4's, \text{ and one } 5.$$

[1] Mathematically, this calculation uses integral calculus. The probability that y falls in the interval between points a and b is the integral over that interval of the function for the curve.

[2] See www.census.gov/hhes/commuting.

In that case, since the mean equals the total of the observations divided by the sample size, the mean equals

$$\frac{0(1) + 1(3) + 2(60) + 3(23) + 4(12) + 5(1)}{100} = \frac{245}{100} = 2.45.$$

This calculation has the form

$$0(0.01) + 1(0.03) + 2(0.60) + 3(0.23) + 4(0.12) + 5(0.01),$$

the sum of the possible outcomes times their probabilities. In fact, for any discrete variable y, the mean of its probability distribution has this form.

Mean of a Probability Distribution (Expected Value)

> The ***mean of the probability distribution*** for a discrete variable y is
>
> $$\mu = \sum yP(y).$$
>
> The sum is taken over all possible values of the variable. This parameter is also called the ***expected value of y*** and denoted by $E(y)$.

For Table 4.1, for example,

$$\begin{aligned}\mu = \sum yP(y) &= 0P(0) + 1P(1) + 2P(2) + 3P(3) + 4P(4) + 5P(5)\\ &= 0(0.01) + 1(0.03) + 2(0.60) + 3(0.23) + 4(0.12) + 5(0.01)\\ &= 2.45.\end{aligned}$$

This is also the *expected value* of y, $E(y) = 2.45$. The terminology reflects that $E(y)$ represents what we expect for the average value of y in a long series of observations.

The ***standard deviation*** of a probability distribution, denoted by σ, measures its variability. The more spread out the distribution, the larger the value of σ. The Empirical Rule (Section 3.3) helps us to interpret σ. If a probability distribution is bell shaped, about 68% of the probability falls between $\mu - \sigma$ and $\mu + \sigma$, about 95% falls between $\mu - 2\sigma$ and $\mu + 2\sigma$, and all or nearly all falls between $\mu - 3\sigma$ and $\mu + 3\sigma$.

The standard deviation is the square root of the ***variance*** of the probability distribution. The variance measures the average squared deviation of an observation from the mean. That is, it is the expected value of $(y - \mu)^2$. In the discrete case, the formula is

$$\sigma^2 = E(y - \mu)^2 = \sum (y - \mu)^2 P(y).$$

We shall not need to calculate σ^2, so we shall not further consider this formula here.

4.3 The Normal Probability Distribution

Some probability distributions are important because they approximate well sample data in the real world. Some are important because of their uses in statistical inference. This section introduces the ***normal probability distribution***, which is important for both reasons.

Normal Distribution

> The ***normal distribution*** is symmetric, bell shaped, and characterized by its mean μ and standard deviation σ. The probability within any particular number of standard deviations of μ is the same for all normal distributions. This probability (rounded off) equals 0.68 within 1 standard deviation, 0.95 within 2 standard deviations, and 0.997 within 3 standard deviations.

Each normal distribution[3] is specified by its mean μ and standard deviation σ. For any real number for μ and any nonnegative number for σ, there is a normal distribution having that mean and standard deviation. Figure 4.3 illustrates this. Essentially the entire distribution falls between $\mu - 3\sigma$ and $\mu + 3\sigma$.

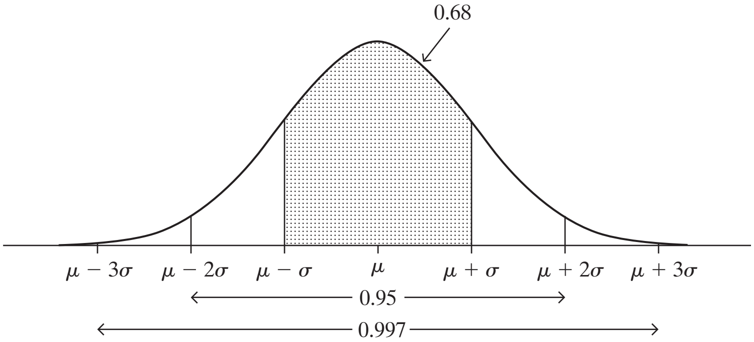


FIGURE 4.3: For Every Normal Distribution, the Probability (Rounded) Equals 0.68 within σ of μ, 0.95 within 2σ of μ, and 0.997 within 3σ of μ

For example, heights of adult females in North America have approximately a normal distribution with $\mu = 65.0$ inches and $\sigma = 3.5$. The probability is nearly 1.0 that a randomly selected female has height between $\mu - 3\sigma = 65.0 - 3(3.5) = 54.5$ inches and $\mu + 3\sigma = 65.0 + 3(3.5) = 75.5$ inches. Adult male height has a normal distribution with $\mu = 70.0$ and $\sigma = 4.0$ inches. So, the probability is nearly 1.0 that a randomly selected male has height between $\mu - 3\sigma = 70.0 - 3(4.0) = 58$ inches and $\mu + 3\sigma = 70.0 + 3(4.0) = 82$ inches. See Figure 4.4.

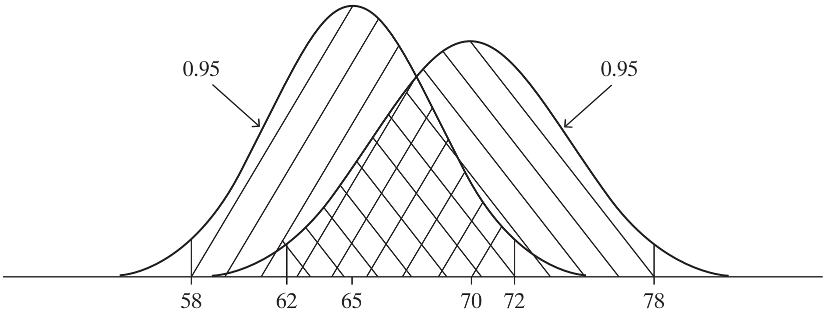


FIGURE 4.4: Normal Distributions for Women's Height ($\mu = 65, \sigma = 3.5$) and for Men's Height ($\mu = 70, \sigma = 4.0$)

FINDING NORMAL PROBABILITIES: TABLES, SOFTWARE, AND APPLETS

For the normal distribution, for each fixed number z, the probability that is within z standard deviations of the mean depends only on the value of z. This is the area under the normal curve between $\mu - z\sigma$ and $\mu + z\sigma$. For every normal distribution, this probability is 0.68 for $z = 1$, 0.95 for $z = 2$, and nearly 1.0 for $z = 3$.

For a normal distribution, the probability concentrated within $z\sigma$ of μ is the same for all normal curves even if z is not a whole number—for instance, $z = 1.43$ instead of 1, 2, or 3. Table A, also shown next to the inside back cover, determines probabilities

[3] More technically, the normal distribution with mean μ and standard deviation σ is represented by a bell-shaped curve that has the formula

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-[(y-\mu)^2/2\sigma^2]}.$$

for any region of values. It tabulates the probability for the values falling in the right tail, at least z standard deviations above the mean. The left margin column of the table lists the values for z to one decimal point, with the second decimal place listed above the columns.

Table 4.2 displays a small excerpt from Table A. The probability for $z = 1.43$ falls in the row labeled 1.4 and in the column labeled .03. It equals 0.0764. This means that for every normal distribution, the right-tail probability above $\mu + 1.43\sigma$ (i.e., more than 1.43 standard deviations above the mean) equals 0.0764.

TABLE 4.2: Part of Table A Displaying Normal Right-Tail Probabilities

	Second Decimal Place of z									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
					....					
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559

Since the entries in Table A are probabilities for the right half of the normal distribution above $\mu + z\sigma$, they fall between 0 and 0.50. By the symmetry of the normal curve, these right-tail probabilities also apply to the left tail below $\mu - z\sigma$. For example, the probability below $\mu - 1.43\sigma$ also equals 0.0764. The left-tail probabilities are called ***cumulative probabilities***.

We can also use statistical software to find normal probabilities. The free software R has a function `pnorm` that gives the cumulative probability falling below $\mu + z\sigma$. For example, *pnorm(2.0)* provides the cumulative probability falling below $\mu + 2.0\sigma$:

```
> pnorm(2.0)   # cumulative probability below mu + 2.0(sigma)
[1] 0.97724987 # right-tail probability = 1 - 0.977 = 0.023
```

In the Stata software, we can use the `display normal` command:

```
. display normal(2.0) # cumulative probability below mu + 2.0(sigma)
.97724987             # right-tail probability = 1 - 0.977 = 0.023
```

We subtract the cumulative probability from 1 to find the right-tail probability above $\mu + 2.0\sigma$. That is, the probability $1 - 0.97725 = 0.02275$ falls more than two standard deviations above the mean. By the symmetry of the normal distribution, this is also the probability falling more than two standard deviations below the mean. The probability falling *within* two standard deviations of the mean is $1 - 2(0.02275) = 0.954$. (Here, we've used rule (1) of the probability rules at the end of Section 4.1, that $P(\text{not } A) = 1 - P(A)$.) You can also find normal probabilities with SPSS and SAS software.

Normal probabilities are also available on the Internet, such as with the easy-to-use *Normal Distribution* applet[4] for which there is a link at `www.artofstat.com/webapps.html`. See Figure 4.5.

[4] This is one of several applets we shall use that were developed by Prof. Bernhard Klingenberg for the text *Statistics: The Art and Science of Learning from Data*, 4th ed., by A. Agresti, C. Franklin, and B. Klingenberg (Pearson, 2017).

