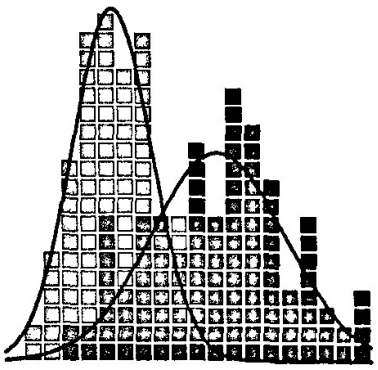
2020

Random Sampling

Lawrence Leemis

Chapter 1

Random Sampling



A *statistic* is a number that is calculated from a sample consisting of data values. One simple example of a statistic is the sample mean. The study of statistics concerns gathering data, analyzing data, drawing conclusions from data, and making decisions from data. Statisticians use mathematical techniques and algorithms to collect, display, and analyze data. Many of their methods rely on probability, so the study of statistics can be thought of as probability applied to data.

Sometimes, you just need to present data in a way that allows you to reach an obvious conclusion. In other cases, sophisticated mathematical models are required in order to draw a conclusion from a data set. We begin with the first case.

1.1 Statistical Graphics

Assuming that data values have been collected in a reasonable fashion (more detail will be provided later about what constitutes "reasonable"), statisticians face the challenge of how to present the data values in a fair, intuitive, and revealing fashion. The graphical display of a data set consists of tables and figures that highlight key features that address a relevant question of interest.

We begin with a simple example of placing a data set of just five observations in a table in order to compare the populations of five countries.

Example 1.1 A data set consisting of the populations of five countries in the year 2000 taken from the appropriate Wikipedia website is displayed in Table 1.1. Although it contains all of the information required to compare the populations, Table 1.1 is a dreadful presentation of this data set for the following reasons.

- The populations are placed horizontally. Although this takes up less vertical space on the page, it makes visual comparisons more difficult. Aligning the populations vertically on the decimal point is a better approach.

Country:	China	Indonesia	Montenegro	Serbia	U.S.A.
Pop.:	1,242,612,266	206,264,595	620,145	9,778,991	281,421,906

Table 1.1: Populations in 2000 (presented poorly).

- The countries are sorted alphabetically even though the interest is in comparing population sizes; it would have been more helpful to sort them by decreasing population size.
- The bland monospace font is hypnotizing for the reader and obscures rather than accentuates the population sizes.
- There are lots of lines separating the fields in the table; it is better to use fewer lines.
- Unnecessary extra ink (such as the colons after the row labels) should be avoided. Abbreviating the population label is not helpful.
- Using all of the digits on population is distracting and makes comparing the populations more difficult. Does the value in the ones digit really matter?

So how can this table be improved? We can improve Table 1.1 by addressing each of the six points above. Table 1.2 is a second attempt at displaying the five data values which, hopefully, you find to be a more intuitive way of presenting the data with the goal of comparing populations. One can easily and immediately see that China has the largest population with over one billion people, followed by the others in decreasing population size. The data values are arranged vertically, aligned on the decimal point, and sorted by population size; fonts have been altered, unnecessary ink has been removed, and distracting digits have been removed by presenting the populations in millions. The essence of the population data set is easily and quickly gleaned from Table 1.2 by using common-sense principles to redesign the table.

Country	Population (millions)
China	1,242.6
U.S.A.	281.4
Indonesia	206.3
Serbia	9.8
Montenegro	0.6

Table 1.2: Populations in 2000.

The display of data in an accurate, meaningful, and intuitive fashion is as much an art as it is a science. What constitutes a "good" table is a matter of preference. The two tables from Example 1.1 have illustrated a bad presentation and a good presentation. The key step in this process is placing yourself in the reader's position and thinking of ways to simplify the presentation of the data to illuminate the aspect of the data that is of interest.

Presenting data in tables, however, is not the only option. The graphical presentation of data provides a much more efficient way to convey the message a data set provides.

The practice of presenting data in graphical form is a field known as *statistical graphics*. One of the early pioneers in the field was William Playfair (1759–1823), who published the *Commercial and Political Atlas* in 1786. His atlas contained graphs that described imports and exports of England and Wales with their trading partners. Before the publication of Playfair's atlas, graphs were rare; after his publication, graphs began to appear with increasing regularity. More recent leaders in the field include William S. Cleveland, Edward R. Tufte, and John W. Tukey. Their books are listed in the preface and are recommended reading materials if this brief overview of statistical graphics given here sparks some interest on your part.

Contemplating how one should construct a graph in order to display quantitative information will help you think about data, which is at the core of statistical practice. Visualizing the data set via a graph is often the first view we get of a data set after the data is collected. It provides a first impression that often guides the next step, which might be using a statistical inference technique to draw a conclusion from the data. The graphs that will be produced in this section are drawn in the free statistical package called R, but there are many other tools available for constructing statistical graphics.

Do not prioritize fancy fonts, color, shading, or highlighting when it comes to statistical graphics. The best graphic is often a simple plot which conveys only the appropriate information in the data set. One overriding principle established by Tufte is to maximize the *data-ink ratio.* Make every bit of ink that you place on a graphic count. Use as little ink as possible to provide as much explanation as is necessary. Do not decorate your graphics with what has come to be known as *chartjunk*, which conveys no information, but is simply a misguided attempt to make the graphic more attractive. Display the data succinctly and clearly, bringing the key information contained in the data set to prominence in your graphic.

There are dozens of decisions that must be made when constructing a graph to display data. Human vision and perception considerations should drive all decisions. A sampling of the related questions includes the following.

- *Aspect ratio.* Is there a reason (by virtue of the meaning of the scales) that the plotting area should be a square? If not, is a tall thin graph (portrait orientation) more appropriate, or is a short wide graph (landscape orientation) more appropriate?

- *Axes.* Should axes be included? Does one of the variables naturally belong on the horizontal axis? Should axes be included on just the bottom and left sides of the plot, or should they be included on three or all four sides of the plot? Should the axes intersect or should there be a gap between them?

- *Scales.* Should the scales on the axes be linear? Should a logarithmic scale for an axis be used? Should a square root scale for an axis be used? Where should the scales begin and end? Is it helpful to have zero included on the scale? (This is not an option for a logarithmic scale.) Should a break be placed on a scale in order to include zero?

- *Axis labels.* Should the axis labels be in the same font style as the manuscript text? Should the axis labels be the same font size as the manuscript text? Should the labels be placed at the ends of the axes or in the center of the axes? Should the vertical axis label be displayed parallel to the axis or rotated clockwise 45° or 90° for easier reading?

- *Tick marks and tick labels.* Should the tick marks extend into the plotting area or out of the plotting area? How many tick marks should be included on each axis? Should all tick marks be the same size? How long should the tick marks be? Should all tick marks be labeled? Should the tick mark labels be in the same font style as the manuscript text? Should the tick mark labels be the same font size as the manuscript text? Should the tick mark labels for the vertical axis be displayed parallel to the axis or rotated clockwise 90° for easier reading?

- *Plotting area.* What symbol is appropriate for plotting a point? Is a legend necessary to describe the meaning of the plotted symbols? What should be done if two points fall on top of one another? Should points be connected with lines? Is placing text in the plotting area helpful? Should reference lines be included in the plotting area? If so, should they be solid, dotted, or dashed? Should the elements placed in the plotting area be black, gray, or colored?

Every graph that you construct must answer these questions. The first example of a statistical graphic comes from sports.

Example 1.2 *March Madness* occurs every spring when a 64-team single-elimination tournament is used to determine the best college basketball team in the United States. One game, which some consider to be the greatest basketball game ever played, occurred in the East Regional final game of the men's tournament on March 28, 1992 between the Blue Devils of Duke and the Wildcats of Kentucky. Duke prevailed 104–103 in an overtime victory, winning on a last-second shot by Christian Laettner. The game featured perfect shooting for Laettner (10 for 10 from the field and 10 for 10 free throws for 31 points), five lead changes in the last 31.5 seconds, and both teams shooting 63% from the field in the second half and overtime. ESPN sportswriter Gene Wojciechowski wrote a book about this game titled *The Last Great Game*. Table 1.3 contains the box score for those who played in the game.

Duke		**Kentucky**	
Christian Laettner	31	Jamal Mashburn	28
Bobby Hurley	22	Sean Woods	21
Thomas Hill	19	Dale Brown	18
Brian Davis	13	John Pelphrey	16
Grant Hill	11	Richie Farmer	9
Antonio Lang	4	Gimel Martinez	5
Cherokee Parks	4	Deron Feldhaus	5
Marty Clark	0	Aminu Timberlake	1
		Nehemiah Braddy	0
		Travis Ford	0
		Andre Riddick	0
Total	104	Total	103

Table 1.3: Duke vs. Kentucky box score.

The box score is helpful for knowing the final score and how points were distributed among the players, but it does not give you any indication of the dynamic, or time-dependent, aspect of the game. Was the game a see-saw battle with many lead changes, or did one team build a big lead and the other battled back? Did one team score a large portion of their points with three-point shots? The box score does not answer these questions; so, we must design a statistical graphic that includes the answers.

The first question concerns the axes on the graphic. One plot that might be of interest is scoring over time. We follow the standard practice of placing time on the horizontal axis. Now we move to the vertical axis. One measure of scoring which allows the viewer to easily capture the flow of the game is to plot the difference between the scores. The step function moves upward for a Duke score and downward for a Kentucky score; so whenever the difference lies above zero, Duke is leading. The fact that the basketball hoops are swapped at halftime is not particularly relevant.

Four more elements are added to the statistical graphic to help the viewer. First, a bit of shading helps highlight the area between the function and the horizontal line associated with a tied game, which helps the reader see how long a lead is held. Second, vertical

lines are added at $t = 20$ minutes (halftime) and $t = 40$ minutes (end of regulation). Third, small dots are added at each scoring time. Finally, vertical axes on both the left and right sides of the graphic make the differences in scores easier to determine.

After some experimentation, the landscape orientation of the plotting area looked the best; the final graphic is shown in Figure 1.1. Several conclusions can be quickly drawn from this statistical graphic that are not apparent from the box score.

- Kentucky scored first with a 3-point shot.
- Kentucky built a lead of 8 points early in the first half.
- Duke lead by 5 points at halftime.
- The two longest scoring droughts occurred early in the second half.
- Duke built their lead to 12 midway through the second half.
- Kentucky brought the game back to within 1 point with a 9 point scoring run.
- There was a scoring frenzy at the very end of the game, which is often the case in a close game because of deliberate fouling, which stops the clock.
- Duke won the game on a buzzer shot.

The small dots on the plot allow the viewer to see the difference between two consecutive successful free throws (such as those by Duke midway through the overtime period) and a two point shot (such as that by Kentucky just prior to the successful free throws). The statistical graphic can be enhanced by labeling the times of key events (such as a key player fouling out) that might influence the momentum of the game.

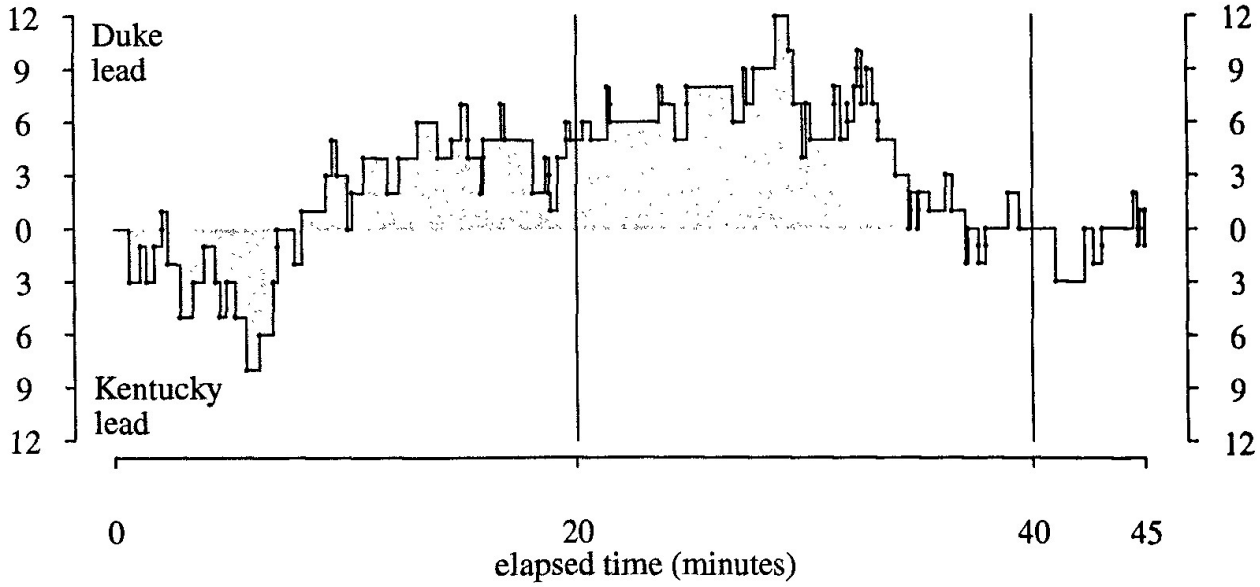


Figure 1.1: Duke vs. Kentucky, March 28, 1992.

The statistical graphic for the scoring in a basketball game allows one to glean the flow of the scoring for an entire basketball game in a glance. Would this type of graphic work for all sports? Consider football, where 1, 2, 3, or 6 points can be scored at a time. You could still produce the same statistical graphic for the scoring as we did for the basketball game, but football has other strategic elements (for example, field position) that are not captured with the single graphic alone.

The basketball graphic portrays quantitative variables on both the horizontal and vertical axes. Occasions arise when one of the variables is quantitative, but the other is qualitative, as is the case in the next example. This example also illustrates the case in which a statistical graphic is employed to help solve a mystery.

Example 1.3 The !Kung hunter-gatherers of Botswana and Namibia have long intervals between births, typically between 3 and 4 years, despite being a noncontracepting and nonabstinent population. Speculations linked the birth spacing to nutritional infertility, because the !Kung diet is sometimes low in calories, but no direct data had been collected to support this hypothesis.

Harvard anthropologists Melvin Konner and Carol Worthman investigated the unusually long birth spacings. Figure 1.2 shows one daylight cycle of interaction between a mother and her 14-day old son. As in the previous statistical graphics, the horizontal axis is time, which runs over the daytime interactions from 7:30 AM to 7:30 PM. The dependent variable here is not quantitative—it is one the following states: nursing, sleeping, holding, and crying. Furthermore, some of these states can occur simultaneously. For this reason, the states are labeled on the left using a text string and their durations are indicated by bars. Gaps between the bars imply that the state is not occurring. Nursing is placed at the top because it plays a central role in the conclusion that was drawn by Konner and Worthman.

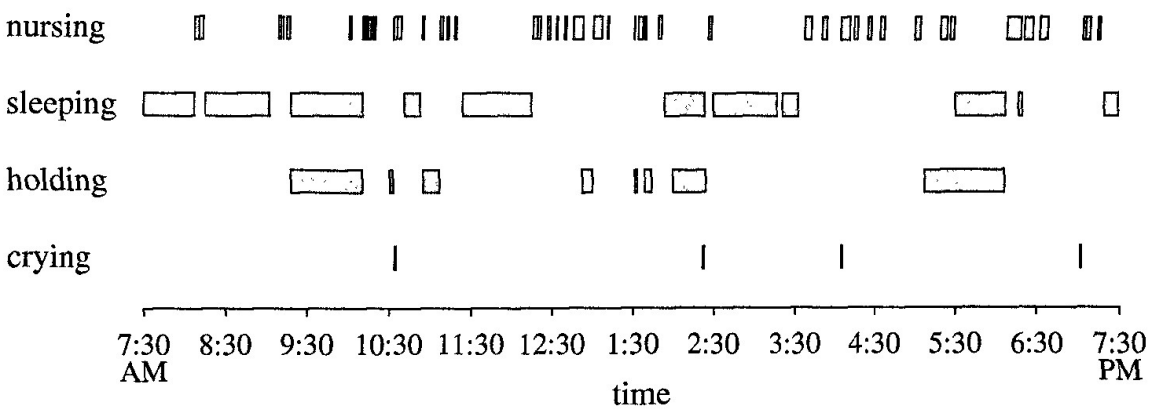


Figure 1.2: !Kung mother and baby daytime interactions.

The first thing that jumps out from the data (particularly to a Western mother) is the high frequency of nursing bouts. During the daytime hours when recordings were collected, there were 46 such bouts. This corresponds to a nursing bout every 15 minutes on average for this baby. There was a nursing bout every 13 minutes on average for all of the babies that they observed. Blood samples were also drawn daily on the nursing mothers. Konner and Worthman used the following logic to reject the conclusion that the low-calorie diet alone produced the long birth spacings: (1) nursing results in the release of the hormone prolactin, (2) prolactin has a half-life in the plasma of 10 to 30 minutes, (3) prolactin has an antigonadotrophic effect, which means that the mother will be less fertile if the prolactin level is high enough. So these frequent nursing bouts result in an elevated level of prolactin that results in the infertility of the mother. It is only late into the second year of life, when the baby's separations from the mother are longer as

the baby spends more time playing, that the mother once again becomes fertile. The investigators used the statistical graphic, blood tests, and some biochemistry to draw their conclusion.

The statistical graphics in the previous two examples have had time on the horizontal axis. The next example considers a plot that shows the relationship between two categorical variables.