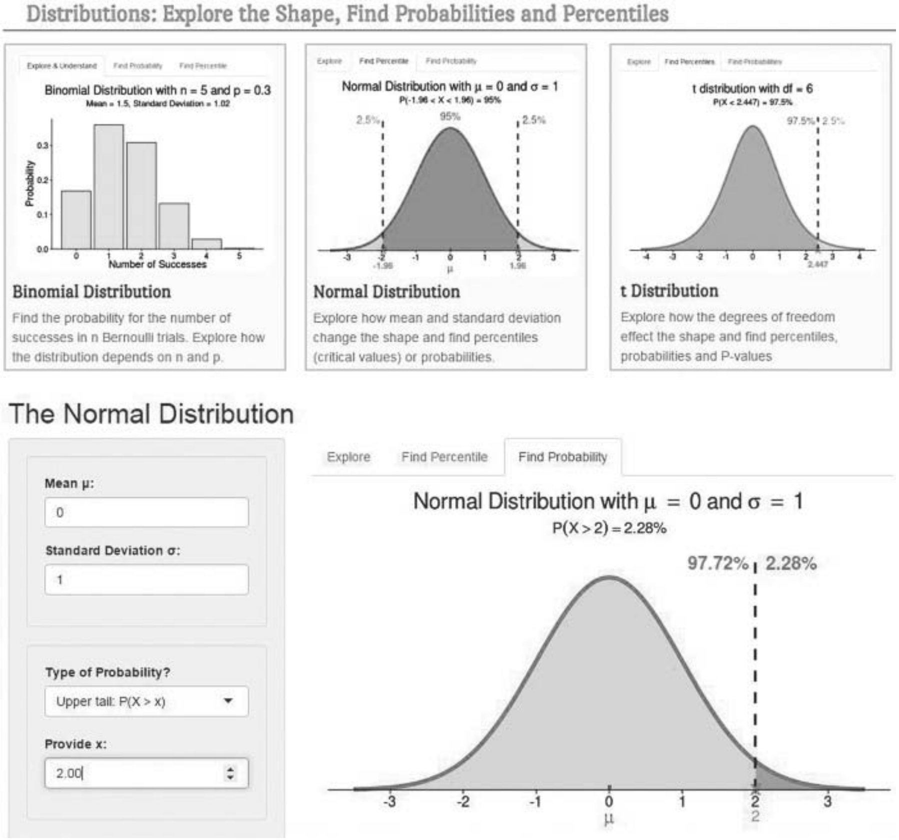


FIGURE 4.5: Go to `www.artofstat.com/webapps.html` for Useful Internet Applets. For normal probabilities, click on *Normal Distribution*. The applet denotes a variable by X and a possible value by x. Tail probabilities do not depend on values entered for μ and σ.

Distributions: Explore the Shape, Find Probabilities and Percentiles

Binomial Distribution

Find the probability for the number of successes in n Bernoulli trials. Explore how the distribution depends on n and p.

Normal Distribution

Explore how mean and standard deviation change the shape and find percentiles (critical values) or probabilities.

t Distribution

Explore how the degrees of freedom effect the shape and find percentiles, probabilities and P-values

The Normal Distribution

Mean μ: 0

Standard Deviation σ: 1

Type of Probability? Upper tail: P(X > x)

Provide x: 2.00

NORMAL PROBABILITIES AND THE EMPIRICAL RULE

Probabilities for the normal distribution apply *approximately* to other bell-shaped distributions. They yield the probabilities for the Empirical Rule. Recall (page 44) that that rule states that for bell-shaped histograms, about 68% of the data fall within one standard deviation of the mean, 95% within two standard deviations, and all or nearly all within three standard deviations. For example, we've just used software to find that for normal distributions the probability falling within two standard deviations of the mean is 0.954. For one and for three standard deviations, we find central probabilities of 0.683 and 0.997, respectively.

The approximate percentages in the Empirical Rule are the actual percentages for the normal distribution, rounded to two decimal places. The Empirical Rule stated the percentages as being *approximate* rather than *exact*. Why? Because that rule referred to *all approximately bell-shaped distributions*, not only the normal distribution. Not all bell-shaped distributions are normal, only those described by the formula shown in the footnote on page 73. We won't need that formula, but we will use probabilities for it throughout the text.

FINDING z-VALUES FOR CERTAIN TAIL PROBABILITIES

Many inferential methods use z-values corresponding to certain normal curve probabilities. This entails the reverse use of Table A or software or applets. Starting with a tail probability, we find the z-value that provides the number of standard deviations that that number falls from the mean.

To illustrate, let's first use Table A to find the z-value having a right-tail probability of 0.025. We look up 0.025 in the body of Table A, which contains tail probabilities. It corresponds to $z = 1.96$ (i.e., we find .025 in the row of Table A labeled 1.9 and in the column labeled .06). This means that a probability of 0.025 falls above $\mu + 1.96\sigma$. Similarly, a probability of 0.025 falls below $\mu - 1.96\sigma$. So, a total probability of $0.025 + 0.025 = 0.050$ falls more than 1.96σ from μ. We saw in the previous subsection that 95% of a normal distribution falls within two standard deviations of the mean. More precisely, 0.954 falls within 2.00 standard deviations, and here we've seen that 0.950 falls within 1.96 standard deviations.

R software has a function `qnorm` that gives the z-value for a particular cumulative probability. The right-tail probability of 0.025 corresponds to a cumulative probability of $1 - 0.025 = 0.975$, for which the z-value is

```
> qnorm(0.975) # q denotes "quantile"; .975 quantile = 97.5 percentile
[1] 1.959964   # The z-value is 1.96, rounded to two decimals
```

Here is how you can find the z-value for a cumulative probability using the Stata software:

```
  . display invnormal(0.975)   /* invnormal = "inverse normal" */
  1.959964
```

The `qnorm` function in R is equivalent to the `invnormal` (inverse normal) function in Stata. You can also find this z-value using an Internet applet, such as Figure 4.6 shows with the *Normal Distribution* applet at `www.artofstat.com/webapps.html`. It is also possible to find z-values with SPSS and SAS software.

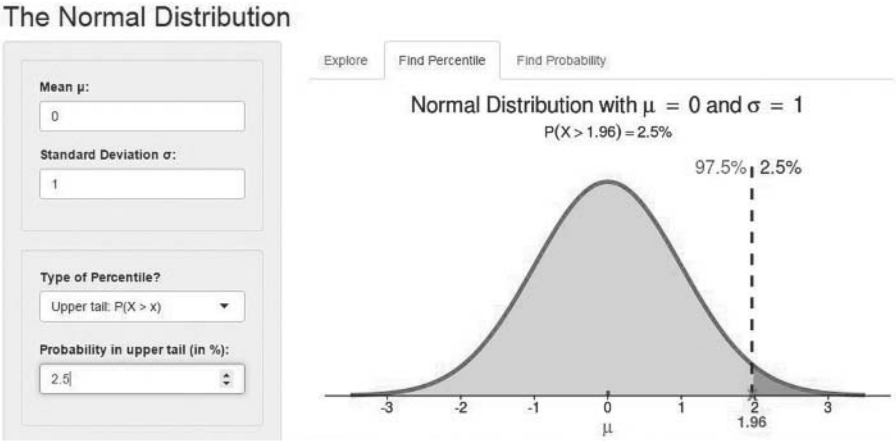


FIGURE 4.6: Using the *Normal Distribution* Applet at `www.artofstat.com/webapps.html` to Find the z-value for a Normal Tail Probability of 0.025 (i.e., 2.5 percent)

The Normal Distribution

Mean μ: 0

Standard Deviation σ: 1

Type of Percentile? Upper tail: P(X > x)

Probability in upper tail (in %): 2.5

Normal Distribution with μ = 0 and σ = 1

P(X > 1.96) = 2.5%

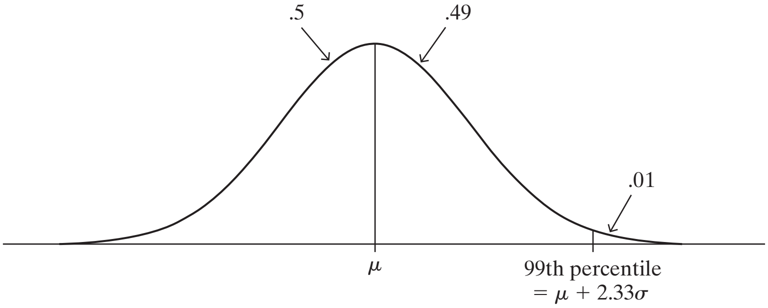
To check that you understand this reasoning, use Table A, software, or the applet to verify that the z-value for a right-tail probability of (1) 0.05 is $z = 1.64$, (2) 0.01 is $z = 2.33$, and (3) 0.005 is $z = 2.58$. Show that 90% of a normal distribution falls between $\mu - 1.64\sigma$ and $\mu + 1.64\sigma$.

Example 4.3

Finding the 99th Percentile of IQ Scores Stanford-Binet IQ scores have approximately a normal distribution with mean = 100 and standard deviation = 16. What is the 99th percentile of IQ scores? In other words, what is the IQ score that falls above 99% of the scores?

To answer this, we need to find the value of z such that $\mu + z\sigma$ falls above 99% of a normal distribution. Now, for $\mu + z\sigma$ to represent the 99th percentile, the probability below $\mu + z\sigma$ must equal 0.99, by the definition of a percentile. So, 1% of the distribution is above the 99th percentile. The right-tail probability equals 0.01, as Figure 4.7 shows.

FIGURE 4.7: The 99th Percentile for a Normal Distribution Has 99% of the Distribution below that Point and 1% above It



With Table A, software, or the Internet, you can find that the z-value for a cumulative probability of 0.99 or right-tail probability of 0.01 is $z = 2.33$. Thus, the 99th percentile is 2.33 standard deviations above the mean. In summary, 99% of any normal distribution is located below $\mu + 2.33\sigma$.

For IQ scores with mean = 100 and standard deviation = 16, the 99th percentile equals

$$\mu + 2.33\sigma = 100 + 2.33(16) = 137.$$

That is, about 99% of IQ scores fall below 137. ■

To check that you understand the reasoning above, show that the 95th percentile of a normal distribution is $\mu + 1.64\sigma$, and show that the 95th percentile for the IQ distribution equals 126.

z-SCORE REPRESENTS THE NUMBER OF STANDARD DEVIATIONS FROM THE MEAN

The z symbol in a normal table refers to the distance between a possible value y of a variable and the mean μ of its probability distribution, in terms of the *number of standard deviations* that y falls from μ.