Example 1.4 All of the statistical graphics presented thus far have been prepared in the R language, which is open-source software. In this example, the R code required to produce the statistical graphic is presented, which only requires a single line of code. R has a built-in data set named `HairEyeColor`, which gives counts of the hair and eye color of 592 statistics students at the University of Delaware. The hair color is classified into four levels: black, brown, red, and blond. The eye color is also classified into four levels: brown, blue, hazel, and green. (The gender of the students was also collected, but will be ignored in the statistical graphic created here.) The hair color and eye color are known to statisticians as *categorical variables*. One way to investigate the relationship between hair color and eye color is a *mosaic plot*, which can be used to visualize the relationship between two or more variables. The single R command given below produces a mosaic plot of the hair and eye color data.

```
mosaicplot(~ Hair + Eye, data = HairEyeColor)
```

The mosaic plot is shown in Figure 1.3. The four hair colors are depicted horizontally; the four eye colors are depicted vertically. The areas of each rectangle are proportional to the counts of each combination of hair and eye color. For example, there are only 5 students in the statistics class with black hair and green eyes, so that rectangle has the smallest area. At the other extreme, there are 119 students with brown hair and brown

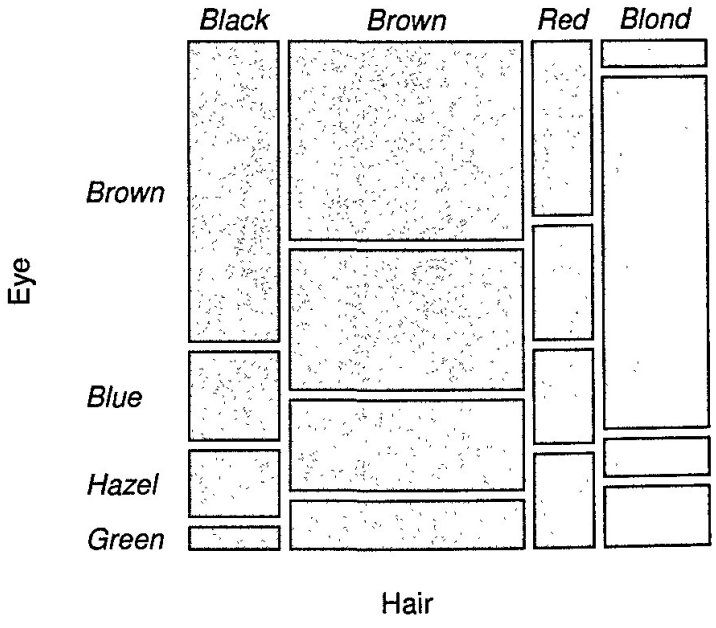


Figure 1.3: Mosaic plot of hair and eye color.

eyes, so that rectangle has the largest area. Clearly, based on the mosaic plot, hair color and eye color are related. If someone has blond hair, their eyes are most likely blue. If someone has black hair, their eyes are most likely brown.

Some statistical graphics don't require axes. *Word clouds*, which require no axes at all, are useful for showing frequencies of words in a text. They can also be used to quickly compare relative sizes of populations, relative number of search engine words, etc.

Example 1.5 A *word cloud* (also known as a wordle, tag cloud, or weighted list) uses the frequency of interesting words to give the viewer a quick visual overview of the content of a book, article, speech, or document. The words are packed as densely as possible into the cloud. Rotating some of the words so that they appear vertically allows the words to be packed more densely. The font size used on the three word clouds that follow is proportional to the frequency of the occurrence of the word in each text. I have arbitrarily chosen the top 50 "interesting" words in the three word clouds illustrated here. (This is a subjective process, but I left out words like "the," "is," "at," and "may.")

The Bible is divided into two parts: the Old Testament and the New Testament. I have used the 39 books for the Old Testament and the 27 books for the New Testament that are in common use by the Catholic, Protestant, Greek Orthodox, and Russian Orthodox churches, translated into the King James Version. The two different versions of the word "Lord" were counted together, making it the most frequent word in the text. The word cloud in Figure 1.4 contains the 50 most frequent interesting words that appear. By just viewing this graphic for a few seconds, it is easy to see what is emphasized in the text.

Figure 1.4: Word cloud of the Bible.

As a second contrasting example, Figure 1.5 contains a word cloud of the top 50 most frequent interesting words in William Shakespeare's play *Hamlet*. Several words appear (for example, "Denmark," "England") that could not have appeared in the first word cloud.

Figure 1.5: Word cloud of William Shakespeare's *Hamlet*.

The previous two word clouds are in stark contrast to the one for Mark Twain's *The Adventures of Huckleberry Finn* in Figure 1.6. Several of the words (for example, "reckon," "warn't," "ain't," "dey," "knowed") were common in the deep south at the time the book was written. The words "raft," "canoe," "river," and "woods" reflect the book's setting along the Mississippi River.

My hope is that these examples have conveyed the value of well-constructed statistical graphics. Their value is enormous in terms of both taking a first look at a data set for a statistician and communicating the information in a data set to non-statisticians. The following example includes

Figure 1.6: Word cloud of Mark Twain's *The Adventures of Huckleberry Finn*.

the R code necessary to produce some elementary statistical graphics that can be applied to a single data set of n values drawn from a univariate population. In this case the data set consists of $n = 100$ experimental estimates of the speed of light.

Example 1.6 We now know that the speed of light in a vacuum, oftentimes denoted by the constant c in Einstein's famous $E = mc^2$ formula, is 299,792.458 kilometers per second. The speed of light is slightly slower for light traveling through air—approximately 90 kilometers per second slower based on the refractive index of air. The speed of light in air also depends on the temperature and pressure of the air, so it is difficult to pin down one value. Before modern science emerged, many believed that the speed of light was infinite, meaning that light transmitted instantaneously. Galileo Galilei was one of the first to believe that the speed of light was finite. The first effort to determine the exact value of c was performed by Danish astronomer Ole Rømer in 1676, using the difference in the periods of Jupiter's innermost moon when the earth was approaching and receding from Jupiter. Christiaan Huygens combined Rømer's observation with an estimate of the diameter of the Earth's orbit to produce the first estimate of c, which was low by 26%. In 1879, Albert Michelson conducted experiments using a device with a rotating mirror (called an interferometer) to estimate the speed of light in air. These experiments resulted in the data set shown in Table 1.4. Each data value is the amount in excess of 299,000 kilometers per second. Each data value given is the average of ten experiments conducted by Michelson. The order of observation is given row-wise.

850	740	900	1070	930	850	950	980	980	880
1000	980	930	650	760	810	1000	1000	960	960
960	940	960	940	880	800	850	880	900	840
830	790	810	880	880	830	800	790	760	800
880	880	880	860	720	720	620	860	970	950
880	910	850	870	840	840	850	840	840	840
890	810	810	820	800	770	760	740	750	760
910	920	890	860	880	720	840	850	850	780
890	840	780	810	760	810	790	810	820	850
870	870	810	740	810	940	950	800	810	870

Table 1.4: Estimates of the speed of light in air (add 299,000 km/sec to each value).

The data displayed in Table 1.4 are rather unwieldy. Since the *order* that the observations were collected is not of particular interest, perhaps the data is better presented in sorted order, as in Table 1.5. This display is much more helpful to the reader. Clearly, the observations range from 299,620 kilometers per second to 300,070 kilometers per second, with the majority of the data values falling between 299,800 kilometers per second and 299,900 kilometers per second.

Although this second table is an improvement over the first, it is still rather difficult for the reader to intuit the *shape* of the probability distribution associated with this data set, even by spending significant time staring at the table of sorted observations. One crude, text-based graphic that can help determine the shape of the probability distribution is known as a *stem-and-leaf plot*. These plots were popularized in the 1970s, when monospace fonts (all numbers and letters use the same amount of horizontal space `like this`) were common on computer terminals and hard copy. A vertical line separates the

620	650	720	720	720	740	740	740	750	760
760	760	760	760	770	780	780	790	790	790
800	800	800	800	800	810	810	810	810	810
810	810	810	810	810	820	820	830	830	840
840	840	840	840	840	840	840	850	850	850
850	850	850	850	850	860	860	860	870	870
870	870	880	880	880	880	880	880	880	880
880	880	890	890	890	900	900	910	910	920
930	930	940	940	940	950	950	950	960	960
960	960	970	980	980	980	1000	1000	1000	1070

Table 1.5: Estimates of the speed of light in air (add 299,000 km/sec to each value).

stem values, which fall to the left of the line, and the leaf values, which fall to the right of the line. The first step in constructing a stem-and-leaf plot is to sort the data in ascending order, as in Table 1.5. The next step is to determine the meaning of the stem values. For the speed of light data, this will be the leftmost digit of the data values from Table 1.5 for the first 96 sorted observations, then 10 for the last four observations. The last step is to create a leaf entry for each data value. For the speed of light data set, we ignore the rightmost digit for each data value (because it is always zero) and use the tens digit of each entry as the leaf. The results of this process for the speed of light data are shown in Figure 1.7. There must be the same amount of horizontal space allocated to each leaf value for the shape of the probability distribution to be meaningful. The stem-and-leaf plot is one of the few statistical graphics in which the data set can be reconstructed from the plot. The stem-and-leaf plot can be viewed as an estimate of the probability density function of Michelson's observations as follows. Rotate your book 90° counterclockwise and look at the leaf values that are above the (now horizontal) line. These leaf values form a crude digital histogram by displaying the various cell frequencies. Not surprisingly, the speed of light estimates have a bell-shaped distribution, centered around the sample mean $\bar{x} = 852$, which is slightly higher than the known population mean of approximately $\mu = 792$. The spread of the probability distribution associated with the observations is partially explained by the error in the measuring device used in the experiment in the late 1800s.

The following two lines of code in R are used to produce a stem-and-leaf plot similar to the one in Figure 1.7.

```
x = scan("michelson.d")
stem(x)
```

```
  6 | 25
  7 | 22244456666678899 9
  8 | 000001111111111223344444444555555556667777888888888999 9
  9 | 001123344455566667888
 10 | 0007
```

Figure 1.7: Stem-and-leaf plot of the estimates of the speed of light in air.

The data set is stored in a file named `michelson.d` and is read into the vector of length $n = 100$ by R's `scan` function. Figure 1.8 shows the stem-and-leaf plot produced by the `stem` function. The `stem` function decided that the five cells in the stem-and-leaf plot plotted by hand in Figure 1.7 were not adequate, and decided internally to plot ten cells instead. I added a legend on the left that describes the meaning of the smallest and largest data values. This common practice helps the viewer interpret the values in the plot. Typing `help(stem)` at the command prompt in R gives the options associated with the `stem` command.

```
6 | 2 = 299,620 km/sec

  6 | 2
  6 | 5
  7 | 222444
  7 | 566666788999
  8 | 00000111111111223344444444
  8 | 55555555666777788888888899
  9 | 0011233444
  9 | 55566667888
 10 | 000
10 | 7 = 300,070 km/sec
 10 | 7
```

Figure 1.8: Another stem-and-leaf plot of the estimates of the speed of light in air.

Another common statistical graphic that captures the shape of the probability distribution of a data set is the *histogram.* Compared with the stem-and-leaf plot, the histogram is a bit more aesthetically pleasing and allows greater flexibility in choosing the cells that contain the data values. Unlike the stem-and-leaf plot, however, you are not able to recreate the data values from a histogram. The R statements

```
x = scan("michelson.d")
hist(x)
```

produce a histogram of the data values, which is plotted in Figure 1.9. The units on

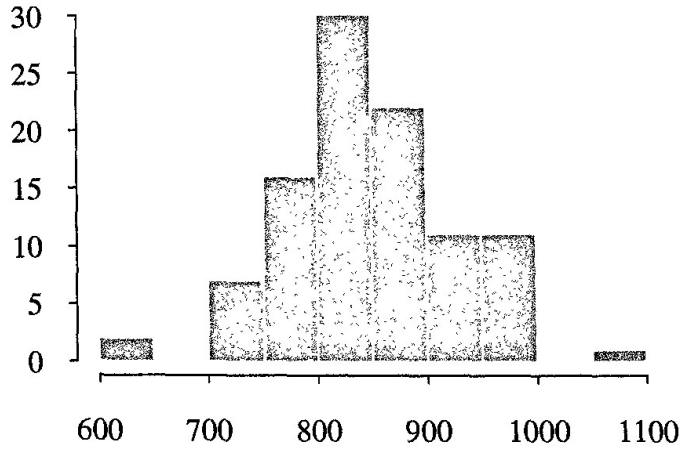


Figure 1.9: Histogram of speed of light estimates (km/sec over 290,000) in air.

the horizontal axis are the speeds in excess of 290,000 kilometers per second and the units on the vertical axis are the counts of observations falling into the cells (600, 650], (650, 700], ..., (1050, 1100]. The units on the vertical axis are the number of observations falling in each bin. Since the histogram is the statistical analog of the probability mass function or probability density function $f(x)$, some statisticians prefer to alter the vertical axis units so that the area under the histogram is one, just as it is for $f(x)$. Simply add the `probability = TRUE` option in the call to `hist` to alter the vertical scale in this fashion. This allows the analyst to superimpose a hypothesized probability density function on top of such a histogram to illustrate a potential population probability distribution that might have produced such a data set.

The main strength of the histogram is that it is one of the better ways of assessing the shape of a probability distribution. This shape can lead to a short list of probability models for the population distribution. Histograms also have several weaknesses. The first weakness concerns the arbitrary grouping of observations into cells. Choosing the number of cells and cell boundaries for a histogram is rather important for the following reasons.

- Choosing too few cells can mask important features of the data set. The default number of cells was chosen by R in Figure 1.9. Sturges's rule suggests using $1 + \log_2 n$ cells.
- Choosing too many cells can highlight the natural *random sampling variability* (that is, the chance fluctuations in the data associated with a finite sample size) rather than the shape of the parent probability distribution.

Even if you choose the right number of cells, shifting the cells slightly to the left or the right can cause subtle or even dramatic differences in the shape of the histogram. For the aesthetics of the histogram, it is preferable to have round numbers as the cell boundaries. The number of cells should increase with the sample size because random sampling variability is less pronounced for larger data sets.

The second weakness associated with histograms is that they are notoriously bad at comparing multiple populations; they do not stack well. One exception to this is the case of two populations, where the two histograms can be placed side-by-side. This case is illustrated in the following example.

Example 1.7 Demography is the statistical study of human populations. One aspect of human populations that can be summarized by two histograms is the age distribution for a particular sub-population. A *population pyramid* or *age structure diagram* consists of two histograms rotated 90° and placed side-by-side. These diagrams illustrate the longevity, birth rate, and probability distribution of the ages of the population by gender, race, etc. When two population pyramids are compared for the same sub-population at two different points in time, they reveal population dynamics due to various factors such as medical advances or immigration.

Figure 1.10 contains a population pyramid for France on January 1, 1960 using data from `www.insee.fr`. There are 100 ages plotted on the vertical axis, and corresponding male populations on the left and female populations on the right, in thousands, on the horizontal axis. The most pronounced features of this statistical graphic are the nearly-symmetric dents in the populations that achieve their lowest level at ages 19 and 44. These dents cannot be attributed to random sampling variability because the sample size is so large. So what caused the dents?

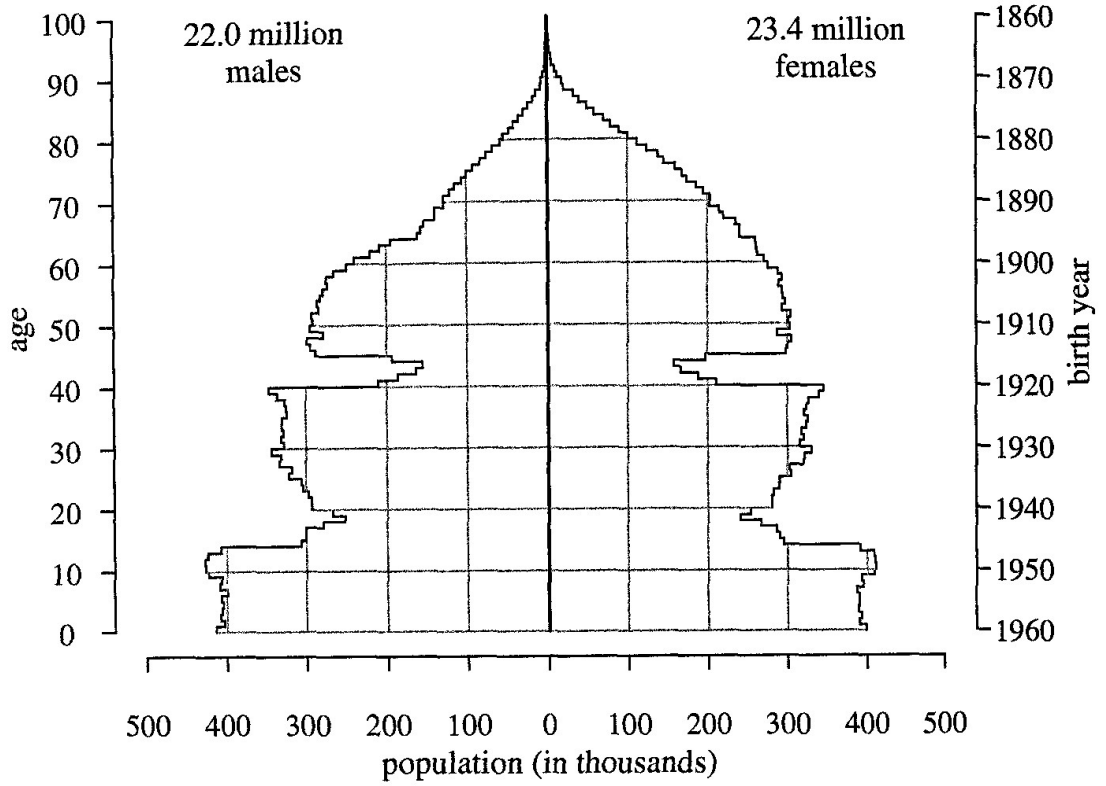


Figure 1.10: Population pyramid for France, January 1, 1960.