For example, scores on each portion of the Scholastic Aptitude Test (SAT) have traditionally been approximately normal with mean $\mu = 500$ and standard deviation $\sigma = 100$. The test score of $y = 650$ has a z-score of $z = 1.50$, because 650 is 1.50 standard deviations above the mean. In other words, $y = 650 = \mu + z\sigma = 500 + z(100)$, where $z = 1.50$.

For sample data, Section 3.4 introduced the z-score as a measure of position. Let's review how to find it. The distance between y and the mean μ equals $y - \mu$. The z-score expresses this difference in units of standard deviations.

***z*-Score**

The z-score for a value y of a variable is the *number of standard deviations* that y falls from the mean. For a probability distribution with mean μ and standard deviation σ, it equals

$$z = \frac{\text{Variable value} - \text{Mean}}{\text{Standard deviation}} = \frac{y - \mu}{\sigma}.$$

To illustrate, when $\mu = 500$ and $\sigma = 100$, a value of $y = 650$ has the z-score of

$$z = \frac{y - \mu}{\sigma} = \frac{650 - 500}{100} = 1.50.$$

Positive z-scores occur when the value for y falls *above* the mean μ. *Negative* z-scores occur when the value for y falls *below* the mean. For example, for SAT scores with $\mu = 500$ and $\sigma = 100$, a value of $y = 350$ has a z-score of

$$z = \frac{y - \mu}{\sigma} = \frac{350 - 500}{100} = -1.50.$$

The test score of 350 is 1.50 standard deviations below the mean. The value $y = 350$ falls below the mean, so the z-score is negative.

The next example shows that z-scores provide a useful way to compare positions for different normal distributions.

Example 4.4

Comparing SAT and ACT Test Scores Suppose that when you applied to college, you took a SAT exam, scoring 550. Your friend took the ACT exam, scoring 30. If the SAT has $\mu = 500$ and $\sigma = 100$ and the ACT has $\mu = 18$ and $\sigma = 6$, then which score is relatively better?

We cannot compare the test scores of 550 and 30 directly, because they have different scales. We convert them to z-scores, analyzing how many standard deviations each falls from the mean. The SAT score of $y = 550$ converts to a z-score of

$$z = \frac{y - \mu}{\sigma} = \frac{550 - 500}{100} = 0.50.$$

The ACT score of $y = 30$ converts to a z-score of $(30 - 18)/6 = 2.0$.

The ACT score of 30 is relatively higher than the SAT score of 650, because 30 is 2.0 standard deviations above its mean whereas 550 is only 0.5 standard deviations above its mean. The SAT and ACT scores both have approximate normal distributions. From Table A, $z = 2.0$ has a right-tail probability of 0.0228 and $z = 0.5$ has a right-tail probability of 0.3085. Of all students taking the ACT, only about 2% scored higher than 30, whereas of all students taking the SAT, about 31% scored higher than 550. In this relative sense, the ACT score is higher. ■

USING z-SCORES TO FIND PROBABILITIES OR y-VALUES

Here's a summary of how we use z-scores:

- If we have a value y and need to find a probability, convert y to a z-score using $z = (y - \mu)/\sigma$, and then convert z to the probability of interest using a table of normal probabilities, software, or the Internet.
- If we have a probability and need to find a value of y, convert the probability to a tail probability (or cumulative probability) and find the z-score (using a normal table, software, or the Internet), and then evaluate $y = \mu + z\sigma$.

For example, we used the equation $z = (y - \mu)/\sigma$ to determine how many standard deviations a SAT test score of 650 fell from the mean of 500, when $\sigma = 100$ (namely, 1.50). Example 4.3 used the equation $y = \mu + z\sigma$ to find a percentile score for a normal distribution of IQ scores.

THE STANDARD NORMAL DISTRIBUTION

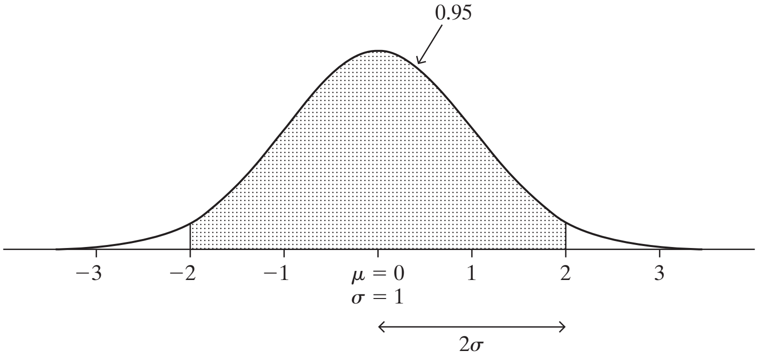
Many inferential statistical methods use a particular normal distribution, called the ***standard normal distribution***.

Standard Normal Distribution

> The ***standard normal distribution*** is the normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

For the standard normal distribution, the number falling z standard deviations above the mean is $\mu + z\sigma = 0 + z(1) = z$. It is simply the z-score itself. For instance, the value of 2 is two standard deviations above the mean, and the value of -1.3 is 1.3 standard deviations below the mean. The original values are the same as the z-scores. See Figure 4.8.

FIGURE 4.8: The Standard Normal Distribution Has Mean 0 and Standard Deviation 1. Its ordinary scores are the same as its z-scores.



When the values for an arbitrary normal distribution are converted to z-scores, those z-scores are centered around 0 and have a standard deviation of 1. The z-scores have the standard normal distribution.

z-Scores and the Standard Normal Distribution

> If a variable has a normal distribution, and if its values are converted to z-scores by subtracting the mean and dividing by the standard deviation, then the z-scores have the standard normal distribution.

Suppose we convert each SAT score y to a z-score by using $z = (y - 500)/100$. For instance, $y = 650$ converts to $z = 1.50$, and $y = 350$ converts to $z = -1.50$. Then, the entire set of z-scores has a normal distribution with a mean of 0 and a standard deviation of 1. This is the standard normal distribution.

Many inferential methods convert values of statistics to z-scores and then to normal curve probabilities. We use z-scores and normal probabilities often throughout the rest of the book.

BIVARIATE PROBABILITY DISTRIBUTIONS: COVARIANCE AND CORRELATION*

Section 3.5 introduced *bivariate* descriptive statistics that apply to a pair of variables. An example is the sample correlation. Likewise, ***bivariate probability distributions*** determine joint probabilities for pairs of random variables. For example, the *bivariate normal distribution* generalizes the bell curve over the real line for a single variable y to a bell-shaped surface in three dimensions over the plane for possible values of two variables (x, y).

Each variable in a bivariate distribution has a mean and a standard deviation. Denote them by (μ_x, σ_x) for x and by (μ_y, σ_y) for y. The way that x and y vary together is described by their ***covariance***, which is defined to be

$$\text{Covariance}(x, y) = E[(x - \mu_x)(y - \mu_y)],$$

which represents the average of the cross products about the population means (weighted by their probabilities). If y tends to fall *above* its mean when x falls *above* its mean, the covariance is *positive*. If y tends to fall *below* its mean when x falls *above* its mean, the covariance is *negative*.

The covariance can be any real number. For interpretation, it is simpler to use

$$\text{Correlation}(x, y) = \frac{\text{Covariance}(x, y)}{(\text{Standard deviation of } x)(\text{Standard deviation of } y)}.$$

But this equals

$$\frac{E[(x - \mu_x)(y - \mu_y)]}{\sigma_x \sigma_y} = E\left[\left(\frac{x - \mu_x}{\sigma_x}\right)\left(\frac{y - \mu_y}{\sigma_y}\right)\right] = E(z_x z_y),$$

where $z_x = (x - \mu_x)/\sigma_x$ denotes the z-score for the variable x and $z_y = (y - \mu_y)/\sigma_y$ denotes the z-score for the variable y. That is, the population correlation equals the average cross product of the z-score for x times the z-score for y. It falls between -1 and $+1$. It is positive when positive z-scores for x tend to occur with positive z-scores for y and when negative z-scores for x tend to occur with negative z-scores for y.

We shall not need to calculate these expectations. We can use software to find sample values, as we showed in Table 3.10 for the correlation.

4.4 Sampling Distributions Describe How Statistics Vary

We've seen that probability distributions summarize probabilities of possible outcomes for a variable. Let's now look at an example that illustrates the connection between statistical inference and probability calculations.

Example 4.5

Predicting an Election from an Exit Poll Television networks sample voters on election day to help them predict the winners early. For the fall 2014 election for Governor of California, CBS News[5] reported results of an exit poll of 1824 voters. They stated that 60.5% of their *sample* reported voting for the Democratic party candidate, Jerry Brown. In this example, the probability distribution for a person's vote would state the probability that a randomly selected voter voted for Brown. This equals the proportion of the *population* of voters who voted for him. When the exit poll was taken, this was an unknown population parameter.

To judge whether this is sufficient information to predict the outcome of the election, the network can ask, "Suppose only half the population voted for Brown. Would it then be surprising that 60.5% of the sampled individuals voted for him?" If this would be very unlikely, the network infers that Brown received more than half the population votes and won the election. The inference about the election outcome is based on finding the probability of the sample result under the supposition that the population parameter, the percentage of voters preferring Brown, equals 50%. ■

[5] See www.cbsnews.com/elections/2014/governor/california/exit/.

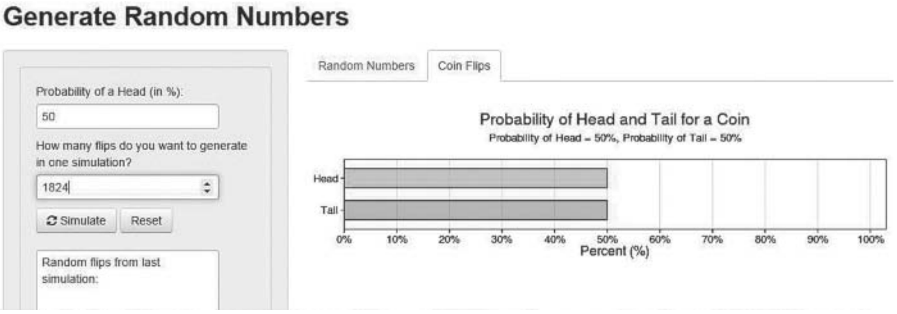
About 7.3 million people voted in this race. The exit poll sampled only 1824 voters, yet TV networks used it to predict that Brown would win. How could there possibly have been enough information from this poll to make a prediction? We next see justification for making a prediction.

SIMULATING THE SAMPLING PROCESS

A ***simulation*** can show us how close an exit poll result tends to be to the population proportion voting for a candidate. One way to simulate the vote of a voter randomly chosen from the population is to select a random number using software. Suppose exactly 50% of the population voted for Brown and 50% voted for the Republican candidate, Neel Kashkari. Identify all 50 two-digit numbers between 00 and 49 as Democratic votes and all 50 two-digit numbers between 50 and 99 as Republican votes. Then, each candidate has a 50% chance of selection on each choice of two-digit random number. For instance, the first two digits of the first column of the random numbers table on page 15 provide the random numbers 10, 53, 24, and 42. So, of the first four voters selected, three voted Democratic (i.e., have numbers between 00 and 49) and one voted Republican. Selecting 1824 two-digit random numbers simulates the process of observing the votes of a random sample of 1824 voters of the much larger population (which is actually treated as infinite in size).

To do this, we can use software that generates random numbers or that uses such numbers to simulate flipping a coin repeatedly, where we regard one outcome (say, head) as representing a person who votes for the Democrat and the other outcome (say, tail) as representing a person who votes for the Republican. Here is how we simulated, using an applet on the Internet. We suggest you try this also, to see how it works.

- Go to `www.artofstat.com/webapps.html` and click on *Random Numbers*.
- Click on *Coin Flips*.
- The box for *Probability of a Head (in %)* should say 50. Then, random numbers between 00 and 49 correspond to head and random numbers between 50 and 99 correspond to tail. In the box for *How many flips do you want to generate in one simulation?* enter 1824. See Figure 4.9.
- Click *Simulate*.



Probability of a Head (in %): 50

How many flips do you want to generate in one simulation? 1824

Random flips from last simulation:

FIGURE 4.9: The *Random Numbers* Applet for Simulating at `www.artofstat.com/webapps.html`. When we click on *Simulate*, we see the results of flipping a coin 1824 times when the probability on each flip of getting a head equals 0.50.

When we performed this simulation, we got 901 heads (Democratic votes) and 923 tails (Republican votes). The sample proportion of Democratic votes was $901/1824 = 0.494$, quite close to the population proportion of 0.50. This particular estimate was good. Were we merely lucky? We repeated the process and simulated

1824 more flips. (In this app, click again on *Simulate*.) This time the sample proportion of Democratic votes was 0.498, also quite good.

Using software,[6] we next performed this process of picking 1824 people 10,000 times so that we could search for a pattern in the results. Figure 4.10 shows a histogram of the 10,000 values of the sample proportion. Nearly all the simulated proportions fell between 0.46 and 0.54, that is, within 0.04 of the population proportion of 0.50. Apparently a sample of size 1824 provides quite a good estimate of a population proportion.

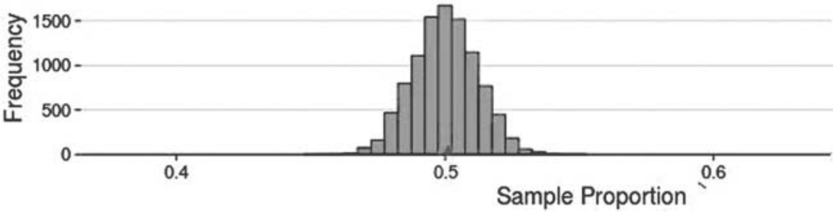


FIGURE 4.10: Results of 10,000 Simulations of the Sample Proportion Favoring the Democratic Candidate, for Random Samples of 1824 Subjects from a Population in which Half Voted for the Democrat and Half Voted for the Republican. In nearly all cases, the sample proportion fell within 0.04 of the population proportion of 0.50 (i.e., between 0.46 and 0.54).

In summary, if half the population of voters had voted for Brown, we would have expected between 46% and 54% of voters in an exit poll of size 1824 to have voted for him. It would have been very unusual to observe 60.5% voting for him, as happened in the actual exit poll. If *less than half* the population had voted for Brown, it would have been even more unusual to have this outcome. This is the basis of the network's exit poll prediction that Brown won the election.

You can perform this simulation using *any* population proportion value, corresponding to flipping a coin in which head and tail have different probabilities. For instance, you could simulate sampling when the population proportion voting for the Democrat is 0.45 by changing the probability of a head in the applet to 45%. Likewise, we could change the size of each random sample in the simulation to study the impact of the sample size. From results of the next section, for a random sample of size 1824 the sample proportion has probability close to 1 of falling within 0.04 of the population proportion, regardless of its value.

REPRESENTING SAMPLING VARIABILITY BY A SAMPLING DISTRIBUTION

Voter preference is a variable, varying among voters. Likewise, so is the sample proportion voting for some candidate a variable: Before the sample is obtained, its value is unknown, and that value varies from sample to sample. If we could select several random samples of size $n = 1824$ each, a certain predictable amount of variation would occur in the sample proportion values. A probability distribution with appearance similar to Figure 4.10 describes the variation that occurs from repeatedly selecting samples of a certain size n and forming a particular statistic. This distribution is called a ***sampling distribution***. It also provides probabilities of the possible values of the statistic for a *single* sample of size n.

Sampling Distribution

> A ***sampling distribution*** of a statistic (such as a sample proportion or a sample mean) is the probability distribution that specifies probabilities for the possible values the statistic can take.

[6] The *Sampling Distribution for the Sample Proportion* applet at `www.artofstat.com/webapps.html` does this efficiently.

Each sample statistic has a sampling distribution. There is a sampling distribution of a sample mean, a sampling distribution of a sample proportion, a sampling distribution of a sample median, and so forth. A sampling distribution is merely a type of probability distribution. Unlike the probability distributions studied so far, a sampling distribution specifies probabilities not for individual observations but for possible values of a statistic computed from the observations. A sampling distribution allows us to calculate, for example, probabilities about the sample proportions of individuals in an exit poll who voted for the different candidates. Before the voters are selected for the exit poll, this is a variable. It has a sampling distribution that describes the probabilities of the possible values.

The sampling distribution is important in inferential statistics because it helps us predict how close a statistic falls to the parameter it estimates. From Figure 4.10, for instance, with a sample of size 1824 the probability is apparently close to 1 that a sample proportion falls within 0.04 of the population proportion.

Example 4.6

Constructing a Sampling Distribution It is sometimes possible to construct the sampling distribution without resorting to simulation or complex mathematical derivations. To illustrate, we construct the sampling distribution of the sample proportion for an exit poll of $n = 4$ voters from a population in which half voted for each candidate. (Such a small n would not be used in practice, but it enables us to more easily explain this process.)

We use a symbol with four entries to represent the votes for a potential sample of size 4. For instance, (R, D, D, R) represents a sample in which the first and fourth subjects voted for the Republican and the second and third subjects voted for the Democrat. The 16 possible samples are

(R, R, R, R) (R, R, R, D) (R, R, D, R) (R, D, R, R)
(D, R, R, R) (R, R, D, D) (R, D, R, D) (R, D, D, R)
(D, R, R, D) (D, R, D, R) (D, D, R, R) (R, D, D, D)
(D, R, D, D) (D, D, R, D) (D, D, D, R) (D, D, D, D)

When half the population voted for each candidate, the 16 samples are equally likely.

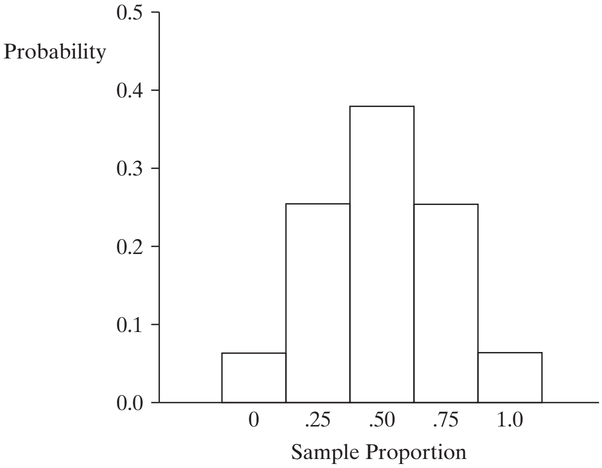
Let's construct the sampling distribution of the sample proportion that voted for the Republican candidate. For a sample of size 4, that proportion can be 0, 0.25, 0.50, 0.75, or 1.0. The proportion 0 occurs with only one of the 16 possible samples, (D, D, D, D), so its probability equals $1/16 = 0.0625$. The proportion 0.25 occurs for four samples, (R, D, D, D), (D, R, D, D), (D, D, R, D), and (D, D, D, R), so its probability equals $4/16 = 0.25$. Based on this reasoning, Table 4.3 shows the probability for each possible sample proportion value.

TABLE 4.3: Sampling Distribution of Sample Proportion Voting Republican, for Random Sample of Size $n = 4$ when Population Proportion Is 0.50. For example, a sample proportion of 1.0 occurs for only 1 of 16 possible samples, namely (R, R, R, R), so its probability is 1/16 = 0.0625.