Using the birth year on the right-hand vertical scale, it can be concluded that the tragic effects of two world wars fought on French soil are the cause of the dents in the population pyramid. The durations of World War I (1914–1918) and World War II (1939–1945) coincide with the dents. There was a decreased birth rate during the wars, and a post-war baby boom after each war. Both male and female births are affected equally by the decline and subsequent bump in the birthrate. Looking more closely at the elderly population at the top of the graphic also indicates that there are significantly more elderly women than men. Could this be due to the increased longevity that women have over men, or are other factors at play? France sustained 1.7 million casualties in World War I and 600,000 casualties in World War II. These casualties would also account for some of the difference between the male and female population sizes. If you would like to see population pyramids for all countries for any year between 1950 and 2100, visit `populationpyramid.net`.

Population pyramids are useful ways of summarizing the age distribution of a population. Some extensions are listed below.

- Although a matter of taste, the gray grid lines inside of the pyramid could be removed, giving more emphasis to the shape of the two histograms.
- Labels are often added to a population pyramid to explain features of the pyramid (for example, dents).
- Population pyramids can be viewed over time, giving a dynamic sense as to how population is changing over time. In the case of the population pyramid of France on January 1, 1960, the dents would float upward as time advances.

- The two sides of the pyramid are not limited to male and female populations. The left side could be for left-handed individuals and the right side could be for right-handed individuals.
- Likewise, the vertical axis need not be age for constructing a pyramid of this type. The vertical axis could be SAT scores for a cohort of students applying to a particular university.
- The horizontal axis on most population pyramids is the population size, as in the previous example. But this need not be universally true. If the histogram on the left corresponds to the population (both male and female) in Europe, and the histogram on the right corresponds to the population (both male and female) in New Zealand, then it makes sense to use percent of population on the horizontal axis in order to assess the difference between the two age distributions.

Histograms are not an appropriate vehicle for comparing more than two populations simultaneously. A *box plot* (also known as the *box and whisker plot*) is a convenient statistical graphic for comparing multiple data sets simultaneously. Five numbers are used to summarize a data set in a box plot:

- the sample minimum (the smallest observation),
- the first quartile (the sample 25th percentile),
- the second quartile (the sample 50th percentile, also known as the sample median),
- the third quartile (the sample 75th percentile),
- the sample maximum (the largest observation).

Like the stem-and-leaf plot and the histogram, a box plot is "nonparametric" or "distribution-free" in the sense that no assumptions are made about the population distribution from which the data was drawn. Box plots can be drawn horizontally or vertically. A scale is typically included near the box plot. Returning to the speed of light data set, a box plot can be drawn with the R commands

```
x = scan("michelson.d")
boxplot(x)
```

which display the vertically-oriented box plot shown in Figure 1.11. The vertical axis displayed is the data values in km/sec in excess of 290,000 km/sec.

The difference between the third quartile and the first quartile, which captures the middle 50% of the data, is known as the *interquartile range* and is often abbreviated IQR. The interquartile range is a measure of the *dispersion*, *variability*, or *spread* of the data values. In a box plot, this is the height of the box when the box plot is arranged vertically as in Figure 1.11.

The symmetry of the data set is also apparent from the box plot. If the sample median (the middle line in the box) is about the average of the two ends of the box and is also about the average of the two extreme observations, then it is reasonable to conclude that the data set was drawn from a nearly symmetric distribution, and therefore the population skewness of the population probability distribution is approximately zero.

The box plot described here is the most common, but there are several variations on box plots. Although the ends of the boxes are universally the first and third quartiles, the whiskers that extend from the box do not always extend to the extremes. One common practice is to let the whiskers extend to a certain sample percentile, then include data values beyond this percentile with dots. Another practice is to place a notch in the box around the sample median value to indicate its

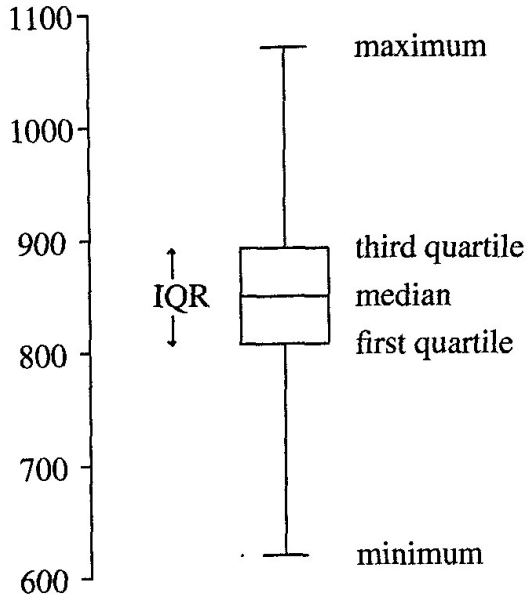


Figure 1.11: Box plot of the estimates of the speed of light in air.

precision. Still another practice is to let the width of the box reflect the sample size. More detail is given on most websites that describe box plots.

The real value of a box plot is not in just describing a single data set, as in Figure 1.11, but rather in comparing two or more data sets. The following example illustrates such a comparison.

Example 1.8 The average daily maximum temperature (in degrees Fahrenheit) for three U.S. cities (Monterey, California; Portland, Oregon; New York City, New York) in the year 2000 is given in Table 1.6.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Monterey	58.4	61.0	61.7	65.5	66.1	68.8	64.8	68.2	73.8	65.7	61.1	62.3
Portland	45.1	50.2	53.9	64.1	66.0	76.4	78.4	78.6	73.9	63.2	49.5	45.5
New York	37.9	43.7	54.9	58.2	71.6	78.8	79.0	78.6	72.9	64.4	50.9	37.2

Table 1.6: Average maximum temperature in three cities in 2000.

Plotting the temperatures for the three cities over time would be cluttered because the curves would intersect one another at several points in time. A more elegant statistical graphic is to compare box plots of the data for the three cities. Although the time dependency is lost, the extreme values and quartiles are easily compared. The R statements

```
m = c(58.4, 61.0, 61.7, 65.5, 66.1, 68.8, 64.8, 68.2, 73.8, 65.7, 61.1, 62.3)
p = c(45.1, 50.2, 53.9, 64.1, 66.0, 76.4, 78.4, 78.6, 73.9, 63.2, 49.5, 45.5)
n = c(37.9, 43.7, 54.9, 58.2, 71.6, 78.8, 79.0, 78.6, 72.9, 64.4, 50.9, 37.2)
boxplot(m, p, n)
```

produce a plot similar to that shown in Figure 1.12, where the three box plots are oriented horizontally. Since there are $n = 12$ data values for each city, the sample median

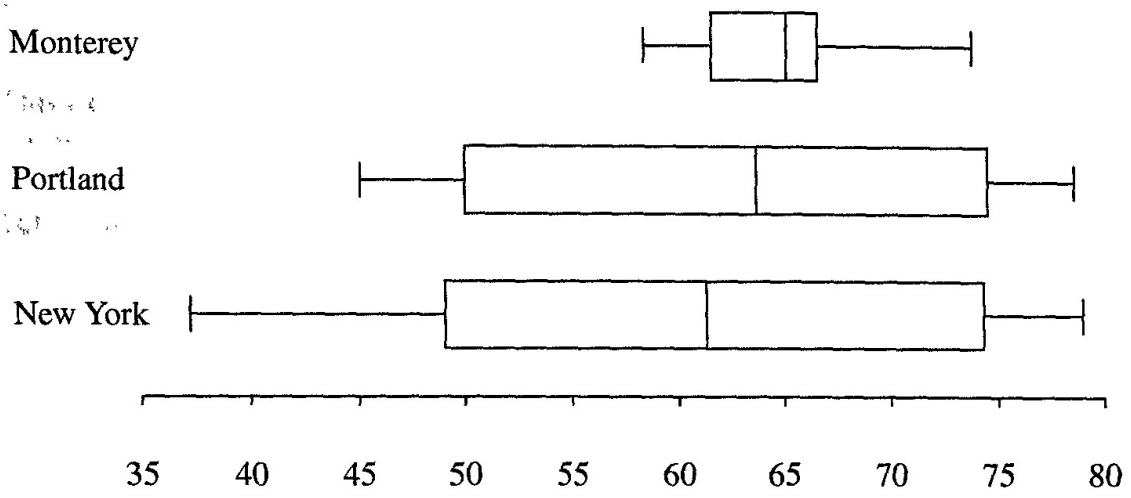


Figure 1.12: Box plots for the monthly average high temperature for three U.S. cities.

is calculated by averaging the two sorted middle values. One immediate conclusion that can be drawn from the box plots is that the sample medians are all quite close. The difference between the highest sample median (Monterey) and the lowest sample median (New York City) is less than four degrees. Even though the central tendency is nearly the same, the variability is drastically different. Monterey, California has one of the world's best climates, as reflected by both the top box plot and the associated housing prices. Portland is next in terms of variability, followed by New York City, whose residents endure the harshest winters and hottest summers of the three cities. What is the cause of the difference in variability? The earth's rotation gives cities on the west coast the advantage of warmer winters and cooler summers because of the damping effect on the temperature of the air that passes over the Pacific Ocean. The effect of the Atlantic Ocean on temperatures in New York City is minimal.

Box plots provide a way of comparing multiple probability distributions simultaneously. They also provide a way of describing a probability distribution that avoids the binning of data (that is, placing data values into cells) that is present in histograms. There is another statistical graphic that avoids binning data. The *empirical cumulative distribution function* is the statistical analog of the cumulative distribution function. The empirical cumulative distribution function is a step function with upward steps of height $1/n$ at each data value.

Another way of thinking about why the empirical cumulative distribution function is a reasonable estimate of $F(x)$ is as follows. If you just had the data values $x_1, x_2, \ldots, x_n$ and wanted to generate a "best guess" for $f(x)$, one option is to create an empirical probability mass function $\hat{f}(x)$ that has mass $1/n$ at each data value. (Statisticians use the hat, or caret, above f to indicate that $\hat{f}(x)$ is an estimator of $f(x)$.) This works fine if each of the data values is unique. If there are d values that are tied, then there will be a mass value of d/n at the tied value. Now what would the empirical cumulative distribution function $\hat{F}(x)$ associated with this empirical probability mass function look like? It would be exactly the one described above. It would have an upward step of height $1/n$ at each unique data value and an upward step of height d/n when there are d tied data values.

The good news about the empirical cumulative distribution function is that no binning is required, which means that an empirical cumulative distribution function is unique for a particular data set. The bad news is that its shape is not quite as distinctive as the histogram.

Example 1.9 Returning to Michelson's $n = 100$ estimates of the speed of light measured in excess of 290,000 km/sec, the empirical cumulative distribution function can be plotted with the R function `plot.ecdf` as shown in the code below.

```
x = scan("michelson.d")
plot.ecdf(x, verticals = TRUE, pch = "")
```

The empirical cumulative distribution function is plotted in Figure 1.13. There are several options for displaying the empirical cumulative distribution function, and the one displayed here has the vertical risers included on the steps. Some prefer these risers left off because the step function, after all, is still a function. This is largely a matter of personal taste.

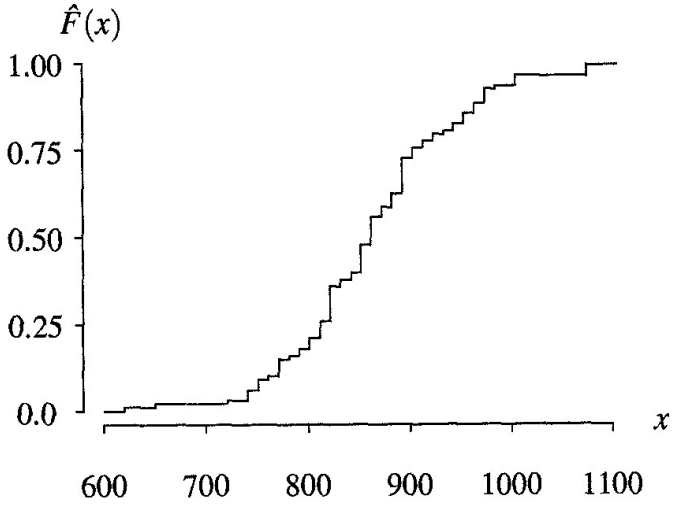


Figure 1.13: Empirical cumulative distribution function for the speed of light data.

Similar to overlaying a probability density function on top of a histogram, one often overlays a theoretical cumulative distribution function on top of an empirical cumulative distribution function. This will be illustrated in the next chapter.

This next example is drawn from business and finance. It illustrates the benefit of a logarithmic scale for displaying certain types of data sets.

Example 1.10 The Dow Jones Industrial Average (DJIA), also known as the Dow 30, was devised by Charles Dow and was initiated on May 26, 1896. The average bears Dow's name and that of statistician and business associate Edward Jones. The DJIA is the average stock price of 30 U.S.-based, publicly traded companies, adjusted for stock splits and the swapping of companies in and out of the average so that it adequately reflects the composition of the domestic stock market. These adjustments are made by altering the average's denominator for historical continuity, which is now much less than 30. The value of the denominator is given every day in the *Wall Street Journal*.

The evolution of the DJIA is not a true reflection of the yield of the 30 stocks because two important factors are not incorporated into the average. First, the average does not factor in dividends that are paid by some of the 30 stocks. Second, the average does not factor in inflation, which erodes the true return that a stock investment provides.

If dividends were factored in, the DJIA would be much higher than it is presently; if inflation were factored in, the DJIA would be much lower than it is presently.

This example develops statistical graphics associated with the DJIA that illustrate various ways to view its evolution over time. The first is a plot of the average annual DJIA closing values during the 20th century. This plot is generated with the R code given below.

```
x = 1901:2000
y = scan("djia")
plot(x, y, type = "l")
```

The file `djia` contains the 100 annual average closing values. The resulting graph is shown in Figure 1.14. Data sets of this nature in which a response variable is plotted over time are known as *time series*. Most economics and statistics departments at universities offer classes titled "time series analysis" in which probabilistic models are developed for describing a time series.

The DJIA had a sample mean closing value of 69.52 during 1901 and a sample mean closing value of 10731.15 during 2000. The linear vertical scale that is used in Figure 1.14 obscures most of the variability of the DJIA during the first half of the century. The graph can be made more meaningful by using a logarithmic scale on the vertical axis. This is accomplished by adding the `log = "y"` parameter to the `plot` command in the R code, resulting in the graph shown in Figure 1.15. Labels have been added to help highlight events that might have influenced the DJIA.

The stock market crash in October of 1929 that initiated the Great Depression is much more pronounced in Figure 1.15. The DJIA had peaked with a close of 381.20 on

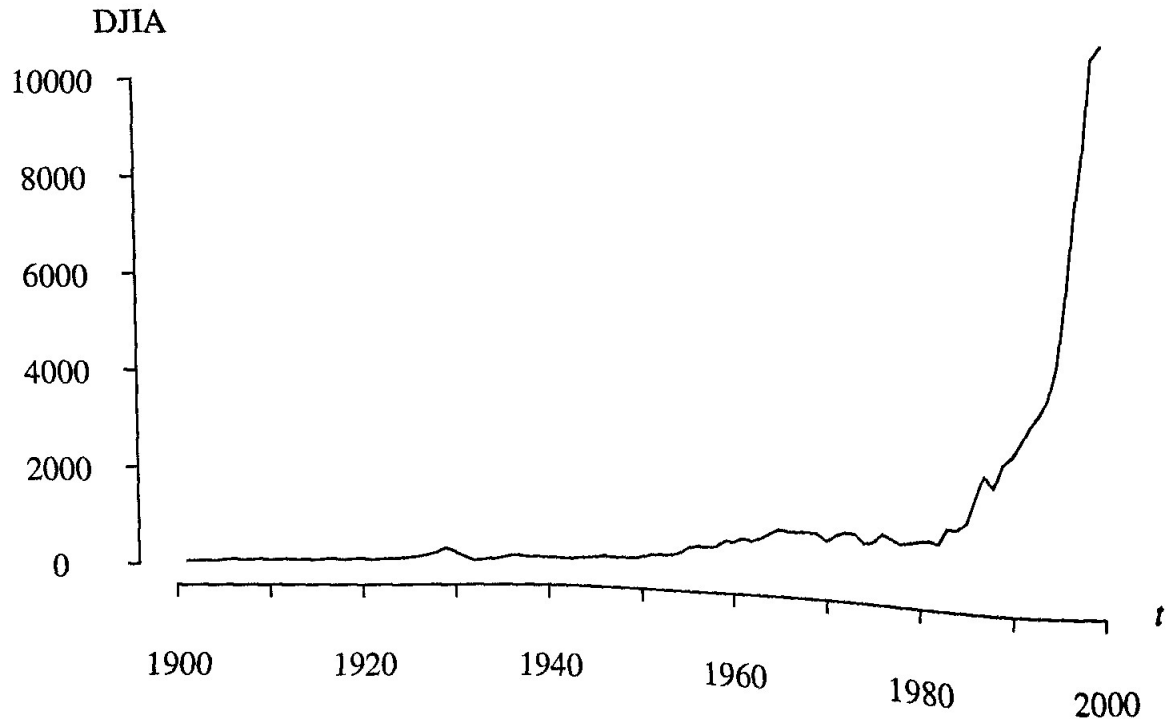


Figure 1.14: Dow Jones Industrial Average, 1901–2000.

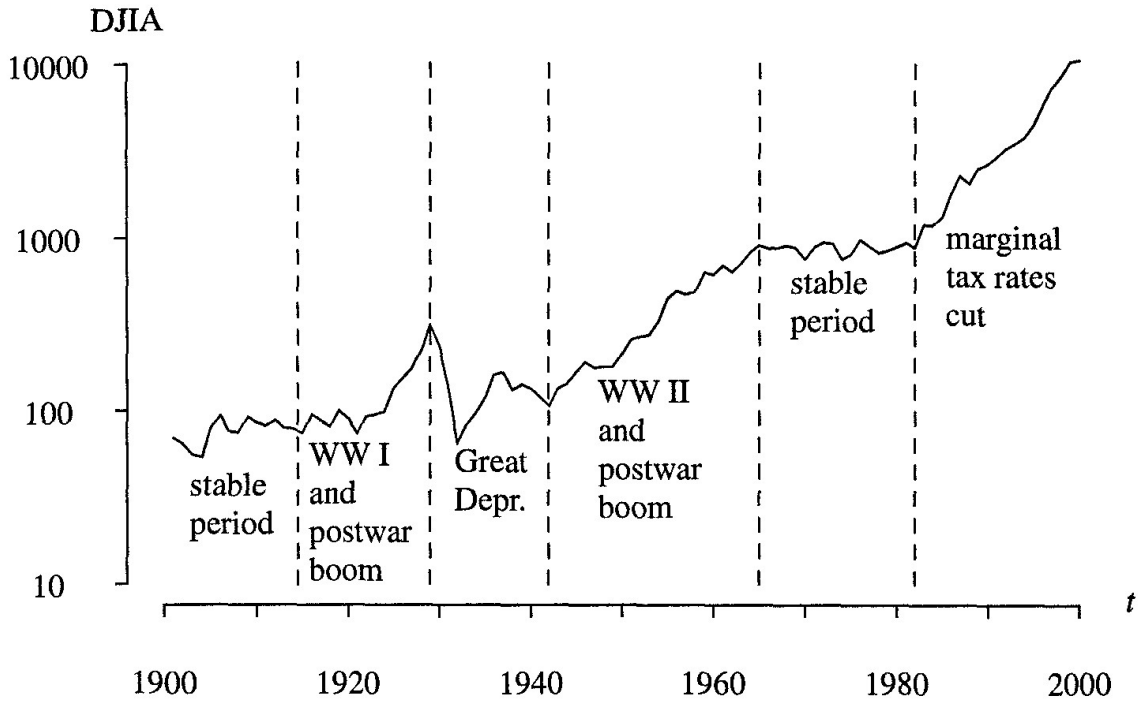


Figure 1.15: Dow Jones Industrial Average, 1901–2000.

September 3, 1929. The market bottomed out on July 8, 1932 when it closed at 41.20, which corresponds to a loss of almost 90%. Each of the two World Wars fought during the twentieth century was followed by a sustained bull market in the DJIA. The top marginal income tax rate was lowered from 70% to 28% and the federal budget was brought into balance in the 1980s and 1990s, resulting in a prolonged growth in the DJIA.

The next statistical graphic provides more detail associated with the DJIA during the first year of the twenty-first century. The DJIA closed on December 29, 2000, the last trading day of the century, at 10787.99. The DJIA closed on December 31, 2001 at 10021.57. This 7.1% decline is in part due to the terrorist attacks on the U.S. on September 11, 2001. The 7.1% decline is an average for the 30 stocks. Some performed better and some performed worse. The stocks that comprised the DJIA during 2001 are given in Table 1.7. Their ticker symbols are given in parentheses. General Electric is the only company that was in the original DJIA from its inception in 1886. A statistical graphic can be devised that captures the following five pieces of information about the 30 stocks comprising the DJIA from December 29, 2000 to December 31, 2001:

- the stock ticker symbol,
- the market sector,
- the absolute market capitalization (by including a legend),
- the relative market capitalization,
- the one-year performance.

This graphic is shown in Figure 1.16. There are 30 rectangles for each of the 30 stocks in the DJIA. The area of each rectangle is a monotonically increasing function of the

Alcoa (AA)	Allied Signal (ALD)	American Express (AXP)
AT&T (T)	Boeing (BA)	Caterpillar (CAT)
Citigroup (C)	Coca-Cola (KO)	DuPont (DD)
Exxon (XOM)	General Electric (GE)	General Motors (GM)
Hewlett-Packard (HPQ)	Home Depot (HD)	Intel (INTC)
IBM (IBM)	International Paper (IP)	Johnson & Johnson (JNJ)
J.P. Morgan (JPM)	Kodak (EK)	McDonalds (MCD)
Merck (MRK)	Microsoft (MSFT)	3M (MMM)
Philip Morris (PM)	Procter & Gamble (PG)	SBC Communications (SBC)
United Technologies (UTX)	WalMart (WMT)	Walt Disney (DIS)

Table 1.7: Dow Jones Industrial Average Companies in 2001.

market capitalization (that is, the value of the publicly-traded shares of stock). The scale of the market capitalization is seen in the legend in the lower-right hand corner. Thinner lines separate individual stocks. Thicker lines separate the stocks by sector, and the sector labels are given outside of the large rectangle in italics. Each of the 30 rectangles contains a ticker symbol to identify the stock and the performance during 2001 as a percentage. Of the 30 stocks, 11 stocks increased throughout the year and 19 stocks decreased. The energy sector was the strongest throughout 2001; the financial sector was the weakest throughout 2001. Rectangular-shaped plots of this nature have generally replaced the more traditional pie diagram because of their ability to easily capture additional information in the smaller rectangles.

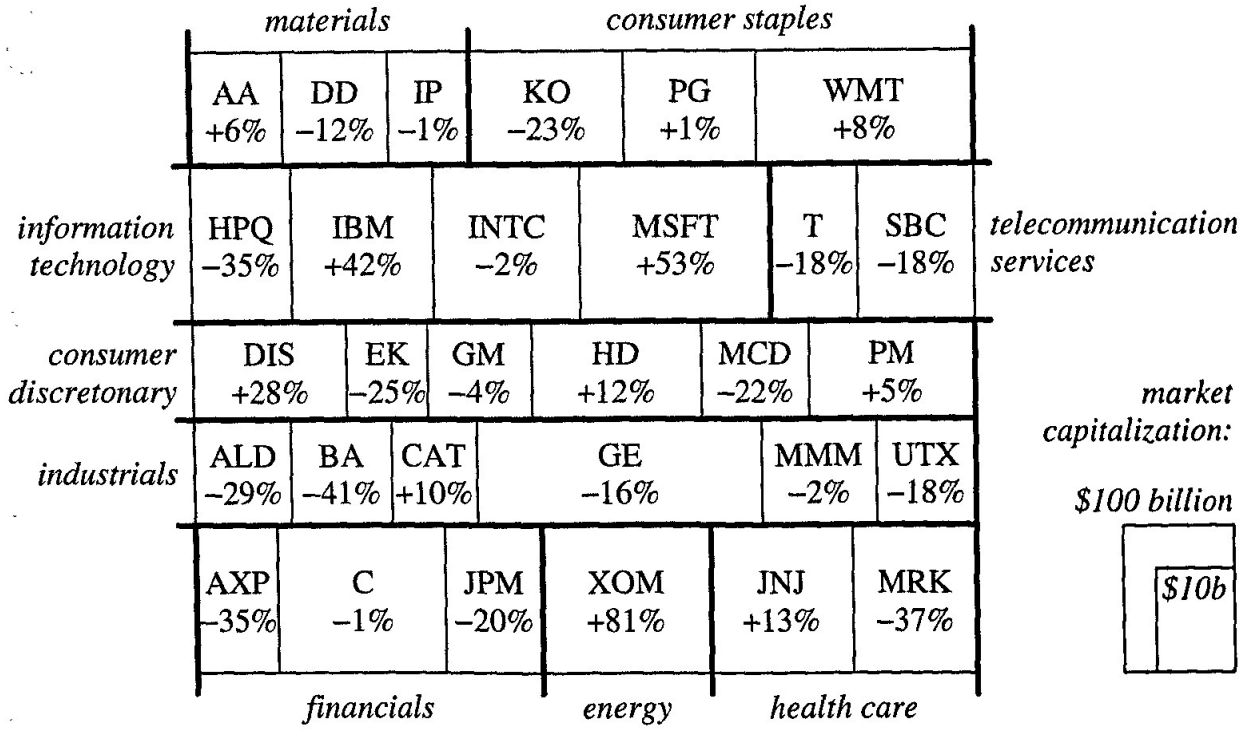


Figure 1.16: Dow Jones Industrial Average component stocks performance in 2001.

The applications of statistics span a wide range of disciplines. So far we have encountered data sets associated with comparing population sizes, showing the dynamics associated with a basketball game, solving a mystery concerning birth spacings, using a mosaic plot to visualize the relationship between hair and eye color, displaying word frequencies in a book, analyzing estimates of the speed of light in air, comparing the age distribution of men and women in France, comparing weather data for three U.S. cities, and displaying stock market data. This section ends with one final example that concerns the display of the estimate of a mixed discrete–continuous probability distribution. Mixed discrete–continuous random variables occur, for example, in queueing (the waiting time for a server), meteorology (the total rainfall in one day), and reliability (the lifetime of a product). In all three of these examples, there is a non-zero probability that the random variable will assume a value of zero, which accounts for a discrete portion of the probability distribution.

Example 1.11 A play-by-play account of the 2016 National Football League (NFL) regular-season games contains the field position after $n = 2593$ kickoffs. Kickoffs are further subdivided into those returned in the field of play and those with discrete categorical outcomes (safeties, touchbacks, out-of-bounds, and touchdowns). When a run-back is attempted by the kickoff returner and the return is concluded in the field of play (usually by tackling the kickoff returner but occasionally by a fumble recovery), the resulting field position is a continuous random variable with a support ranging from 0 to 100, measured as the distance from the return team's end zone. The categorical outcomes are associated with a discrete random variable with four mass values (0, 25, 40, and 100). Table 1.8 highlights the division between the continuous (1047 observations) and discrete (1546 observations) portions of the probability distribution and shows the frequency of the various outcomes during the 2016 season.

Type	Category	Starting field position	Frequency	Probability
Continuous	Returned in the field of play	$(0, 100)$	1047	$\frac{1047}{2593} \cong 0.404$
Discrete	End zone (return team)	0	3	$\frac{3}{2593} \cong 0.001$
	Touchback	25	1518	$\frac{1518}{2593} \cong 0.585$
	Out-of-bounds	40	18	$\frac{18}{2593} \cong 0.007$
	End zone (kicking team)	100	7	$\frac{7}{2593} \cong 0.003$
		Total:	2593	$\frac{2593}{2593} = 1.000$

Table 1.8: NFL 2016 regular-season kickoff starting field positions.

Let the random variable X be the starting field position following a NFL kickoff during the 2016 season measured in yards from the return team's end zone (regardless of whether a turn-over occurs). We wish to construct a statistical graphic that captures an estimate of the probability distribution of the starting field position X from the $n = 2593$ data values. Since the probability distribution of X is a mixed discrete–continuous random variable, a reasonable estimate of the contribution of the two parts of the probability distribution is the finite mixture

$$\hat{f}(x) = \frac{1047}{2593}\hat{f}_C(x) + \frac{1546}{2593}\hat{f}_D(x),$$

where $\hat{f}_C(x)$ and $\hat{f}_D(x)$ are

$$\hat{f}_C(x) = \begin{cases} \text{kernel density function} \\ \text{of outcomes returned} \\ \text{in the field-of-play} \end{cases} \quad \text{and} \quad \hat{f}_D(x) = \begin{cases} 3/1546 & x = 0 \\ 1518/1546 & x = 25 \\ 18/1546 & x = 40 \\ 7/1546 & x = 100 \end{cases}$$

and the hats denote estimators. Figure 1.17 displays the estimator for the 2016 NFL data. Even though a histogram would have worked perfectly fine for the estimate of the continuous portion of X, using a *kernel density function* emphasizes the continuous nature of the spotting of the ball on the field of play. Two different vertical scales (the scale for the continuous portion is on the left, labeled PDF for probability density function, and the scale for the discrete portion is on the right, labeled PMF for probability mass function) were necessary to avoid having the continuous portion crunched down to the horizontal axis. The scales were selected to reflect the approximately 40/60 split between the continuous and discrete portions of the probability distribution. A square root scale on the discrete axis make is easier to differentiate between the three discrete-but-unlikely outcomes. Two distinct modes are evident in the continuous portion. The first—with a mode at the 22 yard-line—represents distances most often attained before a returner is tackled. The second—just past mid-field—is a consequence of the 54 onside kick attempts during the 2016 season.

Well-designed statistical graphics are of benefit to statisticians and non-statisticians alike. They display aspects of a data set that are often difficult to see by viewing the raw data or by viewing the data in tables. Statistical graphics are capable of displaying multiple variables simultaneously, and it is often apparent from a display how the variables are related to one another.

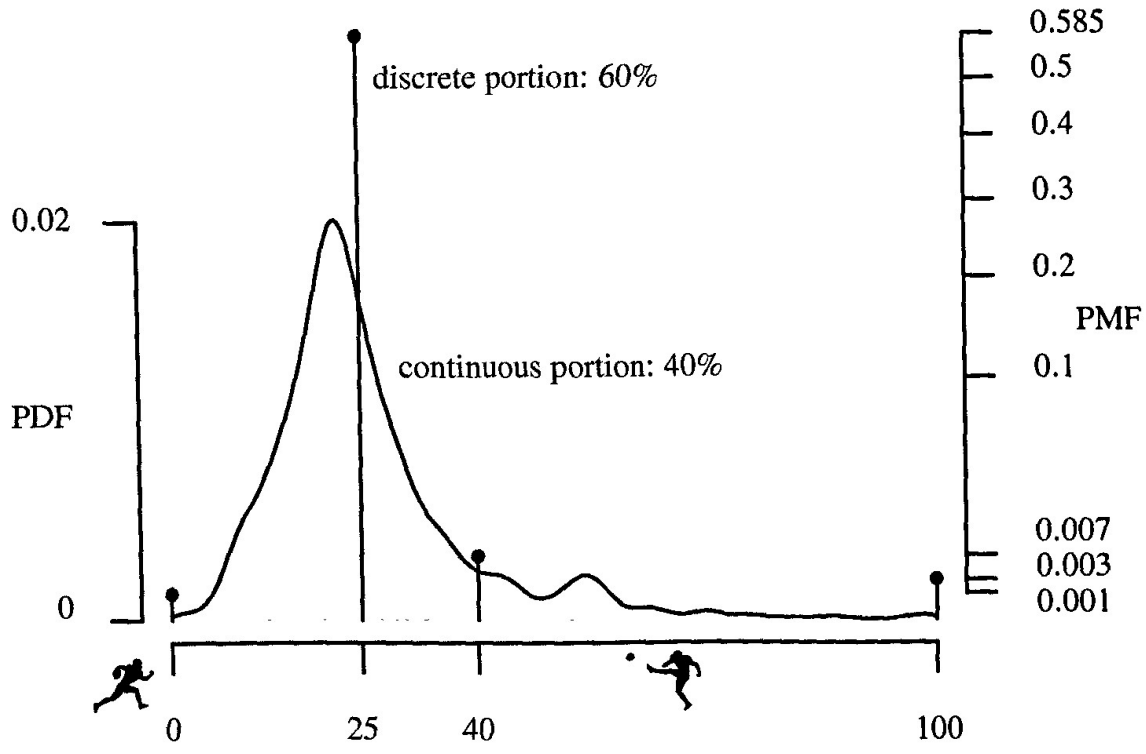


Figure 1.17: 2016 NFL regular season starting field position (yards into the field of play).

1.2 Random Sampling, Statistics, and Sampling Distributions

Statistical graphics give you a toolbox of techniques that are useful for visually summarizing a data set. Unfortunately, two knowledgeable people might draw opposite conclusions from a well-designed statistical graphic. This begs for a mathematical framework, which draws from probability theory, that can be used to remove personal opinion from the process. Two examples of questions that can be addressed by statistics are given below.

- How long, on average, does a particular brand of 60-watt light bulbs last under commonly-encountered environmental conditions?
- Should radiation or chemotherapy or both be used to treat a particular cancer?

There are lots of loose ends that have not been determined with these questions. Are the light bulbs burned continuously? Are the light bulbs used in an optimal environment (like a living room) or a high temperature or high vibration environment (like a rocket launch)? Are all people having the cancer women with diabetes, high blood pressure, and low cholesterol of approximately the same age? In order to answer the statistical questions, it is important to first pin down the details associated with the setting.

Statistical models used to describe data behave in a similar manner to mathematical models in economics, chemistry, or physics. A good model will be a close match to what is actually observed. A bell-shaped histogram, for example, is evidence that a normal population probability model could be an appropriate modeling assumption. The histogram would provide even more evidence, however, if there were $n = 1000$ observations producing the bell-shaped histogram rather than only $n = 20$ observations.

We have thus far avoided the question of how data values are collected. Now is the time to address that question. There are many different sampling mechanisms that can be used to cull the data. To keep the mathematics simple initially, we assume that *univariate data* is being collected on n subjects. This leads to the following definition of a random sample.

Definition 1.1 Let $X_1, X_2, \ldots, X_n$ be mutually independent random variables, each with the same but possibly unknown probability distribution described by $f_X(x)$. Realizations of the random variables $X_1, X_2, \ldots, X_n$ constitute a *random sample.*

There are several loose ends associated with the definition of a random sample that are outlined below.

- The integer n is known as the *sample size.*
- Some textbook authors refer to a random sample as a *simple random sample* (SRS).
- Mutually independent random variables, each with the same probability distribution, are often described using the abbreviation *iid* for "independent and identically distributed."
- Definition 1.1 implies that the joint distribution of the random sample can be found by

$$f(x_1, x_2, \ldots, x_n) = f_X(x_1) f_X(x_2) \ldots f_X(x_n).$$

- Definition 1.1 applies equally well to random sampling from discrete populations, continuous populations, and mixed discrete–continuous populations.

- The random sampling of vectors, that is, multivariate random variables, is considered in an advanced course.

- The assumption of mutual independence implies that each value must be sampled in a manner so that it is not influenced by any of the other values.

- Taking a "random sample" in industrial applications oftentimes requires significant effort. If a farmer delivers a truckload of potatoes to a potato chip company, the n potatoes sampled for quality should be selected from random positions in the truck, perhaps generated by a random number generator.

- Beware of selection bias. The famous headline "Dewey Defeats Truman" from the *Chicago Tribune* after the 1948 U.S. Presidential election was partially based on polling. If these polls were conducted by phone and more Republicans had phones than Democrats, then the estimate of the probability of Dewey defeating Truman would be biased.