Sample Proportion	Probability
0.00	0.0625
0.25	0.2500
0.50	0.3750
0.75	0.2500
1.00	0.0625

Figure 4.11 portrays the sampling distribution of the sample proportion for $n = 4$. It is much more spread out than the one in Figure 4.10 for samples of size $n = 1824$, which falls nearly entirely between 0.46 and 0.54. With such a small sample ($n = 4$), the sample proportion need not be near the population proportion. This is not surprising. In practice, samples are usually much larger than $n = 4$. We used a small value in this example, so it was simpler to write down all the potential samples and find probabilities for the sampling distribution.[7] ■

FIGURE 4.11: Sampling Distribution of Sample Proportion Voting Republican, for Random Sample of Size $n = 4$ when Population Proportion Is 0.50



Suppose we denoted the two possible outcomes by 0 for Democrat and by 1 for Republican. From Section 3.2 (page 40), *the proportion of times that 1 occurs is the sample mean of the data*. For instance, for the sample (0, 1, 0, 0) in which only the second subject voted for the Republican, the sample mean equals $(0 + 1 + 0 + 0)/4 = 1/4 = 0.25$, the sample proportion voting for the Republican. So, Figure 4.11 is also an example of a sampling distribution of a sample mean. Section 4.5 presents properties of the sampling distribution of a sample mean.

REPEATED SAMPLING INTERPRETATION OF SAMPLING DISTRIBUTIONS

Sampling distributions portray the sampling variability that occurs in collecting data and using sample statistics to estimate parameters. If different polling organizations each take their own exit poll and estimate the population proportion voting for the Republican candidate, they will get different estimates, because the samples have different people. Likewise, Figure 4.10 describes the variability in sample proportion values that occurs in selecting a huge number of samples of size $n = 1824$ and constructing a histogram of the sample proportions. By contrast, Figure 4.11 describes the variability for a huge number of samples of size $n = 4$.

A sampling distribution of a statistic for n observations is the relative frequency distribution for that statistic resulting from repeatedly taking samples of size n, each time calculating the statistic value. It's possible to form such a distribution empirically, as in Figure 4.10, by repeated sampling or through simulation. In practice, this is not necessary. The form of sampling distributions is often known theoretically, as shown in the previous example and in the next section. We can then find probabilities about the value of the sample statistic for one random sample of the given size n.

[7] Section 6.7 presents a formula for probabilities in this sampling distribution, called the *binomial distribution*, but we do not need the formula here.

4.5 Sampling Distributions of Sample Means

Because the sample mean $\bar{y}$ is used so much, with the sample proportion also being a sample mean, its sampling distribution merits special attention. In practice, when we analyze data and find $\bar{y}$, we do not know how close it falls to the population mean μ, because we do not know the value of μ. Using information about the spread of the sampling distribution, though, we can predict how close it falls. For example, the sampling distribution might tell us that with high probability, $\bar{y}$ falls within 10 units of μ.

This section presents two main results about the sampling distribution of the sample mean. One provides formulas for the center and spread of the sampling distribution. The other describes its shape.

MEAN AND STANDARD ERROR OF SAMPLING DISTRIBUTION OF $\bar{y}$

The sample mean $\bar{y}$ is a variable, because its value varies from sample to sample. For random samples, it fluctuates around the population mean μ, sometimes being smaller and sometimes being larger. In fact, the mean of the sampling distribution of $\bar{y}$ equals μ. If we repeatedly took samples, then in the long run, the mean of the sample means would equal the population mean μ.

The spread of the sampling distribution of $\bar{y}$ is described by its standard deviation, which is called the ***standard error*** of $\bar{y}$.

Standard Error

> The standard deviation of the sampling distribution of $\bar{y}$ is called the ***standard error*** of $\bar{y}$ and is denoted by $\sigma_{\bar{y}}$.

The standard error describes how much $\bar{y}$ varies from sample to sample. Suppose we repeatedly selected samples of size n from the population, finding $\bar{y}$ for each set of n observations. Then, in the long run, the standard deviation of the $\bar{y}$-values would equal the standard error. The symbol $\sigma_{\bar{y}}$ (instead of σ) and the terminology *standard error* (instead of *standard deviation*) distinguish this measure from the standard deviation σ of the population distribution.

In practice, we do not need to take samples repeatedly to find the standard error of $\bar{y}$, because a formula is available. For a random sample of size n, the standard error of $\bar{y}$ depends on n and the population standard deviation σ by[8]

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}.$$

Figure 4.12 displays a population distribution having $\sigma = 10$ and shows the sampling distribution of $\bar{y}$ for $n = 100$. When $n = 100$, the standard error is $\sigma_{\bar{y}} = \sigma/\sqrt{n} = 10/\sqrt{100} = 1.0$. The sampling distribution has only a tenth of the spread of the population distribution. This means that individual observations tend to vary much more than sample means vary from sample to sample.

In summary, the following result describes the center and spread of the sampling distribution of $\bar{y}$:

Mean and Standard Error of $\bar{y}$

> For sampling a population, the sampling distribution of $\bar{y}$ states the probabilities for the possible values of $\bar{y}$. For a random sample of size n from a population having mean μ and standard deviation σ, the sampling distribution of $\bar{y}$ has mean μ and standard error $\sigma_{\bar{y}} = \sigma/\sqrt{n}$.

[8] Exercise 4.58 shows the basis of this formula.

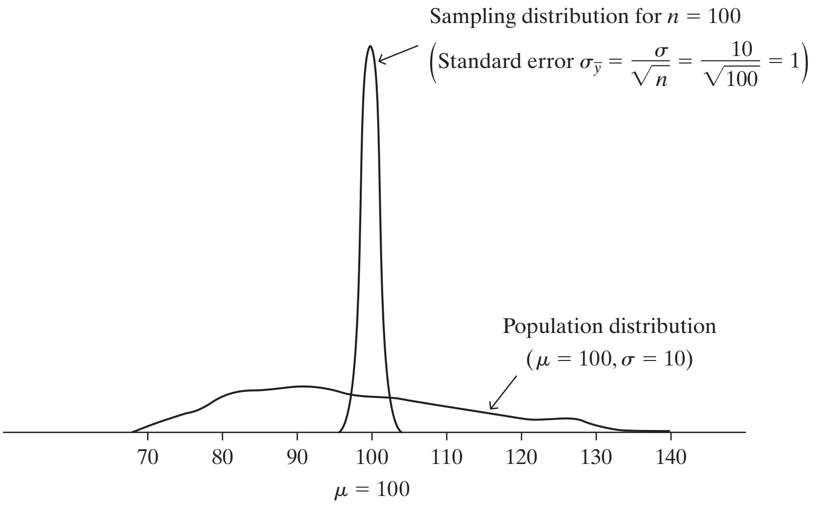


FIGURE 4.12: A Population Distribution and the Sampling Distribution of $\bar{y}$ for $n = 100$

Example 4.7 **Standard Error of Sample Proportion in Exit Poll** Following Example 4.5 (page 80), we conducted a simulation to investigate how much variability to expect from sample to sample in an exit poll of 1824 voters. Instead of conducting a simulation, we can get similar information directly by finding a standard error. Knowing the standard error helps us answer the following question: If half the population voted for each candidate, how much would a sample proportion for an exit poll of 1824 voters tend to vary from sample to sample?

Let the variable y denote a vote outcome. As at the end of Example 4.6, we let $y = 0$ represent a vote for the Democrat and $y = 1$ represent a vote for the Republican. Figure 4.13 shows the population distribution for which half the population voted for each, so that $P(0) = 0.50$ and $P(1) = 0.50$. The mean of the distribution equals 0.50, which is the population proportion voting for each. (Or, from the formula near the end of Section 4.2, $\mu = \sum yP(y) = 0(0.50) + 1(0.50) = 0.50$.) The squared deviation of y from the mean, $(y - \mu)^2$, equals $(0 - 0.50)^2 = 0.25$ when $y = 0$, and it equals $(1 - 0.50)^2 = 0.25$ when $y = 1$. The variance is the expected value of this squared deviation. Thus, it equals $\sigma^2 = 0.25$. So, the standard deviation of the population distribution of y is $\sigma = \sqrt{0.25} = 0.50$.

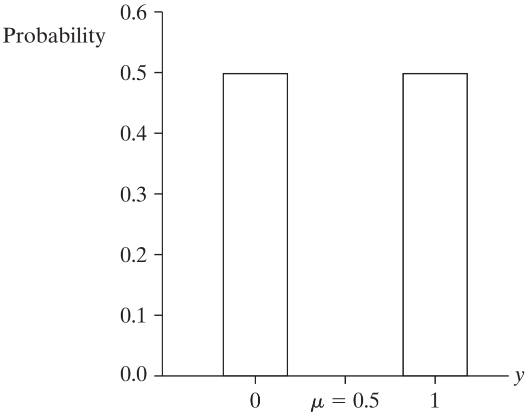


FIGURE 4.13: The Population Distribution when $y = 0$ or 1, with Probability 0.50 Each. This is the probability distribution for a vote, with 0 = vote for Democratic candidate and 1 = vote for Republican candidate.

For a sample, the mean of the 0 and 1 values is the sample proportion of votes for the Republican. Its sampling distribution has mean that is the mean of the population distribution of y, namely, $\mu = 0.50$. For repeated samples of a fixed size n, the sample proportions fluctuate around 0.50, being larger about half the time and smaller half the time. The standard deviation of the sampling distribution is the standard error. For a sample of size 1824, this is

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{0.50}{\sqrt{1824}} = 0.0117.$$

A result from later in this section says that this sampling distribution is bell shaped. Thus, with probability close to 1.0 the sample proportion falls within three standard errors of μ, that is, within $3(0.0117) = 0.035$ of 0.50, or between about 0.46 and 0.54. For a random sample of size 1824 from a population in which 50% voted for each candidate, it would be surprising if fewer than 46% or more than 54% voted for one of them. We've now seen how to get this result either using simulation, as shown in Figure 4.10, or using the information about the mean and standard error of the sampling distribution. ■

EFFECT OF SAMPLE SIZE ON SAMPLING DISTRIBUTION AND PRECISION OF ESTIMATES

The standard error gets smaller as the sample size n gets larger. The reason for this is that the denominator ($\sqrt{n}$) of the standard error formula $\sigma_{\bar{y}} = \sigma/\sqrt{n}$ increases as n increases. For instance, when the population standard deviation $\sigma = 0.50$, we've just seen that the standard error is 0.0117 when $n = 1824$. When $n = 100$, a less typical size for a poll, the standard error is

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{0.50}{\sqrt{n}} = \frac{0.50}{\sqrt{100}} = 0.050.$$

With $n = 100$, since three standard errors equal $3(0.050) = 0.15$, the probability is very high that the sample proportion falls within 0.15 of 0.50, or between 0.35 and 0.65.