- Beware also of response bias. Questions like "Do you use LSD?" or "Did you cheat on the French exam?" might not yield an honest response, which would lead to a biased estimator of the associated probabilities.

- Count the costs associated with collecting data. Sometimes a simple survey question can produce a data value cheaply. On the other hand, automobile safety data might involve crashing a vehicle into a wall to obtain data. *Destructive testing* destroys a test unit; *nondestructive testing* retains a test unit.

- Many statisticians follow the convention that

$$X_1, X_2, \ldots, X_n$$

  are the data values in the abstract—they are random variables and hence described by uppercase letters. Realizations of these random variables that assume specific numerical values, however, are denoted by

$$x_1, x_2, \ldots, x_n.$$

  This convention will be followed in this text.

- The data values $x_1, x_2, \ldots, x_n$ are known as a "data set."

Once a random sample has been collected, there are two potential next steps. The first is to construct one or more statistical graphics, as introduced in the previous section. The second is to compute one or more *statistics*, which are defined next.

Definition 1.2 Consider the data values $X_1, X_2, \ldots, X_n$. A *statistic* is some function of the data values that does not depend on any unknown parameter(s).

The key to this definition of a statistic is that no unknown parameters are involved. So, for example, the expressions

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \qquad n \cdot X_{(1)} \qquad \frac{X_{(1)} + X_{(n)}}{2} \qquad \prod_{i=1}^{n} X_i$$

are all statistics because they do not involve any unknown parameters. (Recall that the order statistic $X_{(1)}$ is the smallest member of a data set and the order statistic $X_{(n)}$ is the largest member of a data set.) On the other hand, the expressions

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \qquad \theta \cdot X_{(1)} \qquad \frac{\bar{X}}{\mu} \qquad \frac{X_7-\mu}{S/\sqrt{n}}$$

are not statistics because they involve the unknown parameters μ, σ, and θ.

This is an appropriate time to delineate the difference between probability and statistics. Figure 1.18 contains two ovals. The larger oval on the left represents the *population*. As one particular instance, the population might consist of the weight, in pounds, of every person in the world. The smaller oval on the right represents a *sample* of n values taken from the population. Using the current instance, the sample might be the weights of n people sampled at random and without replacement from the population. Populations and samples are fundamentally different entities:

- the population in the left-hand oval is often described by *parameters*, such as the population mean μ and the population variance σ^2, which are fixed constants;
- the sample in the right-hand oval can be described by *statistics*, such as the sample mean $\bar{X}$ and the sample variance S^2, which are random variables that vary from one sample to the next.

The arrow that points to the right represents the application of probability theory. Here is a typical probability problem:

> Ten people crowd into a small elevator. Their weights, in pounds, are mutually independent and identically distributed normal random variables, each with population mean $\mu = 140$ pounds and population standard deviation $\sigma = 30$ pounds. What is the probability that the elevator capacity of 1500 pounds is exceeded?

For this particular probability problem, the information about the population probability distribution, which is $N(140, 900)$, is known. The question asks about a sample statistic $X_1 + X_2 + \cdots + X_{10}$,

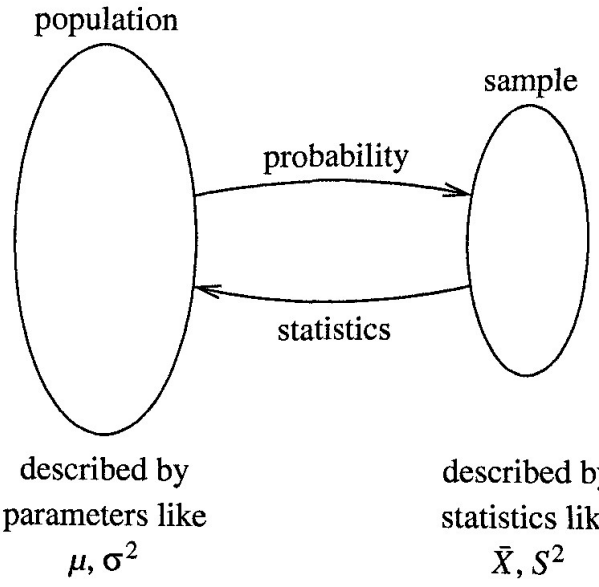


Figure 1.18: The difference between probability and statistics.

where X_i is the weight of the ith person on the elevator. More specifically, the question asks for $P(X_1 + X_2 + \cdots + X_{10} > 1500)$. Referring back to Figure 1.18, knowledge about the probability distribution of the population is used to answer a question concerning a statistic calculated from a sample. The arrow that points to the left represents the application of statistical theory. Here is a typical statistics problem:

> Ten people crowd into a small elevator. Their weights are $x_1 = 220, x_2 = 107, x_3 = 155, \ldots, x_{10} = 129$ pounds. If nothing is known about the population probability distribution of the weights, give estimates of the population mean μ and the population variance σ^2, along with some indication of the precision of the estimates.

In this particular setting, *nothing* is known about the population of weights. Questions are being asked about the population based only on the ten data values. Referring back to Figure 1.18, the data values in the sample are being used to answer questions about the population. This process is often referred to as *statistical inference* because a conclusion is being inferred about the population based on the sample. In some settings, it might be reasonable to assume that the ten values constitute a random sample.

The interest in computing a statistic is oftentimes to gain information about some unknown parameter. For example,

- the sample mean $\bar{X}$ is often used to estimate the population mean μ,
- the sample median M is often used to estimate the population median $x_{0.5}$,
- the sample variance S^2 is often used to estimate the population variance σ^2,
- the sample proportion $\hat{p}$ is often used to estimate the population proportion p.

An important distinction should be made between statistics and the quantities that they are estimating: the statistics (like $\bar{X}$ and S) are random variables, but the parameters that they are estimating (like μ and σ) are constants, which are typically unknown. Statistics take on different values from one sample to the next; population parameters assume just a single value. The sample mean $\bar{X}$ and the sample variance S^2 are formally defined in the next two sections.

Since a statistic is a random variable, its probability distribution is often of interest. The probability distribution of a statistic is called its *sampling distribution*. The following three examples concern a single random experiment—rolling a fair die five times—and highlight the probability distribution of three different statistics that can be gleaned from the five data values. The sample size $n = 5$ is rather small, but the simplicity of this setting allows us to calculate exact sampling distributions of the statistics.

> **Example 1.12** Consider the random variables X_1, X_2, X_3, X_4, X_5, which are the outcomes of five rolls of a fair die. What is the sampling distribution of the statistic $\bar{X}$, the sample mean?
>
> The way that the data is collected (rolling a fair die five times) indicates that the five random variables are mutually independent, so the five values constitute a random sample. The first step in finding the probability distribution of the sample mean $\bar{X}$ is to determine the support (possible values) of the random variable $\bar{X}$. Since the numerator of
>
> $$\bar{X} = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$$

can assume the values $5, 6, \ldots, 30$, the support of $\bar{X}$ is

$$\mathcal{A} = \left\{ x \,\middle|\, x = 1, \frac{6}{5}, \frac{7}{5}, \ldots, \frac{29}{5}, 6 \right\}.$$

The next step is to determine the probabilities associated with each element in the support. Begin with $\bar{X} = 1$. There is only one way to achieve a sample mean of 1, which is the outcome $(X_1, X_2, X_3, X_4, X_5) = (1, 1, 1, 1, 1)$, so

$$P(\bar{X} = 1) = \frac{1}{6^5}.$$

Now consider $\bar{X} = 6/5$. This value for $\bar{X}$ can only be achieved with 4 ones and a single 2, for example the outcome $(X_1, X_2, X_3, X_4, X_5) = (1, 2, 1, 1, 1)$. Since the 2 can occur on any one of the five rolls,

$$P\left(\bar{X} = \frac{6}{5}\right) = \frac{5}{6^5}.$$

Next consider $\bar{X} = 7/5$. There are two ways to achieve this sample mean: 4 ones and a single 3, or 3 ones and 2 twos. So the probability that $\bar{X} = 7/5$ is

$$P\left(\bar{X} = \frac{7}{5}\right) = \frac{5}{6^5} + \frac{10}{6^5} = \frac{15}{6^5}$$

because there are $\binom{5}{2} = 10$ ways to place the 2 twos in the sequence of five rolls. One way to proceed is to continue in this fashion, which is a mind-numbing exercise. A much more efficient way to proceed is to use the following APPL code to calculate the probability mass function of $\bar{X}$.

```
X := UniformDiscreteRV(1, 6);
Y := ConvolutionIID(X, 5);
g := [[x -> x / 5], [5, 30]];
Xbar := Transform(Y, g);
```

This code returns the symmetric probability mass function for $\bar{X}$ as

$$f_{\bar{X}}(x) = \begin{cases} 1/7776 & x = 1 \\ 5/7776 & x = 6/5 \\ 15/7776 & x = 7/5 \\ 35/7776 & x = 8/5 \\ \vdots & \vdots \\ 1/7776 & x = 6. \end{cases}$$

The probability mass function is plotted in Figure 1.19. Even though the sample size of $n = 5$ is quite small, the effects of the central limit theorem are already being seen in the somewhat bell-shaped probability mass function.

In a statistical setting, you typically get only one instance of the statistic $\bar{X}$. For example, if one rolls a fair die five times, they might get

$$x_1 = 2, x_2 = 6, x_3 = 1, x_4 = 3, x_5 = 2,$$

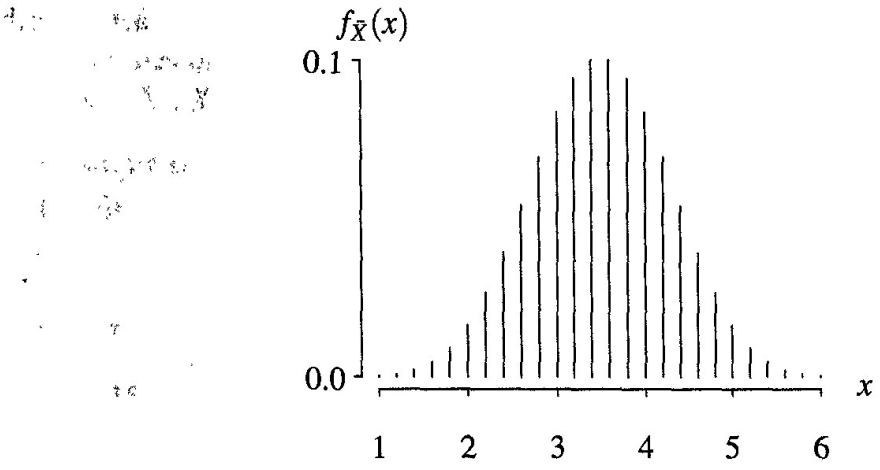


Figure 1.19: Sampling distribution of the sample mean.

which corresponds to $\bar{x} = 14/5 = 2.8$. Knowing where $\bar{x} = 2.8$ falls in the sampling distribution of $\bar{X}$ can be helpful in drawing conclusions concerning these particular rolls of the fair dice.

Here is one particular instance. Let's say a dice manufacturer claims that their dice are fair. Your friend, on the other hand, claims that the dice are loaded and producing too many sixes. As a budding statistician, you decide to purchase a die and roll it five times. If the results of your random experiment are

$$x_1 = 6, x_2 = 6, x_3 = 6, x_4 = 6, x_5 = 6,$$

you would certainly side with your friend. But could the die indeed have been fair as the manufacturer claimed? Possibly, but your work on determining the sampling distribution of $\bar{X}$ under the assumption that the manufacturer is telling the truth indicates that the outcome that you achieved occurs only one time in 7776, making the manufacturer's claim seem rather dubious. We may reject the "null" hypothesis that the die is fair because the likelihood of observing all sixes is extremely small if the die were indeed fair.

As a thought experiment, what would be our conclusion if we rolled all fours? It is for situations like this in which different statistics become valuable tools to support or refute various types of hypotheses.

The next example considers that same random experiment, rolling a fair die five times, but this time uses a different test statistic.

Example 1.13 Consider again the random variables X_1, X_2, X_3, X_4, X_5, which are the outcomes of five rolls of a fair die. What is the sampling distribution of the statistic $X_{(5)} = \max\{X_1, X_2, \ldots, X_5\}$?

The fifth order statistic satisfies the definition of a statistic because it is a function of the data alone and does not involve any unknown parameters. As before, the first step in determining the sampling distribution of the statistic is to determine its support. The largest of five rolls of a fair die has support

$$\mathcal{A} = \{x | x = 1, 2, 3, 4, 5, 6\}.$$

The next step is to assign probabilities to each of the six values in the support. The only way to obtain a maximum of one is to roll 5 ones, so

$$P\left(X_{(5)}=1\right)=P(X_1=X_2=X_3=X_4=X_5=1)=\frac{1}{6^5}.$$

There are multiple ways for the largest value rolled to be a two. One way is to roll all twos (and there is only one way to do so) and the other ways are various sequences of ones and twos. Using combinations to count all of the possibilities associated with the largest outcome being two gives

$$P\left(X_{(5)}=2\right)=\frac{\binom{5}{5}+\binom{5}{4}+\binom{5}{3}+\binom{5}{2}+\binom{5}{1}}{6^5}=\frac{1+5+10+10+5}{6^5}=\frac{31}{7776}.$$

One can continue in this fashion or use the APPL code given below to calculate the probability mass function of $X_{(5)}$.

```
X := UniformDiscreteRV(1, 6);
Y := MaximumIID(X, 5);
```

This code returns the probability mass function for $X_{(5)}$ as

$$f_{X_{(5)}}(x)=\begin{cases}1/7776 & x=1\\ 31/7776 & x=2\\ 211/7776 & x=3\\ 781/7776 & x=4\\ 2101/7776 & x=5\\ 4651/7776 & x=6.\end{cases}$$

This probability mass function is plotted in Figure 1.20. Not surprisingly, the most likely maximum is $X_{(5)}=6$.

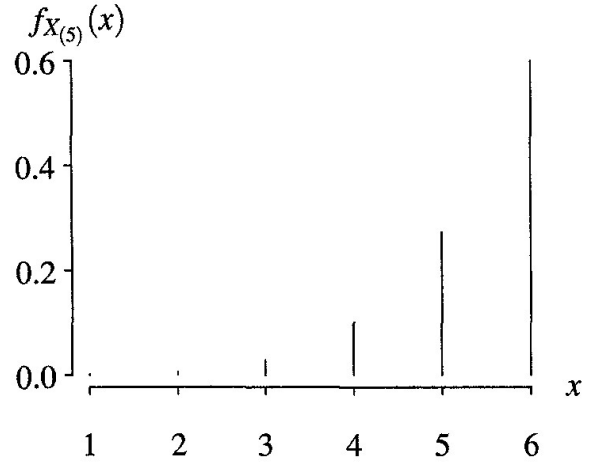


Figure 1.20: Sampling distribution of the sample maximum.

The previous two examples have illustrated two different statistics, the sample mean and the sample maximum, associated with the same random experiment. The next example introduces a new, and somewhat more obscure statistic, the *sample range*, which is used in a field known as *statistical quality control.*

Example 1.14 Consider once again the random variables X_1, X_2, X_3, X_4, X_5, which are the outcomes of five rolls of a fair die. What is the sampling distribution of the statistic $R = \max\{X_1, X_2, \ldots, X_5\} - \min\{X_1, X_2, \ldots, X_5\}$?

The sample range $R = X_{(5)} - X_{(1)}$ satisfies the definition of a statistic given in Definition 1.2 because it is a function of the data only and does not involve any unknown parameters. As before, the first step in determining the sampling distribution of the statistic is to determine its support. The difference between the largest outcome of the five rolls and the smallest outcome of the five rolls has support

$$\mathcal{A} = \{x | x = 0, 1, 2, 3, 4, 5\}.$$

The next step is to assign probabilities to each of the six values in the support. The only way to obtain a sample range of $R = 0$ is to roll five identical values, so

$$P(R=0) = P(X_1 = X_2 = X_3 = X_4 = X_5) = \frac{6}{6^5}.$$

There are multiple ways to obtain a sample range of $R = 1$. Examples include the outcomes $(X_1, X_2, X_3, X_4, X_5) = (1, 1, 1, 2, 1)$ and $(X_1, X_2, X_3, X_4, X_5) = (4, 5, 4, 5, 4)$. Using combinations to count all of the possibilities associated with a sample range of $R = 1$, we obtain

$$P(R=1) = 5 \cdot \frac{\binom{5}{4} + \binom{5}{3} + \binom{5}{2} + \binom{5}{1}}{6^5} = \frac{150}{7776}.$$

One can continue in this fashion or use the APPL code given below to calculate the probability mass function of R.

```
X := UniformDiscreteRV(1, 6);
R := RangeStat(X, 5);
```

This code returns the probability mass function for R as

$$f_R(x) = \begin{cases} 6/7776 & x = 0 \\ 150/7776 & x = 1 \\ 720/7776 & x = 2 \\ 1710/7776 & x = 3 \\ 2640/7776 & x = 4 \\ 2550/7776 & x = 5. \end{cases}$$

Some might prefer the notation $f_R(r)$ for this probability mass function, but we use $f_R(x)$ to have a consistent index x with the previous two examples. This probability mass function is plotted in Figure 1.21.

The previous three examples have illustrated three statistics and their associated sampling distributions. One key insight here is that when a random experiment is conducted, we get to calculate just one instance of the test statistic. The importance of the sampling distribution is to let you know whether the value of the statistic that you observe is common or rare. For example, if you roll a fair die five times and get all ones, then the three statistics give

$$\bar{x} = 1 \qquad x_{(5)} = 1 \qquad r = 0.$$

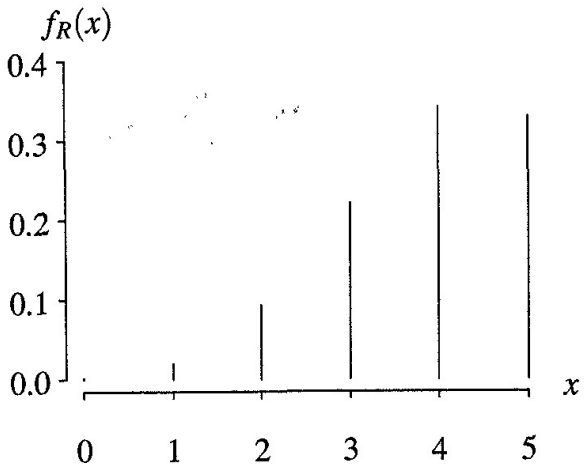


Figure 1.21: Sampling distribution of the sample range.

Looking at the three sampling distributions from the previous three examples, the values of these particular three statistics correspond to very unlikely events because the sampling distributions tell us

$$P(\bar{X}=1)=\frac{1}{7776} \qquad P(X_{(5)}=1)=\frac{1}{7776} \qquad P(R=0)=\frac{6}{7776}.$$

Different statistics are used to detect different types of "rare" events. For example, if you roll a fair die five times and get all sixes, then the three statistics are

$$\bar{x}=6 \qquad x_{(5)}=6 \qquad r=0.$$

Looking at the three sampling distributions from the previous three examples, $\bar{x}=6$ and $r=0$ are extraordinarily unlikely events, but $x_{(5)}=6$ occurs quite often. The sampling distributions tell us that

$$P(\bar{X}=6)=\frac{1}{7776} \qquad P(X_{(5)}=6)=\frac{4651}{7776} \qquad P(R=0)=\frac{6}{7776}.$$

The take-away message here is that certain types of statistics can be selected to detect one particular type of rarity over another. Finally, to consider a sequence of rolls that is a bit more mainstream, the rolls $(x_1,x_2,x_3,x_4,x_5)=(1,4,5,2,4)$ result in the three statistics

$$\bar{x}=3.2 \qquad x_{(5)}=5 \qquad r=4.$$

Looking at the graphs of the three sampling distributions from the previous three examples, none of these statistics point to this particular outcome as particularly rare because

$$P(\bar{X}=3.2)=\frac{735}{7776}\cong 0.09 \qquad P(X_{(5)}=5)=\frac{2101}{7776}\cong 0.27 \qquad P(R=4)=\frac{2640}{7776}\cong 0.34.$$

The next two sections introduce two broad classes of statistics that arise in many statistical problems. The first class consists of statistics that reflect the central tendency of the population probability distribution. The second class consists of statistics that reflect the dispersion of the population probability distribution. Having estimates of the central tendency and the dispersion is important because they allow a statistician to quantify both the center of the population probability distribution and how far from that center one can expect random variables to stray.

1.3 Estimating Central Tendency

As indicated in the previous section, statistics can be defined to estimate certain characteristics of a population distribution. One aspect of a population that is nearly always of interest is the central tendency of the population distribution. Several statistics that reflect this central tendency are formally defined in this section. We begin with the sample mean.

Sample mean

The *sample mean* is the most intuitive measure of central tendency. People naturally average data values in order to get a sense of the center of a probability distribution.

Definition 1.3 Let $x_1, x_2, \ldots, x_n$ be experimental values associated with the random variables $X_1, X_2, \ldots, X_n$. The *sample mean* is

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

As indicated previously, $\bar{X}$ is used in the abstract when there are no specific data values. When specific data values have been collected, the lower case version $\bar{x}$ is used to denote the sample mean. The sample mean is sometimes called the *sample arithmetic mean.*

Example 1.15 Ten kindergarten children from ten different families are polled to find the number of children that are in their family. The resulting values, $x_1, x_2, \ldots, x_{10}$ are

$$3, 1, 5, 1, 3, 2, 1, 1, 3, 2.$$

Calculate the sample mean.

The sample mean is

$$\bar{x} = \frac{3+1+5+1+3+2+1+1+3+2}{10} = \frac{22}{10} = 2.2 \text{ children.}$$

This calculation is straightforward and can be conducted in R with the statements given below.

```
x = c(3, 1, 5, 1, 3, 2, 1, 1, 3, 2)
mean(x)
```

Of course, polling ten different children would most likely result in a different sample mean. The experimental sample mean $\bar{x}$ given above is one instance from the sampling distribution of the random variable $\bar{X}$.

Another way of thinking about a sample mean is to consider it to be a special case of a *weighted average*, in which each of the data values is given a weight of $1/n$. If the data values constitute a random sample, then there is no reason to give more weight to one value over another. Returning to the kindergarten sibling data from the previous example, the sample mean could be written as

$$\bar{x} = \frac{3+1+5+1+3+2+1+1+3+2}{10} = 1\cdot\frac{4}{10} + 2\cdot\frac{2}{10} + 3\cdot\frac{3}{10} + 5\cdot\frac{1}{10}.$$

This way of thinking emphasizes the fact that the sample mean is a weighted average, where the weights reflect the relative frequency of a particular data value. Compare the expression on the

far right with the formula for the population mean $E[X]$ for a discrete probability distribution from probability theory:

$$E[X] = \sum_{\mathcal{A}} x f(x),$$

where $\mathcal{A}$ is the support and $f(x)$ is the probability mass function. The weights $4/10$, $2/10$, $3/10$, and $1/10$ play the role of $f(x)$ from probability theory.

There is still another way to think about the sample mean. In order to develop this formulation, the notion of an empirical probability distribution must be defined.

Definition 1.4 Let $x_1, x_2, \ldots, x_n$ be experimental values associated with the random variables $X_1, X_2, \ldots, X_n$. The *empirical probability distribution* associated with $x_1, x_2, \ldots, x_n$ is the discrete probability distribution defined by assigning probability $1/n$ to each x_i value.

This empirical probability distribution can be expressed as either an *empirical probability mass function*, denoted by $\hat{f}(x)$, or an *empirical cumulative distribution function*, denoted by $\hat{F}(x)$, which are defined next. The empirical cumulative distribution function was introduced in the statistical graphics section as a way to avoid binning observations into cells when constructing a histogram.

Definition 1.5 Let $x_1, x_2, \ldots, x_n$ be experimental values associated with the random variables $X_1, X_2, \ldots, X_n$. The *empirical probability mass function* associated with $x_1, x_2, \ldots, x_n$ is

$$\hat{f}(x) = \frac{\text{number of } x_i \text{ equal to } x}{n}.$$

The *empirical cumulative distribution function* associated with $x_1, x_2, \ldots, x_n$ is

$$\hat{F}(x) = \frac{\text{number of } x_i \text{ less than or equal to } x}{n}.$$

The empirical probability distribution, regardless of whether it is expressed in either of its equivalent forms as $\hat{f}(x)$ or $\hat{F}(x)$, is our best guess for the population probability distribution based on the data values $x_1, x_2, \ldots, x_n$.

Let's return to the discussion of the sample mean. The empirical probability distribution associated with the data set has a population mean, which is typically called the "plug-in estimator of the population mean." Using the formula for the population mean from probability, the formula for the plug-in estimator of the population mean is

$$\hat{\mu} = \sum_{\mathcal{A}} x \hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

where $\mathcal{A}$ is the support of the population distribution. This is, once again, the formula for the sample mean.

So, regardless of whether you simply use the defining formula, think of the sample mean as a weighted average, or use the plug-in estimator of the population mean, the same value results for the sample mean.

Since the sample mean $\bar{X}$ is a random variable, we can calculate its sampling distribution. This sampling distribution depends on the population probability distribution from which the data values are drawn. The two examples that follow consider the sampling distribution of the sample mean for observations drawn from a discrete population and a continuous population.

Example 1.16 Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson(λ) distribution, where λ is a positive unknown parameter. What is the sampling distribution of $\bar{X}$?

One could easily envision a real-world scenario in which averaging observations sampled from a Poisson population could occur, for example,

- averaging the number of customers that arrive to a drive-up window at a fast food restaurant during the lunch hour for five consecutive weekdays,
- averaging the number of potholes per mile on a particular stretch of highway, and
- averaging the number of web hits per day at a popular website during February.

The first step in finding the probability distribution of the sample mean $\bar{X}$ is to determine the support of the random variable $\bar{X}$. The Poisson population distribution has support on the nonnegative integers, so the numerator of

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

can also assume the values $0, 1, 2, \ldots$. Therefore, the support of $\bar{X}$ is

$$\mathcal{A} = \left\{ x \,\middle|\, x = 0, \frac{1}{n}, \frac{2}{n}, \ldots \right\}.$$

The next step is to determine the probabilities associated with each element in the support. The random variable X_i has probability mass function

$$f_{X_i}(x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad x = 0, 1, 2, \ldots$$

for $i = 1, 2, \ldots, n$. The numerator in the sample mean

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

consists of the sum of mutually independent and identically distributed Poisson(λ) random variables because $X_1, X_2, \ldots, X_n$ is a random sample. Using a result from probability theory that can be proved by the moment generating function technique, the numerator $X_1 + X_2 + \cdots + X_n$ is Poisson($n\lambda$) with probability mass function

$$f_{X_1+X_2+\cdots+X_n}(x) = \frac{(n\lambda)^x e^{-n\lambda}}{x!} \qquad x = 0, 1, 2, \ldots.$$

Finally, dividing the numerator of $\bar{X}$ by n gives the probability mass function

$$f_{\bar{X}}(x) = \frac{(n\lambda)^{nx} e^{-n\lambda}}{(nx)!} \qquad x = 0, \frac{1}{n}, \frac{2}{n}, \ldots$$

by the transformation technique.

Now consider the sampling distribution of $\bar{X}$ for a particular sample size n and a particular population mean λ, say $n = 5$ and $\lambda = 2$. Figure 1.22 is a graph of the probability mass function of the first seven support values of the population from which the data values are drawn, that is, a Poisson(2) distribution. The graph of $f_X(x)$ continues to

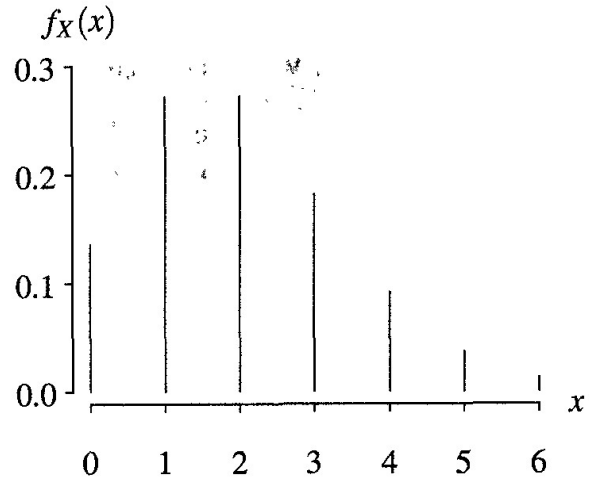


Figure 1.22: Population probability mass function.

decline as x increases. Using the formula for $f_{\bar{X}}(x)$, the probability mass function of $\bar{X}$ is

$$f_{\bar{X}}(x) = \frac{10^{5x}e^{-10}}{(5x)!} \qquad x = 0, \frac{1}{5}, \frac{2}{5}, \ldots .$$

Figure 1.23 is a graph of the probability mass function of the sampling distribution of the statistic $\bar{X}$ when $n = 5$ and $\lambda = 2$. The horizontal scales are identical, but the vertical scales differ on the two graphs. There are four observations that can be made concerning these two probability mass functions:

- Both the probability distribution of X_i and the probability distribution of $\bar{X}$ have the same expected value: $E[X_i] = E[\bar{X}] = 2$ in this particular example. As will be seen subsequently, this result, stated more generally as $E[\bar{X}] = \mu$, is true for any population distribution that has a finite population mean.
- The population variance of the sampling distribution of $\bar{X}$ is less than the population variance of the population distribution. Although X_i and $\bar{X}$ have the same

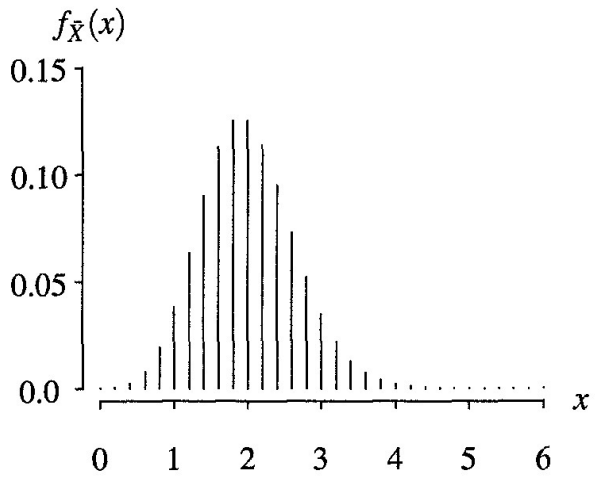


Figure 1.23: Probability mass function (sampling distribution) of the sample mean.

population mean, averaging the $n = 5$ observations decreases the population variance of $\bar{X}$ relative to X_i.

- The support of $\bar{X}$ is finer than the support of X_i. The data values can assume the values $0, 1, 2, \ldots$, but the sample mean can assume the values $0, 1/5, 2/5, \ldots$.
- Even though the sample size is only $n = 5$, the central limit theorem is evident in the distribution of $\bar{X}$ as it has more of a bell shape than the population distribution. The limiting distribution of $\bar{X}$ in this example is normal.

The sampling distribution of $\bar{X}$ when the data values are drawn from a continuous population is determined in a similar fashion, as illustrated in the next example.

Example 1.17 Let $X_1, X_2, \ldots, X_n$ be a random sample from a gamma(λ, κ) distribution, where λ and κ are positive unknown scale and shape parameters. Find the sampling distribution of $\bar{X}$. Also find $P(\bar{X} < 2)$ for a sample size of $n = 4$ when $\lambda = 1$ and $\kappa = 3$.

The probability density function of X_i sampled from a gamma(λ, κ) population is

$$f_{X_i}(x) = \frac{\lambda^\kappa x^{\kappa-1} e^{-\lambda x}}{\Gamma(\kappa)} \qquad x > 0$$

for $i = 1, 2, \ldots, n$. The corresponding moment generating function is

$$M_{X_i}(t) = \left(\frac{\lambda}{\lambda - t}\right)^\kappa \qquad t < \lambda$$

for $i = 1, 2, \ldots, n$. Since the observations are a random sample, $X_1, X_2, \ldots, X_n$ are mutually independent and identically distributed random variables. Hence, the moment generating function of $\bar{X}$ is

$$\begin{aligned}
M_{\bar{X}}(t) &= E\left[e^{t\bar{X}}\right] \\
&= E\left[e^{t(X_1+X_2+\cdots+X_n)/n}\right] \\
&= M_{X_1+X_2+\cdots+X_n}(t/n) \\
&= M_{X_1}(t/n) M_{X_2}(t/n) \ldots M_{X_n}(t/n) \\
&= \left(\frac{\lambda}{\lambda - t/n}\right)^\kappa \left(\frac{\lambda}{\lambda - t/n}\right)^\kappa \cdots \left(\frac{\lambda}{\lambda - t/n}\right)^\kappa \\
&= \left(\frac{n\lambda}{n\lambda - t}\right)^{n\kappa} \qquad t < n\lambda.
\end{aligned}$$

This moment generating function can be recognized as that of a gamma($n\lambda$, $n\kappa$) random variable.

Figure 1.24 is a plot of the population probability density function for $\lambda = 1$ and $\kappa = 3$. The $n = 4$ data values are sampled from this population probability distribution. Since $\bar{X} \sim$ gamma($n\lambda$, $n\kappa$), $\lambda = 1$, $\kappa = 3$, and $n = 4$,

$$\bar{X} \sim \text{gamma}(4, 12).$$

Figure 1.25 contains a plot of the probability density function of the sample mean $\bar{X} \sim$ gamma(4, 12). The horizontal scales in Figures 1.24 and 1.25 are identical, but the vertical scales differ on the two graphs. The same effect as in the previous example (when the random sampling was from a Poisson population) takes place here:

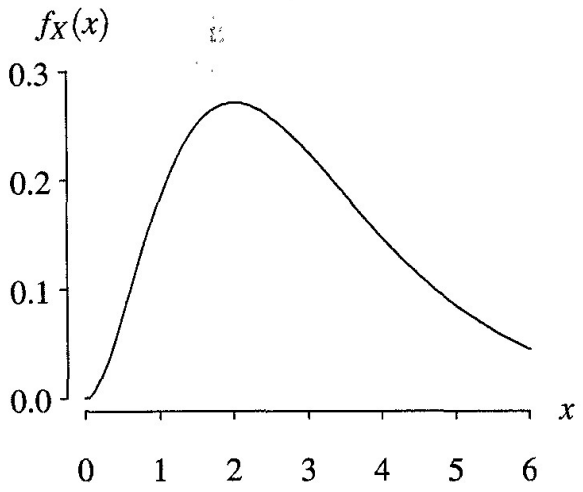


Figure 1.24: Population probability density function.

- The population probability distribution and the probability distribution of $\bar{X}$ have the same central value, which in this case is $E[X_i] = E[\bar{X}] = 3$.
- The probability distribution of $\bar{X}$ has a smaller population variance than the population probability distribution.
- The probability distribution of $\bar{X}$ looks more bell-shaped than the population probability distribution because of the central limit theorem. The probability density function of $\bar{X}$ is nearly symmetric. The limiting distribution of $\bar{X}$ is normal.