Figure 4.14 shows the sampling distributions of the sample proportion when $n = 100$ and when $n = 1824$. As n increases, the standard error decreases and the sampling distribution gets narrower. This means that the sample proportion tends to fall closer to the population proportion. It's more likely that the sample proportion closely approximates a population proportion when $n = 1824$ than when $n = 100$. This agrees with our intuition that larger samples provide more precise estimates of population characteristics.

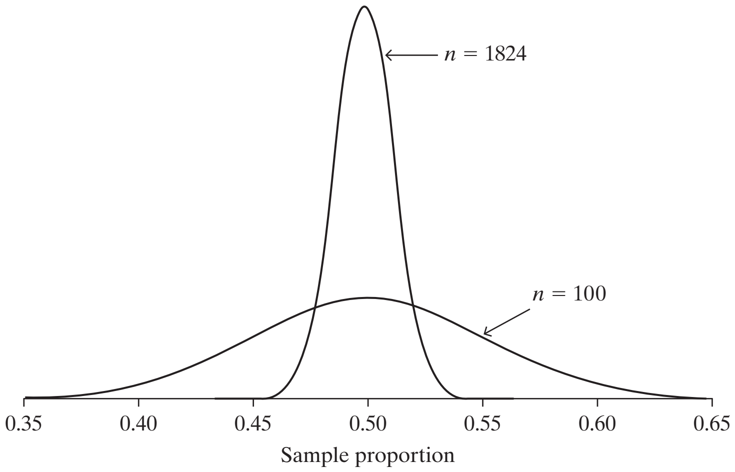


FIGURE 4.14: The Sampling Distributions of the Sample Proportion, when $n = 100$ and when $n = 1824$. These refer to sampling from the population distribution in Figure 4.13.

In summary, error occurs when we estimate μ by $\bar{y}$, because we sampled only part of the population. This error, which is the ***sampling error***, tends to decrease as the sample size n increases. The standard error is fundamental to inferential procedures that predict the sampling error in using $\bar{y}$ to estimate μ.

SAMPLING DISTRIBUTION OF SAMPLE MEAN IS APPROXIMATELY NORMAL

For the population distribution for the vote in an election, shown in Figure 4.13, the outcome has only two possible values. It is highly discrete. Nevertheless, the two sampling distributions shown in Figure 4.14 have bell shapes. This is a consequence of the second main result of this section, which describes the *shape* of the sampling distribution of $\bar{y}$. This result can be proven mathematically, and it is often called the *Central Limit Theorem*.

Central Limit Theorem

For random sampling with a large sample size n, the sampling distribution of the sample mean $\bar{y}$ is approximately a normal distribution.

Here are some implications and interpretations of this result:

- The bell shape of the sampling distribution applies *no matter what the shape* of the population distribution. This is remarkable. For large random samples, the sampling distribution of $\bar{y}$ has a normal bell shape even if the population distribution is very skewed or highly discrete such as the binary distribution in Figure 4.13. We'll learn how this enables us to make inferences even when the population distribution is highly irregular. This is helpful, because many social science variables are very skewed or highly discrete.
  Figure 4.15 displays sampling distributions of $\bar{y}$ for four different shapes for the population distribution, shown at the top of the figure. Below them are

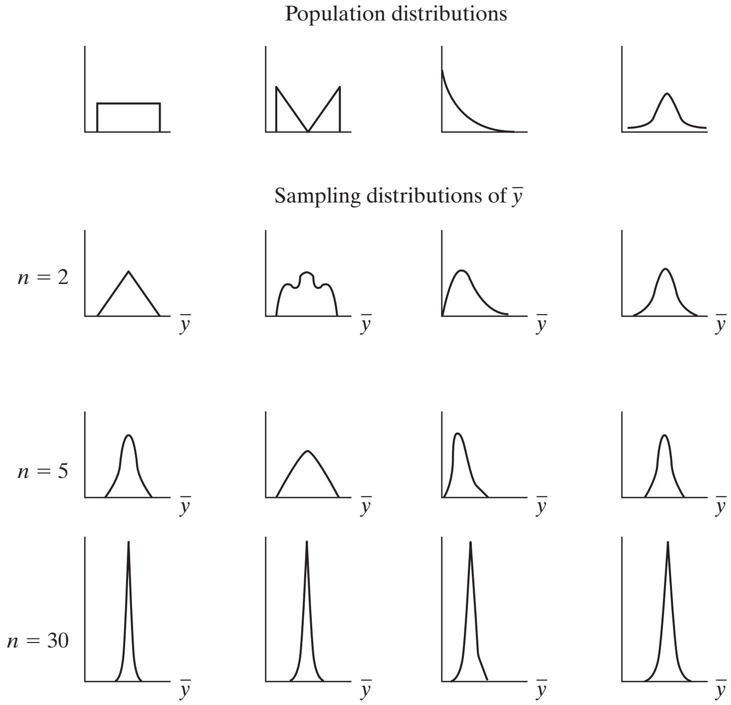


FIGURE 4.15: Four Different Population Distributions and the Corresponding Sampling Distributions of $\bar{y}$. As n increases, the sampling distributions get narrower and have more of a bell shape.

portrayed the sampling distributions for random samples of sizes $n = 2$, 5, and 30. As n increases, the sampling distribution has more of a bell shape.

- How large n must be before the sampling distribution is bell shaped largely depends on the skewness of the population distribution. If the *population* distribution is bell shaped, then the sampling distribution is bell shaped for *all* sample sizes. The rightmost panel of Figure 4.15 illustrates this. More skewed distributions require larger sample sizes. For most cases, n of about 30 is sufficient (although it may not be large enough for precise inference). So, in practice, with random sampling the sampling distribution of $\bar{y}$ is nearly always approximately bell shaped.
- Knowing that the sampling distribution of $\bar{y}$ can be approximated by a normal distribution helps us to find probabilities for possible values of $\bar{y}$. For instance, $\bar{y}$ almost certainly falls within $3\sigma_{\bar{y}} = 3\sigma/\sqrt{n}$ of μ. Reasoning of this nature is vital to inferential statistical methods.

Example 4.8

Simulating a Sampling Distribution You can verify the Central Limit Theorem empirically by repeatedly selecting random samples, calculating $\bar{y}$ for each sample of n observations. Then, the histogram of the $\bar{y}$-values is approximately a normal curve.

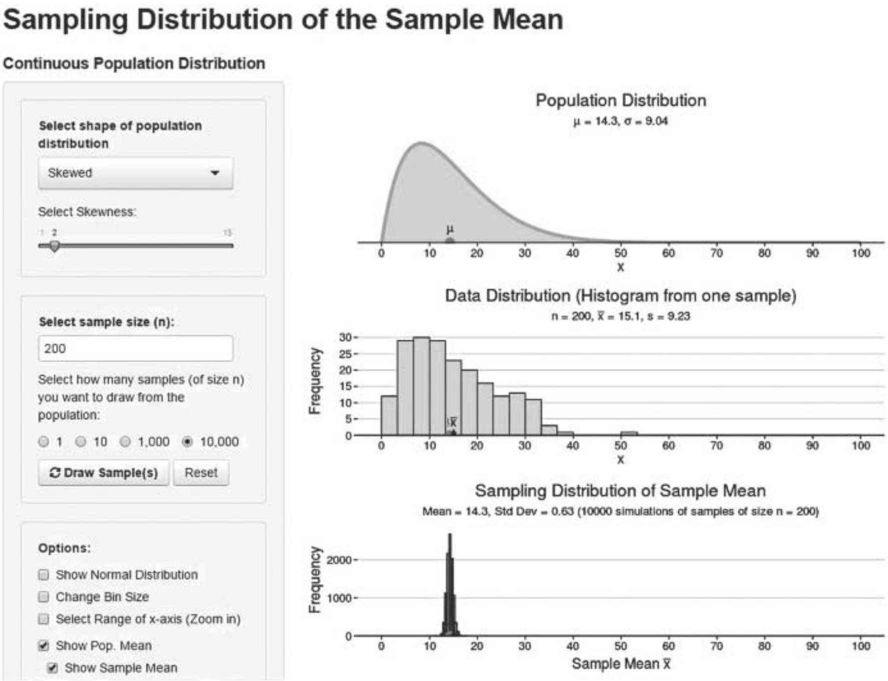
- Go to `www.artofstat.com/webapps.html` and click on the *Sampling Distribution for the Sample Mean* applet for continuous variables.
- Select a skewed population distribution. You can specify how skewed the distribution is. Here, we'll use the value 2 for skewness.
- We'll consider what happens for sample sizes of $n = 200$, relatively modest for a social science study. Enter 200 in the *Select sample size* box. When you click on *Draw sample*, the applet will randomly sample 200 observations, find the sample mean and standard deviation, and plot a histogram.
- Next, change the number of samples of size $n = 200$ that you draw from 1 to 10,000. When you again click on *Draw Sample*, the applet will select 10,000 samples, each of size 200. It will find the sample mean for each sample of 200 observations, overall then finding 10,000 sample means and plotting their histogram. See Figure 4.16 (page 90) for a result. It shows the skewed population distribution on top, the sample data distribution for one of the samples of size 200 below that, and the empirical sampling distribution for the 10,000 sample means at the bottom.

Even though the population distribution in Figure 4.16 is skewed, the sampling distribution is bell shaped. It is also much less spread out, because its spread is described by the standard error, which is the population standard deviation divided by $\sqrt{200} = 14.1$. ■

Example 4.9

Is Sample Mean Income of Migrant Workers Close to Population Mean? For the population of migrant workers doing agricultural labor in Florida, suppose that weekly income has a distribution that is skewed to the right with a mean of $\mu = \$380$ and a standard deviation of $\sigma = \$80$. A researcher, unaware of these values, plans to randomly sample 100 migrant workers and use the sample mean income $\bar{y}$ to estimate μ. What is the sampling distribution of the sample mean? Where is $\bar{y}$ likely to fall, relative to μ? What is the probability that $\bar{y}$ overestimates μ by more than \$20, falling above \$400?

By the Central Limit Theorem, the sampling distribution of the sample mean $\bar{y}$ is approximately normal, even though the population distribution is skewed. The



Sampling Distribution of the Sample Mean

Continuous Population Distribution

Select shape of population distribution: Skewed

Select Skewness: 2

Select sample size (n): 200

Select how many samples (of size n) you want to draw from the population:

○ 1 ○ 10 ○ 1,000 ● 10,000

Draw Sample(s) Reset

Options:

□ Show Normal Distribution
□ Change Bin Size
□ Select Range of x-axis (Zoom in)
☑ Show Pop. Mean
☑ Show Sample Mean

FIGURE 4.16: An Applet for Simulating a Sampling Distribution. Here, in clicking on *Draw Sample*, we take 10,000 samples of size 200 each. The graphic shows the population distribution, the sample data distribution for one sample of size 200, and the empirical sampling distribution that shows the 10,000 values of $\bar{y}$ for the 10,000 samples of size $n = 200$ each.

sampling distribution has the same mean as the population distribution, namely, $\mu =$ \$380. Its standard error is

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{80}{\sqrt{100}} = 8.0 \text{ dollars}.$$

Thus, it is highly likely that $\bar{y}$ falls within about \$24 (three standard errors) of μ, that is, between about \$356 and \$404.

For the normal sampling distribution with mean 380 and standard error 8, the possible $\bar{y}$ value of 400 has a z-score of

$$z = (400 - 380)/8 = 2.5.$$

From a table of normal probabilities (such as Table A) or software, the corresponding right-tail probability above 400 is 0.0062. It is very unlikely that the sample mean would fall above \$400. ■

This example is unrealistic, because we assumed knowledge of the population mean μ. In practice, μ would be unknown. However, the sampling distribution of $\bar{y}$ provides the probability that the sample mean falls within a certain distance of the population mean μ, even when μ is unknown. We illustrate by finding the probability that the sample mean weekly income $\bar{y}$ falls within \$10 of the unknown true mean income μ for all such workers.

Now, the sampling distribution of $\bar{y}$ is approximately normal in shape and is centered about μ. We have just seen that when $n = 100$, the standard error is $\sigma_{\bar{y}} = \$8.0$. Hence, the probability that $\bar{y}$ falls within \$10 of μ is the probability that a normally distributed variable falls within $10/8 = 1.25$ standard deviations of its mean. That is, the number of standard errors that $\mu + 10$ (or $\mu - 10$) falls from μ is

$$z = \frac{(\mu + 10) - \mu}{8} = \frac{10}{8} = 1.25.$$

See Figure 4.17. From a normal table, the probability that $\bar{y}$ falls *more than* 1.25 standard errors from μ (in either direction) is $2(0.1056) = 0.21$. Thus, the probability that $\bar{y}$ falls no more than \$10 from μ equals $1 - 0.21 = 0.79$.

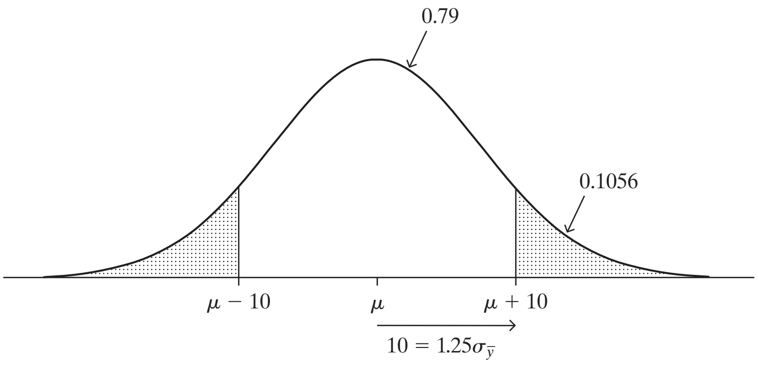


FIGURE 4.17: Sampling Distribution of $\bar{y}$ for Unknown μ and Standard Error $\sigma_{\bar{y}} = 8$

This example is still unrealistic, because we assumed knowledge of the population standard deviation σ. In practice, we'd need to estimate σ. The next chapter shows that to conduct inference, we estimate σ by the sample standard deviation s.

To get a feel for the Central Limit Theorem and how sampling distributions become more bell shaped as n increases, we suggest that you try out an applet on the Internet, as in Exercises 4.38 and 4.39.

4.6 Review: Population, Sample Data, and Sampling Distributions

Sampling distributions are fundamental to statistical inference and to methodology presented in the rest of this text. Because of this, we now review the distinction between sampling distributions and the two types of distributions presented in Section 3.1—the ***population*** distribution and the ***sample data*** distribution.

Here is a capsule description of the three types of distribution:

- ***Population distribution***: This is the distribution from which we select the sample. It is usually unknown. We make inferences about its characteristics, such as the parameters μ and σ that describe its center and spread.

- ***Sample data distribution***: This is the distribution of data that we actually observe, that is, the sample observations $y_1, y_2, \ldots, y_n$. We describe it by statistics such as the sample mean $\bar{y}$ and sample standard deviation s. The larger the sample size n, the closer the sample data distribution resembles the population distribution, and the closer the sample statistics such as $\bar{y}$ fall to the population parameters such as μ.

- ***Sampling distribution*** of a statistic: This is the probability distribution for the possible values of a sample statistic, such as $\bar{y}$. A sampling distribution describes the variability that occurs in the statistic's value among samples of a certain size. This distribution determines the probability that the statistic falls within a certain distance of the population parameter it estimates.

In Figure 4.16 on page 90, the *population distribution* is the skewed distribution shown at the top. The distribution in the middle of the figure is a *sample data distribution* based on one particular sample of $n = 200$ observations. It has similar appearance to the population distribution, also being somewhat skewed to the right. It has $\bar{y} = 13.4$ and $s = 8.9$, similar to $\mu = 14.3$ and $\sigma = 9.0$ for the population. The distribution at the bottom of the figure describes the *sampling distribution* of the sample mean for random samples of size 200. It is an empirical sampling distribution, showing a histogram of 10,000 values of the sample mean for 10,000 random samples of size $n = 200$ each. It is bell shaped, as a consequence of the Central Limit Theorem, and very narrow, as a consequence of the standard error formula $\sigma_{\bar{y}} = \sigma/\sqrt{n}$. Following is an example in which the three distributions would have shape like those in Figure 4.16.

Example 4.10

Three Distributions for a General Social Survey Item In 2014, the GSS asked about the number of hours a week spent on the Internet, excluding e-mail. The *sample data distribution* for the $n = 1399$ subjects in the sample was very highly skewed to the right. It is described by the sample mean $\bar{y} = 11.6$ and sample standard deviation $s = 15.0$.

Because the GSS cannot sample the entire population of adult Americans (about 250 million people), we don't know the *population distribution*. Because the sample data distribution had a large sample size, probably the population distribution looks like it. Most likely the population distribution would also be highly skewed to the right. Its mean and standard deviation would be similar to the sample values. Values such as $\mu = 12.0$ and $\sigma = 14.0$ would be plausible.

If the GSS repeatedly took random samples[9] of 1399 adult Americans, the sample mean time $\bar{y}$ spent on the Internet would vary from survey to survey. The *sampling distribution* describes how $\bar{y}$ would vary. For example, if the population has mean $\mu = 12.0$ and standard deviation $\sigma = 14.0$, then the sampling distribution of $\bar{y}$ also has mean 12.0, and it has a standard error of

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{14.0}{\sqrt{1399}} = 0.37.$$

Unlike the population and sample data distributions, the sampling distribution would be bell shaped and narrow. Nearly all of that distribution would fall within $3(0.37) = 1.12$ of the mean of 12.0. So, it's very likely that any sample of size 1399 would have a sample mean within 1.12 of 12.0. In summary, the sample data and population distributions are highly skewed and spread out, whereas the sampling distribution of $\bar{y}$ is bell shaped and has nearly all its probability in a narrow range. ■

Example 4.11

Three Distributions for Exit Poll Example We consider, once again, the variable $y =$ vote in the 2014 California gubernatorial election for a randomly selected voter. Let $y = 1$ for Jerry Brown (the Democrat) and $y = 0$ for Neel Kashkari (the Republican). In fact, of the 7,317,581 adult residents of California who voted, 60.0% voted for Brown. So, the probability distribution for y has probability 0.600 at $y = 1$ and probability 0.400 at $y = 0$. The mean of this distribution is $\mu = 0.600$, which is the population proportion of votes for Brown. From a formula we'll study in the next chapter, the standard deviation of this two-point distribution is $\sigma = 0.490$.

The population distribution of candidate preference consists of the 7,317,581 values for y, of which 40.0% are 0 and 60.0% are 1. This distribution is described by the parameters $\mu = 0.600$ and $\sigma = 0.490$. Figure 4.18 portrays this distribution, which is highly discrete (binary). It is not at all bell shaped.