The final part of the question is to determine the probability that the sample mean is less than 2 for sample size $n = 4$ and population parameters $\lambda = 1$ and $\kappa = 3$. One way to calculate this probability is to integrate the probability density function over the appropriate limits. Since $\bar{X} \sim \text{gamma}(4, 12)$,

$$P(\bar{X} < 2) = \int_0^2 \frac{(n\lambda)^{n\kappa} x^{n\kappa-1} e^{-n\lambda x}}{\Gamma(n\kappa)} dx = \int_0^2 \frac{4^{12} x^{11} e^{-4x}}{\Gamma(12)} dx.$$

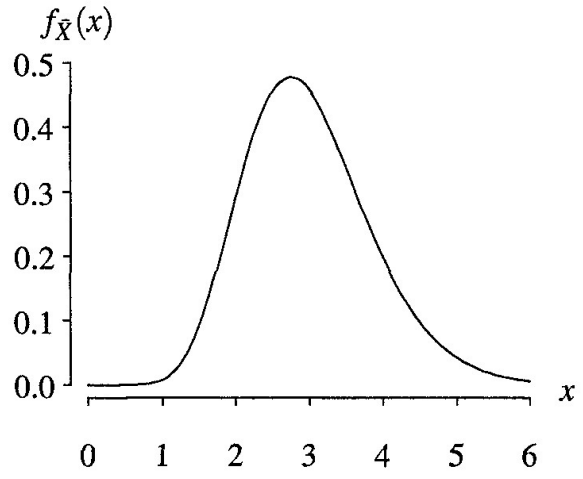


Figure 1.25: Probability density function (sampling distribution) of the sample mean.

This integral can be computed by hand using integration by parts repeatedly or can be calculated using a computer algebra system, giving the required probability as

$$P(\bar{X} < 2) = 1 - \frac{412782941}{155925} e^{-8} \cong 0.1119.$$

R can also be used to calculate the probability that the sample mean is less than 2. Using the `pgamma` function, which returns the cumulative distribution function of a gamma random variable, the single statement

```
pgamma(2, 12, 4)
```

also returns $P(\bar{X} < 2) \cong 0.1119$. Notice that R switches the order of the parameters as arguments relative to the convention gamma(λ, κ) used here.

Finally, to determine whether the derivation and associated numerical value are correct, a Monte Carlo simulation experiment can be conducted to estimate the probability that the sample mean is less than 2. The following R code generates one million sample means and prints the fraction of those sample means that are less than 2.

```
nrep = 1000000
count = 0
for (i in 1:nrep) {
  xbar = mean(rgamma(4, 3, 1))
  if (xbar < 2) count = count + 1
}
print(count / nrep)
```

After a call to `set.seed(3)` to initialize the random number stream, five runs of this simulation yield the following estimates of $P(\bar{X} < 2)$:

0.1118 0.1119 0.1120 0.1117 0.1125.

Since these values hover about the analytic value $P(\bar{X} < 2) \cong 0.1119$, the Monte Carlo simulation supports our analytic solution.

The two previous examples have shown that the probability distribution of the sample mean depends on the probability distribution associated with the population. Every population probability distribution that the data values are drawn from requires a separate derivation—some simple and others quite intricate—to determine the probability distribution of $\bar{X}$. One piece of good news is that the expected value of $\bar{X}$ and the population variance of $\bar{X}$ can be computed with the same formulas for practically all probability distributions. Assuming that $X_1, X_2, \ldots, X_n$ constitute a random sample from some population distribution (discrete or continuous) with finite population mean μ and finite population variance σ^2, then the sample mean $\bar{X}$ has expected value

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}(n\mu) = \mu$$

and population variance

$$V[\bar{X}] = V\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V[X_i] = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}.$$

The first of these equations indicates that the sample mean $\bar{X}$ is on target for estimating the population mean μ. So the first equation addresses the *accuracy* of $\bar{X}$ in estimating μ. Since the expected value of the sample mean is the population mean, statisticians say that $\bar{X}$ is an *unbiased estimator* of μ. (Unbiased estimators will presented formally in Chapter 2.) The second of these equations indicates that the variability of the sample mean $\bar{X}$ decreases as n increases. So the second equation addresses the *precision* of $\bar{X}$ in estimating μ. The sample mean $\bar{X}$ is often said to be a more *precise* estimator of μ as the sample size increases. This constitutes a derivation of the following result.

Theorem 1.1 Let $X_1, X_2, \ldots, X_n$ be a random sample from a population distribution with finite population mean μ and finite population variance σ^2. The sample mean $\bar{X}$ has population mean

$$E[\bar{X}] = \mu$$

and population variance

$$V[\bar{X}] = \frac{\sigma^2}{n}.$$

Having a good understanding of the behavior of $\bar{X}$ is important when drawing conclusions based on sample means, as illustrated in the next example.

Example 1.18 The six states with the highest age-adjusted incidence of kidney cancer in the United States in the years 2012–2016 are shown in Figure 1.26 using data from the Centers for Disease Control website. If you happen to be reading this book and live in one of those six states, you might be grabbing your belly right now and thinking "Oh no, I am living in a kill zone, I need to move away!" But where should you move? Figure 1.27 shows the six states with the lowest age-adjusted incidence of kidney cancer in the United States in the years 2012–2016. It seems like any one of these states would provide a much more hospitable home for your kidneys.

It is important to establish that the kidney cancer incidence rate is indeed a sample mean $\bar{X}$. Think of the kidney cancer status of each resident of a state as a Bernoulli ran-

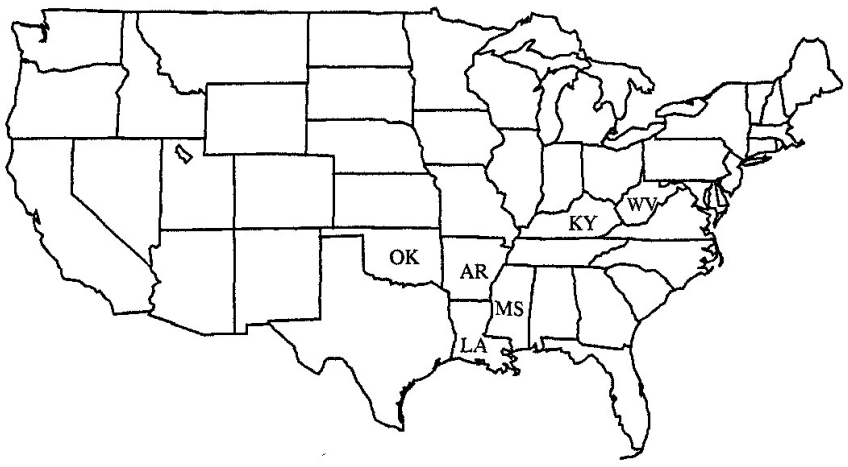


Figure 1.26: Six states with the highest incidence of kidney cancer in 2012–2016.

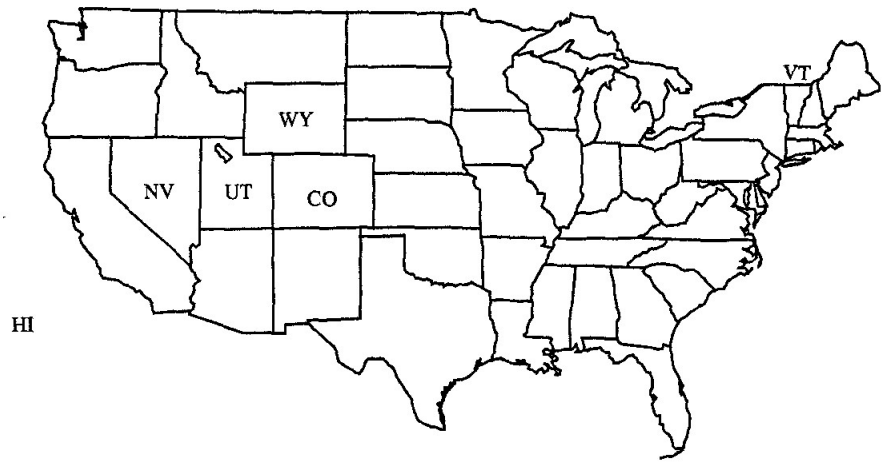


Figure 1.27: Six states with the lowest incidence of kidney cancer in 2012–2016.

dom variable X_i, where $X_i = 0$ corresponds to not being diagnosed with kidney cancer in the years 2012–2016 and $X_i = 1$ corresponds to being diagnosed with kidney cancer in the years 2012–2016, for $i = 1, 2, \ldots, n$, where n is the population of the state. The sample mean $\bar{X}$ gives the estimated probability or incidence rate of kidney cancer for a particular state. Epidemiologists typically express the incidence rate for a rare cancer in terms of the number of cases per 100,000 in order to avoid writing too many leading zeros. This change of scale does not change the fact that the incidence rate behaves like an average from a statistical point of view.

So what is going on here? Are the states with the high kidney cancer incidence rates really less safe, or are we being tricked by random sampling variability? Another statistical graphic can lend some insight. The kidney cancer annual incidence rates for the states are plotted on the vertical axis against the population on the horizontal axis (which uses a logarithmic scale) in Figure 1.28. The weighted annual incidence rate

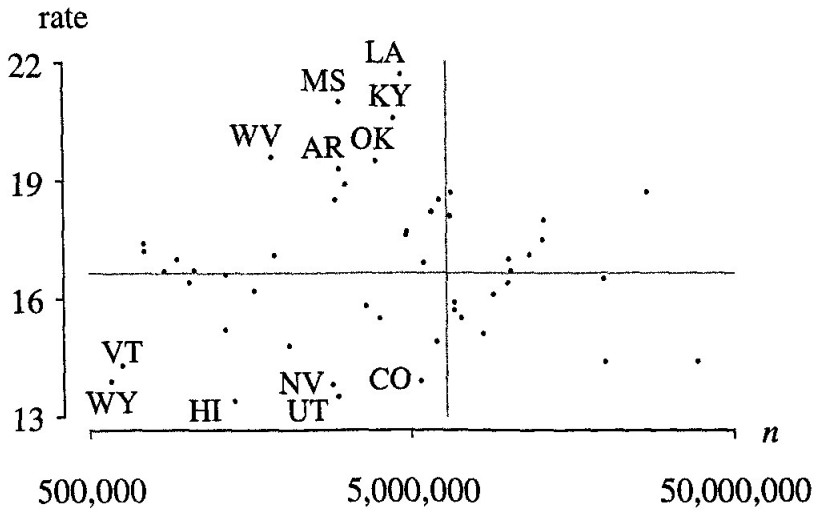


Figure 1.28: Population versus annual kidney incidence rate per 100,000 in 2012–2016.

for the entire U.S. over the five-year period is 16.7 incidences per 100,000 population, indicated by the horizontal line on the plot. The average population of a state over the five-year period is 6.4 million residents, indicated by the vertical line on the plot. The six states with the highest and lowest kidney cancer incidence rates in 2012–2106 are identified on the plot. One curiosity that appears immediately is that the states with high and low incidence rates tend to, on average, be smaller states. The most populous states don't show up on either list.

So the fact that smaller states tend to appear more often on the list of states with high and low kidney cancer incidence rates brings us back to the equation $V[\bar{X}] = \sigma^2/n$. Smaller states have a smaller value of n and are thus more susceptible to random sampling variability of $\bar{X}$ and are thus more likely to show up on the high incidence rate and low incidence rate lists. So it is quite possible that the six states with the highest and lowest kidney cancer incidence rates are really no more or less risky than any others. Looking at the data for subsequent years would help confirm whether or not there is a pattern developing, or if the results are simply due to random sampling variability of $\bar{X}$.

Professor Howard Wainer refers to $V[\bar{X}] = \sigma^2/n$ as a "dangerous equation" in his book *Picturing the Uncertain World.* He also considers kidney cancer incidence rates, but this time by county rather than state. Some counties have just a few hundred residents, so having no kidney cancers gives them a kidney cancer incidence rate of zero, which could potentially wrongly classify them as "safe." Having just a single kidney cancer in a small county, however, could potentially wrongly classify them as "unsafe." A large county with millions of residents will almost never be classified as safe or unsafe because of the n in the denominator of $V[\bar{X}] = \sigma^2/n$. He cites several other examples where the lack of knowledge concerning the effect of n on the population variance of $\bar{X}$ "has led to billions of dollars of loss over centuries, yielding untold hardship."

To visualize the effect of n on the sampling distribution of $\bar{X}$ in a Monte Carlo framework, consider a random sample of n data values drawn from a $N(10, 1)$ population. Ten values of $\bar{x}$ are plotted for $n = 2, 4, 8, 16, 32, 64$ in Figure 1.29, which has a logarithmic horizontal axis. The R code to generate and plot the sample means follows. The `plot` function sets up the axes, and the `points` function plots the points generated by

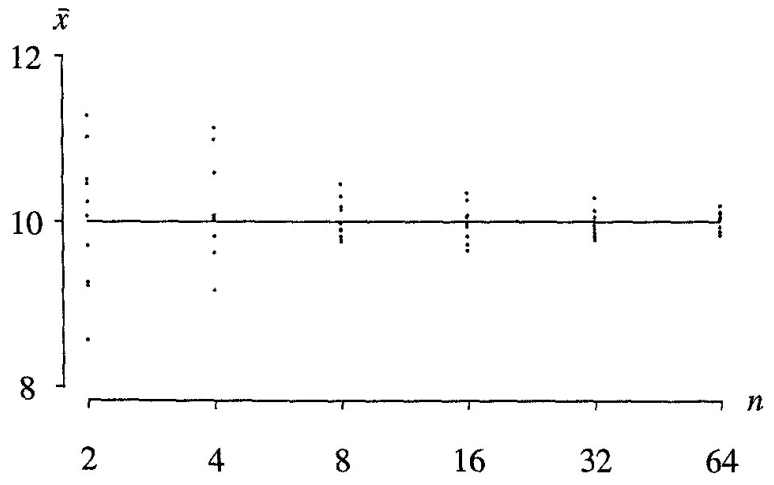


Figure 1.29: Monte Carlo experiment plotting $\bar{x}$ for several values of n.

`rnorm`, which generates random variates from a normal population. The two take-away messages are immediate: (a) since $\bar{X}$ is an unbiased estimate of μ, all averages have an expected value of $\mu = 10$, and (b) extreme values of the sample mean will occur at the smaller sample sizes because the sample mean has a larger population variance for smaller values of n.

```
set.seed(8)
plot(c(2, 64), c(10, 10), type = "l", xlim = c(2, 64),
                                      ylim = c(8, 12), log = "x")
nrep = 10
for (n in c(2, 4, 8, 16, 32, 64)) {
  for (j in 1:nrep) {
    xbar = mean(rnorm(n, 10, 1))
    points(n, xbar)
  }
}
```

The previous example concerned cancer incidence rates, but could apply equally well to any number of settings. Here are three examples.

- If a small high school has stellar average SAT scores, it could be that the high school is particularly good or it could be that the high school is particularly small.
- If a small hospital has an unusually high infection rate for patients, it could be that the hospital is careless with respect to sanitation or it could be that the hospital is particularly small.
- If your friend invests in just two stocks and brags about his average annual return, it could be that he is a brilliant investor or it could be that the number of stocks he invested in is particularly small.

The key take-away point of the previous example involving kidney cancer rates, stated in two equivalent fashions, is

- the sample mean $\bar{X}$ is a more precise estimator of μ for larger sample sizes
- small sample sizes can yield more extreme values of $\bar{X}$ than large sample sizes.

One last point should be made about the equation $V[\bar{X}] = \sigma^2/n$ from Theorem 1.1. Taking the positive square root of both sides of this equation results in

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}},$$

which can be read as "the standard error of the sample mean is the ratio of the population standard deviation to the square root of n." The term "standard error" here is synonymous with standard deviation. This equation implies that if you want to halve the standard deviation of $\bar{X}$, you must quadruple the sample size n. Even more extreme, if you want to decrease the standard deviation of $\bar{X}$ by a factor of 10, you must collect 100 times as many data values. The $\sqrt{n}$ will appear in the denominator of many expressions throughout this book and it will cause problems when lots of precision is required and data values are expensive to collect. This relationship between n and $\sigma_{\bar{X}}$ is an instance of the *law of diminishing returns*.

As a final example to conclude this subsection on the sample mean, consider the effect of sampling *dependent* observations. The mutual independence assumption from the previous examples, which was helpful in determining the sampling distribution of the sample mean, is lost because we are no longer dealing with a random sample. The formula for $\bar{X}$ remains the same, but the calculation of the probability distribution of $\bar{X}$ is more difficult. The next example considers the case of $n = 2$ dependent observations.

Example 1.19 Consider an experiment that consists of sampling a person at random from the community and asking them the following two (somewhat personal) questions:

1. Are you a medical doctor?
2. Is your annual salary greater than \$100,000?

The responses to these $n = 2$ questions are positively correlated because medical doctors tend to have higher salaries than the general population. Let X_1 be 0 for "no" and 1 for "yes" to the first question. Likewise, let X_2 be 0 for "no" and 1 for "yes" to the second question. The responses to the questions have now been defined as the dependent random variables X_1 and X_2. The sample mean is

$$\bar{X} = \frac{X_1 + X_2}{2}.$$

Table 1.9 contains the joint probability mass function of the random variables X_1 and X_2, where p_1, p_2, p_3, and p_4 are unknown probabilities that sum to one. What is the probability mass function of $\bar{X}$?

x_1 \ x_2	0	1
0	p_1	p_2
1	p_3	p_4

Table 1.9: Joint probability mass function for X_1 and X_2.

The dependence between X_1 and X_2 indicates that the probabilities for each of the possible values for $\bar{X}$ needs to be assigned to the appropriate component probabilities. Since the sample size of $n = 2$ is so small, there are only three different values for $\bar{X}$:

- $\bar{X} = 0$, which corresponds to $X_1 = 0$ and $X_2 = 0$,
- $\bar{X} = 1/2$, which corresponds to $X_1 = 0$ and $X_2 = 1$, or $X_1 = 1$ and $X_2 = 0$,
- $\bar{X} = 1$, which corresponds to $X_1 = 1$ and $X_2 = 1$.

So the sampling distribution of $\bar{X}$ in this case is described by the probability mass function

$$f_{\bar{X}}(x) = \begin{cases} p_1 & x = 0 \\ p_2 + p_3 & x = 1/2 \\ p_4 & x = 1. \end{cases}$$

The sample mean is the most common statistical measure of central tendency. The sample median is considered next.

Sample median

The sample mean is the "gold standard" in terms of estimating the central tendency of a population probability distribution and is used in a vast majority of applications in which central tendency is of interest. Occasions arise, however, when the *sample median* is a better measure of central tendency.

Definition 1.6 Let $x_1, x_2, \ldots, x_n$ be experimental values associated with the random variables $X_1, X_2, \ldots, X_n$. The *sample median* is

$$M = \begin{cases} X_{((n+1)/2)} & n \text{ odd} \\ \left(X_{(n/2)} + X_{(n/2+1)}\right)/2 & n \text{ even,} \end{cases}$$

where $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ are the order statistics (the data values sorted into ascending order).

If n is odd, the sample median is just the middle sorted value; if n is even, the sample median is the average of the two middle sorted values.

Economists frequently use the sample median, rather than the sample mean, when reporting statistics concerning certain economic measures, such as incomes or house prices. To see why this is the case, consider a small M.S. program in operations research that graduates just $n = 5$ students in one particular academic year. The students assume positions in industry and report the following annual salaries:

$$\$71{,}000 \qquad \$65{,}000 \qquad \$74{,}000 \qquad \$194{,}000 \qquad \$73{,}000.$$

Now which would be a more accurate way to report the salary data in a recruiting brochure for the new class of operations researchers: use the sample mean $\bar{x} = \$95{,}400$ or use the sample median $m = \$73{,}000$ as the measure of central tendency? The student who graduated and took a salary of \$194,000 might have joined a family business or had a lucrative overseas offering. The other four salaries are fairly tightly clustered around the sample median $m = \$73{,}000$. The one high salary is a rarity, so it can either be considered an outlier or it can be an observation from a very long right-hand tail of the population probability distribution. In either case, reporting the sample median is the appropriate statistic to go in the brochure for next year. It gives the students the most accurate assessment of what their salary will be when they finish the M.S. program.

Determining the sampling distribution of the sample median can vary from simple to very complex. The two examples that follow span the two extremes.

Example 1.20 Let $X_1, X_2, \ldots, X_9$ be a random sample from a $U(0, 1)$ distribution. Find the sampling distribution of the sample median.

Unlike the three examples associated with determining the sampling distribution of the sample mean, this time the population distribution does not have any unknown parameters. The probability density function of X_i drawn from a $U(0, 1)$ population is

$$f_{X_i}(x) = 1 \qquad 0 < x < 1$$

for $i = 1, 2, \ldots, 9$. The corresponding cumulative distribution function on the support of X_i is

$$F_{X_i}(x) = x \qquad 0 < x < 1$$

for $i = 1, 2, \ldots, 9$. The observations are mutually independent and identically distributed random variables because they constitute a random sample. So the distribution of the sample median M, which is $X_{(5)}$ because $n = 9$ is odd, can be found using

the formula for the probability density function of the kth order statistic drawn from a continuous population,

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!}[F(x)]^{k-1} f(x)[1-F(x)]^{n-k} \qquad a < x < b$$

for $k = 1, 2, \ldots, n$, where a and b are the lower and upper limits of the support of the population probability distribution. Applying this formula to our sample of $n = 9$ observations from a $U(0, 1)$ population gives

$$f_M(x) = \frac{9!}{(5-1)!(9-5)!} x^{5-1} \cdot 1 \cdot (1-x)^{9-5} = 630x^4(1-x)^4 \qquad 0 < x < 1.$$

Notice that this probability density function is symmetric about $x = 1/2$. A Monte Carlo simulation experiment can be used to support our analytic work. The following R code generates 100 sample medians from 100 samples of size $n = 9$ drawn from a $U(0, 1)$ population distribution, plots a histogram, and overlays the histogram with the sampling distribution of the sample median derived above.

```
nrep = 100
medians = numeric(nrep)
for (i in 1:nrep) {
  x = runif(9)
  medians[i] = median(x)
}
hist(medians, probability = TRUE)
curve(630 * x ^ 4 * (1 - x) ^ 4, 0, 1, add = TRUE)
```

Executing this code after a call to `set.seed(7)` yields the graph shown in Figure 1.30.

For both the analytic values represented by the curve and the sample values represented by the histogram, the effect of choosing the fifth largest of the nine values is to push the probability distribution away from the extremes at 0 and 1 toward the center of the distribution at $1/2$. But are the histogram and the curve close enough to support our analytic work? The problem illustrated here is that we chose only `nrep = 100` replications of the simulation experiment, resulting in a rather noisy histogram. Random

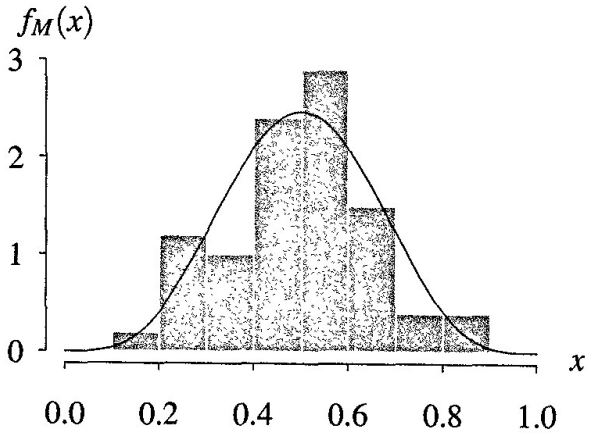


Figure 1.30: Sampling distribution of the sample median (100 replications).

sampling variability applies to Monte Carlo simulation as well as to collecting data. Figure 1.31 uses the same code, but this time with `nrep = 200000`. R chooses more cells for the histogram because of the larger number of replications; the histogram is much smoother this time. We now achieve a good match between the sampling distribution of M and its estimate via Monte Carlo simulation. This time our analytic work is supported by the simulation. The bell shape of the sampling distribution of M is not due to the central limit theorem, but rather due to the choice of the middle order statistic from a symmetric population probability distribution. Now that the sampling distribution of the sample median has been derived and supported by Monte Carlo simulation, it is often of value to know the expected value and population variance of the statistic of interest. The APPL statements below calculate the probability density function of the sample median M and its expected value and its population variance.

```
X := UniformRV(0, 1);
M := OrderStat(X, 9, 5);
Mean(M);
Variance(M);
```

The statements yield

$$E[M] = \frac{1}{2} \qquad \text{and} \qquad V[M] = \frac{1}{44}.$$

Notice that the expected value of the sample median equals the population median (this is $1/2$ by inspection because of the symmetry of the $U(0, 1)$ distribution). This is a good property for an estimator to have because the estimator is "on target" for estimating the population quantity. This property will be defined carefully in the next chapter, but for now $E[M] = E\left[X_{(5)}\right] = 1/2$ is stated in words as "the sample median is an unbiased estimator of the population median." The population variance of M is an indication of how far the sample median might stray from its target. We want the population variance of M to be as small as possible. One way to decrease the population variance of M is to increase the sample size n.

So the sample median seems like a reasonable estimator of the population median. But for this particular population distribution, the $U(0, 1)$ distribution, the population

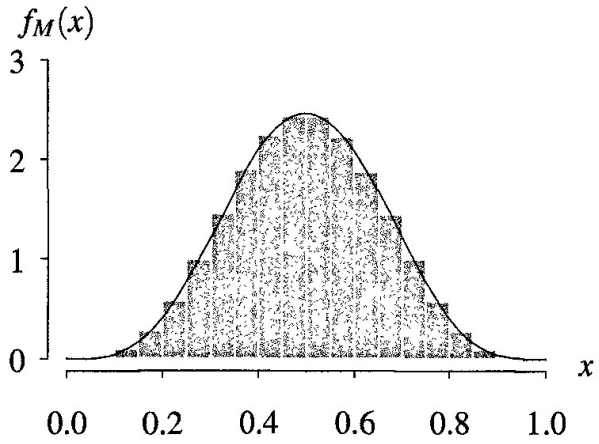


Figure 1.31: Sampling distribution of the sample median (200,000 replications).

median and the population mean both equal $1/2$. Would it be better to use the sample mean $\bar{X}$ to estimate the population median? One way to pin down the choice is to consider the population variance of the estimates. From above, the population variance of the sample median is

$$V[M] = \frac{1}{44},$$

but the population variance of the sample mean $\bar{X}$ by Theorem 1.1 is

$$V[\bar{X}] = \sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n} = \frac{1/12}{9} = \frac{1}{108}.$$

So the sample mean is more tightly clustered about the population median of $1/2$ than the sample median, and is therefore the preferred estimator of the population median for this particular symmetric population distribution. This will not be the case in general.

The previous example had two factors which made the analytic work tractable: an odd value for n and a particularly simple population distribution. In the next example, we remove both of those advantages and see the extra work associated with an even n and a more complicated population distribution.

Example 1.21 Let $X_1, X_2, \ldots, X_6$ be a random sample from a population having probability density function

$$f(x) = 2x \qquad 0 < x < 1.$$

Find the sampling distribution of the sample median.

As in the previous example, the population distribution does not involve any parameters. In contrast to the previous example, there are two complicating factors at play in this question: the *even* sample size $n = 6$ and the *slightly* more complicated population distribution. The even sample size implies that two adjacent order statistics will be averaged in order to arrive at the sample median. As you will see, these two extra factors create lots of extra work in deriving the sampling distribution of the sample median. The problem provides a good review, however, of the joint distribution of order statistics and the transformation technique. The random variable X_i has probability density function

$$f_{X_i}(x) = 2x \qquad 0 < x < 1$$

for $i = 1, 2, \ldots, 6$. The associated cumulative distribution function of X_i on its support is

$$F_{X_i}(x) = \int_0^x 2w\,dw = \left[w^2\right]_0^x = x^2 \qquad 0 < x < 1$$

for $i = 1, 2, \ldots, 6$. Since n is even, the sample median is calculated by averaging $X_{(3)}$ and $X_{(4)}$. Unfortunately, $X_{(3)}$ and $X_{(4)}$ are dependent random variables. So we begin the process of finding the probability density function of the sample median by finding the joint probability density function of $X_{(3)}$ and $X_{(4)}$. Using the same heuristic argument that gave us the probability density function of a single order statistic drawn from a continuous population, the joint probability density function of two order statistics $X_{(i)}$ and $X_{(j)}$, which is $f_{X_{(i)},X_{(j)}}(x_{(i)}, x_{(j)})$ for $i < j$, is given by the expression

$$\frac{n!}{(i-1)!(j-i-1)!(n-j)!}\left[F(x_{(i)})\right]^{i-1} f(x_{(i)})\left[F(x_{(j)}) - F(x_{(i)})\right]^{j-i-1} f(x_{(j)})\left[1 - F(x_{(j)})\right]^{n-j}$$

for $a < x_{(i)} < x_{(j)} < b$, where a and b are the lower and upper bounds on the support of the population distribution. Applying this formula to the population distribution described here, the joint probability density function of $X_{(3)}$ and $X_{(4)}$ is

$$f_{X_{(3)},X_{(4)}}(x_{(3)},x_{(4)}) = \frac{6!}{2!0!2!}\left[x_{(3)}^2\right]^2 2x_{(3)}\left[x_{(4)}^2 - x_{(3)}^2\right]^0 2x_{(4)}\left[1-x_{(4)}^2\right]^2$$

for $0 < x_{(3)} < x_{(4)} < 1$. This simplifies to

$$f_{X_{(3)},X_{(4)}}(x_{(3)},x_{(4)}) = 720x_{(3)}^5 x_{(4)}\left(1-x_{(4)}^2\right)^2 \qquad 0 < x_{(3)} < x_{(4)} < 1.$$

Now the sample median by Definition 1.6 is the average of $X_{(3)}$ and $X_{(4)}$, that is,

$$M = \frac{X_{(3)} + X_{(4)}}{2}.$$

To determine the probability density function of the sample median, we will use the transformation technique, which requires a "dummy" transformation. So the transformation consists of $Y_1 = g_1(X_{(3)}, X_{(4)}) = (X_{(3)} + X_{(4)})/2$, the sample median, and the dummy transformation $Y_2 = g_2(X_{(3)}, X_{(4)}) = (X_{(4)} - X_{(3)})/2$ because these particular functions can be solved in closed form for $X_{(3)}$ and $X_{(4)}$ and the associated Jacobian is tractable. The fact that this is a *linear* transformation ensures that the Jacobian is nonzero. The transformation

$$y_1 = g_1(x_{(3)},x_{(4)}) = \frac{x_{(3)}+x_{(4)}}{2} \qquad \text{and} \qquad y_2 = g_2(x_{(3)},x_{(4)}) = \frac{x_{(4)}-x_{(3)}}{2}$$

is illustrated in Figure 1.32 and is a bivariate one-to-one transformation from

$$\mathcal{A} = \{(x_{(3)},x_{(4)}) \mid 0 < x_{(3)} < x_{(4)} < 1\}$$

to

$$\mathcal{B} = \{(y_1,y_2) \mid y_2 > 0, y_2 < y_1 < 1-y_2\}.$$

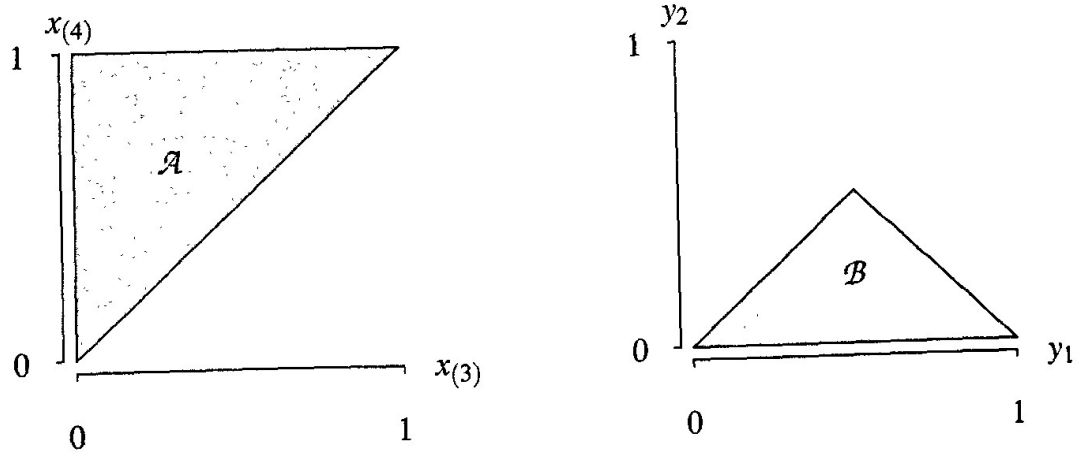


Figure 1.32: The support of $X_{(3)}$ and $X_{(4)}$ and the support of Y_1 and Y_2.

These functions can be solved in closed form for $x_{(3)}$ and $x_{(4)}$ as

$$x_{(3)} = g_1^{-1}(y_1, y_2) = y_1 - y_2 \qquad \text{and} \qquad x_{(4)} = g_2^{-1}(y_1, y_2) = y_1 + y_2$$

with associated Jacobian

$$J = \begin{vmatrix} 1 & -1 \\ 1 & 1 \end{vmatrix} = 2.$$

Applying the transformation technique, the joint probability density function of the random variables Y_1 and Y_2 is

$$f_{Y_1,Y_2}(y_1, y_2) = 720(y_1 - y_2)^5(y_1 + y_2)\left[1 - (y_1 + y_2)^2\right]^2 |2| \qquad (y_1, y_2) \in \mathcal{B},$$

which simplifies to

$$f_{Y_1,Y_2}(y_1, y_2) = 1440(y_1 - y_2)^5(y_1 + y_2)\left[1 - (y_1 + y_2)^2\right]^2 \qquad (y_1, y_2) \in \mathcal{B}.$$

Integrating y_2 out of the joint probability density function gives the marginal distribution of Y_1, which is

$$f_{Y_1}(y_1) = \begin{cases} \displaystyle\int_0^{y_1} 1440(y_1 - y_2)^5(y_1 + y_2)\left(1 - (y_1 + y_2)^2\right)^2 dy_2 & 0 < y_1 < 1/2 \\ \displaystyle\int_0^{1-y_1} 1440(y_1 - y_2)^5(y_1 + y_2)\left(1 - (y_1 + y_2)^2\right)^2 dy_2 & 1/2 < y_1 < 1, \end{cases}$$

or

$$f_{Y_1}(y_1) = \begin{cases} \frac{40960}{77}y_1^{11} - \frac{5200}{7}y_1^9 + \frac{1920}{7}y_1^7 & 0 < y_1 < 1/2 \\ -\frac{40960}{77}y_1^{11} + \frac{15280}{7}y_1^9 - \frac{28800}{7}y_1^7 + 7680y_1^5 - \\ \quad \frac{61440}{7}y_1^4 + 4800y_1^3 - \frac{10240}{7}y_1^2 + 240y_1 - \frac{1280}{77} & 1/2 < y_1 < 1, \end{cases}$$

which is the probability density function of the sample median. Replacing Y_1 with the sample median M, this could also be written as $f_M(x)$, which is a probability density function defined on the support $0 < x < 1$. This derivation had so many opportunities for a mathematical error that it is probably worthwhile conducting a Monte Carlo simulation check of this sampling distribution of the sample median. The following R code generates 200,000 sample medians for a sample of size $n = 6$ drawn from a population distribution with cumulative distribution function

$$F_X(x) = \begin{cases} 0 & x \le 0 \\ x^2 & 0 < x < 1 \\ 1 & x \ge 1. \end{cases}$$

The inverse cumulative distribution function is

$$F_X^{-1}(u) = \sqrt{u} \qquad 0 < u < 1,$$

so random variates are generated via

$$x \leftarrow \sqrt{u},$$

where u is a random number, that is, a realization of a $U(0, 1)$ random variable. The code that follows places a sorted sample of $n = 6$ values into the R vector x, then averages the two middle values to arrive at a sample median m.