[9] In reality, the GSS uses a multistage cluster sample, so the true standard error is a bit larger than $\sigma/\sqrt{n}$. For purposes of illustration, we'll treat GSS data as if they come from a simple random sample, keeping in mind that in practice some adjustment is necessary as explained at the GSS website, sda.berkeley.edu/GSS.

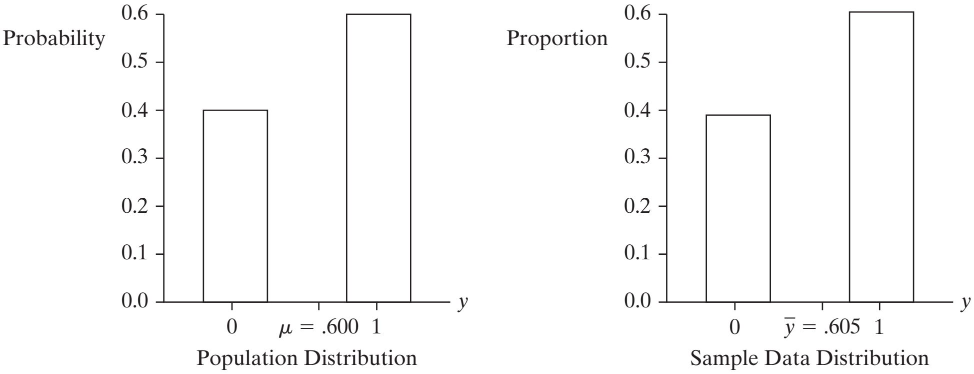


FIGURE 4.18: The Population Distribution (7,317,581 votes) and the Sample Data Distribution ($n = 1824$ votes) in the 2014 California Gubernatorial Election, where $1 =$ Vote for Brown and $0 =$ Vote for Kashkari

Before all the votes were counted, the population distribution was unknown. When polls closed, CBS News reported results of an exit poll of size $n = 1824$ to predict the outcome. A histogram of the 1824 votes in the sample describes the sample data distribution. Of the 1824 voters, 60.5% said they voted for Brown (i.e., have $y = 1$) and 39.5% said they voted for Kashkari ($y = 0$). Figure 4.18 also displays the histogram of these sample data values. Like the population distribution, the sample data distribution concentrates at $y = 0$ and $y = 1$. It is described by sample statistics such as $\bar{y} = 0.605$, which is the sample proportion voting for Brown. The larger the sample size, the more this sample data distribution tends to resemble the population distribution, since the sample observations are a random subset of the population values. If the entire population is sampled, as when all the votes are counted, then the two distributions are identical.

For a random sample of size $n = 1824$, the sampling distribution of $\bar{y}$ is approximately a normal distribution. Its mean is $\mu = 0.600$, and its standard error is

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{0.490}{\sqrt{1824}} = 0.011.$$

Figure 4.19 portrays this sampling distribution, relative to the population distribution of votes.

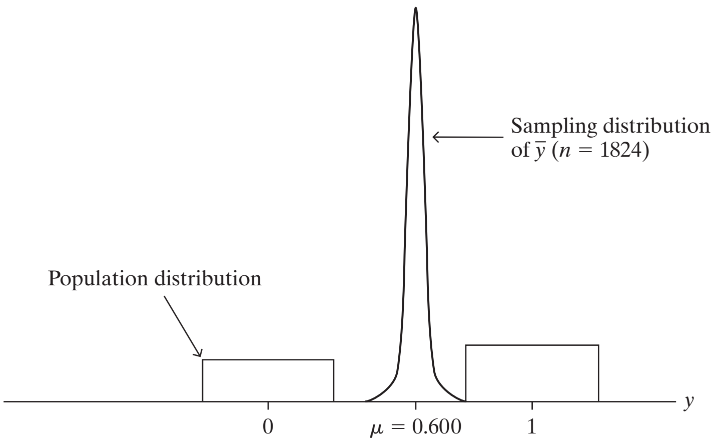


FIGURE 4.19: The Population Distribution (where $y = 1$ is Vote for Brown and $y = 0$ is Vote for Kashkari) and the Sampling Distribution of $\bar{y}$ for $n = 1824$

The population distribution and sample data distribution of votes are concentrated at the values 0 and 1. The sampling distribution looks completely different, being much less spread out and bell shaped. The population and sample data distributions of the vote are not bell shaped. They are highly discrete, concentrated at 0 and 1. For $n = 1824$, the sample proportion can take a large number of values between 0 and 1, and its sampling distribution is essentially continuous, being bell shaped by the Central Limit Theorem. Although the individual values of y are 0 and 1, according to the sampling distribution it is practically impossible that a random sample of size 1824 has a sample mean anywhere near 0 or 1; nearly all the probability falls between 0.57 and 0.63, that is, within three standard errors of the mean $\mu = 0.600$. ■

THE KEY ROLE OF SAMPLING DISTRIBUTIONS IN STATISTICAL INFERENCE

By the Central Limit Theorem, we can often use the normal distribution to find probabilities about $\bar{y}$. The next two chapters show how statistical inferences rely on this theorem.

The result about sample means having approximately normal sampling distributions is important also because similar results hold for many other statistics. For instance, most sample statistics used to estimate population parameters have approximately normal sampling distributions, for large random samples. This is the primary reason for the key role of the normal distribution in statistical science.

4.7 Chapter Summary

For an observation in a random sample or a randomized experiment, the ***probability*** of a particular outcome is the proportion of times that the outcome would occur in a very long sequence of observations.

- A ***probability distribution*** specifies probabilities for the possible values of a variable. We let $P(y)$ denote the probability of the value y. The probabilities are nonnegative and sum to 1.0.
- Probability distributions have summary parameters, such as the mean μ and standard deviation σ. The mean for a probability distribution of a discrete variable is
$$\mu = \sum yP(y).$$
This is also called the ***expected value*** of y.
- The ***normal distribution*** has a graph that is a symmetric bell-shaped curve specified by the mean μ and standard deviation σ. For any z, the probability falling within z standard deviations of the mean is the same for every normal distribution.
- In a probability distribution, the ***z-score*** for a value y is
$$z = (y - \mu)/\sigma.$$
It measures the number of standard deviations that y falls from the mean μ. For a normal distribution, the z-scores have the ***standard normal distribution***, which has mean = 0 and standard deviation = 1.
- A ***sampling distribution*** is a probability distribution of a sample statistic, such as the sample mean or sample proportion. It specifies probabilities for the possible values of the statistic for samples of the particular size n.
- The sampling distribution of the sample mean $\bar{y}$ centers at the population mean μ. Its standard deviation, called the ***standard error***, relates to the standard deviation σ of the population by $\sigma_{\bar{y}} = \sigma/\sqrt{n}$. As the sample size n increases,

the standard error decreases, so the sample mean tends to be closer to the population mean.

- The ***Central Limit Theorem*** states that for large random samples on a variable, the sampling distribution of the sample mean is approximately a normal distribution. This holds no matter what the shape of the population distribution, both for continuous variables and for discrete variables. The result applies also to proportions, since the sample proportion is a special case of the sample mean for observations coded as 0 and 1 (such as for two candidates in an election).

The bell shape for the sampling distribution of many statistics is the main reason for the importance of the normal distribution. The next two chapters show how the Central Limit Theorem is the basis of methods of statistical inference.

Exercises

Practicing the Basics

4.1. In a General Social Survey, in response to the question "Do you believe in heaven?" 1127 people answered "yes" and 199 answered "no."

(a) Estimate the probability that a randomly selected adult in the United States believes in heaven.

(b) Estimate the probability that an American adult does not believe in heaven.

(c) Of those who believe in heaven, about 84% believe in hell. Estimate the probability that a randomly chosen American adult believes in both heaven and hell.

4.2. Software for statistical inference methods often sets the default probability of a correct inference to be 0.95. Suppose we make an inference about the population proportion of people who support legalization of marijuana, and we consider this separately for men and for women. Let A denote the outcome that the inference about men is correct, and let B denote the outcome that the inference about women is correct. Treating these as independent samples and inferences, find the probability that *both* inferences are correct.

4.3. A recent GSS asked subjects whether they are a member of an environmental group and whether they would be very willing to pay much higher prices to protect the environment. Table 4.4 shows results.

(a) Estimate the probability that a randomly selected American adult is a member of an environmental group.

(b) Show that the estimated probability of being very willing to pay much higher prices to protect the environment is (i) 0.312, given that the person is a member of an environmental group, (ii) 0.086, given that the person is not a member of an environmental group.

(c) Show that the estimated probability that a person is both a member of an environmental group *and* very willing to pay much higher prices to protect the environment is 0.027 (i) directly using the counts in the table, (ii) using the probability estimates from (a) and (b).

(d) Show that the estimated probability that a person answers yes to both questions or no to both questions is 0.862.

TABLE 4.4

		Pay Higher Prices		
		Yes	No	Total
Member of Environmental Group	Yes	30	66	96
	No	88	933	1021
Total		118	999	1117

4.4. Let y = number of languages in which a person is fluent. According to Statistics Canada, for residents of Canada y has probability distribution $P(0) = 0.02$, $P(1) = 0.81$, and $P(2) = 0.17$, with negligible probability for higher values of y.

(a) Is y a discrete, or a continuous, variable? Why?

(b) Construct a table showing the probability distribution of y.

(c) Find the probability that a Canadian is *not* multilingual.

(d) Find the mean of this probability distribution.

4.5. Let y denote the number of people known personally who were victims of homicide within the past 12 months. According to results from recent General Social Surveys, for a randomly chosen person in the United States the probability distribution of y is approximately $P(0) = 0.91$, $P(1) = 0.06$, $P(2) = 0.02$, and $P(3) = 0.01$.

(a) Explain why it is not valid to find the mean of this probability distribution as $(0 + 1 + 2 + 3)/4 = 1.5$.

(b) Find the correct mean of the probability distribution.

4.6. A ticket for a statewide lottery costs $1. With probability 0.0000001, you win a million dollars ($1,000,000), and with probability 0.9999999 you win nothing. Let y