Chapter 3
Statistics Concepts

In this chapter we discuss selected topics on probability. We define and graph several basic probability distributions. We review estimation, testing, and sampling from populations. The discussion here is at an intermediate technical level and at a speed appropriate for review of material learned in the prerequisite course.

3.1 A Brief Introduction to Probability

The quality of inferences are commonly conveyed by probabilities. Therefore, before discussing inferential techniques later in this chapter, we briefly digress to discuss *probability* in this section and *random variables* in Section 3.2.

If A is any *event*, $P(A)$ represents the probability of occurrence of A. Always, $0 \le P(A) \le 1$. The odds in favor of the occurrence of event A are

$$\frac{P(A)}{1 - P(A)} \tag{3.1}$$

and the odds against the occurrence of event A are

$$\frac{1 - P(A)}{P(A)} \tag{3.2}$$

Thus, if $P(A) = \frac{3}{4}$, then the odds in favor of A are 3, also referred to as 3 to 1, and the odds against A are $\frac{1}{3}$.

If B is a second event, $A \cup B$ represents the event that "either A or B occurs", that is, the *union* of A and B, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{3.3}$$

R.M. Heiberger, B. Holland, *Statistical Analysis and Data Display*,
Springer Texts in Statistics, DOI 10.1007/978-1-4939-2122-5_3

where $A \cap B$ is the event that "both A and B occur", that is the *intersection* of A and B. Events A and B are said to be *mutually exclusive* events if they cannot both occur; in this case, $A \cap B = \emptyset$ (the impossible event) and so $P(A \cap B) = 0$. Events A and B are said to be *independent* events if the occurrence or nonoccurrence of one of them does not affect the probability of occurrence of the other one; for independent events,

$$P(A \cap B) = P(A)\,P(B)$$

The *conditional probability* of B given A, written $P(B \mid A)$, is the probability of occurrence of B given that A occurs. If $P(A) \neq 0$,

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

Note that $P(B \mid A) = P(B)$ if A and B are independent events, but not otherwise.

To illustrate these ideas, imagine a box containing six white and four red billiard balls, identical to the touch. Suppose we select two balls from the box and let A = "the first ball is white" and B = "the second ball is white". A and B are independent events if the first ball is replaced in the box prior to drawing the second ball, but not otherwise. Let us assume that the first ball is not replaced so that the two events are dependent. Various sets of events are listed with their probabilities in Table 3.1.

In this table we demonstrate two ways to calculate the probability $\frac{78}{90}$ that we get a white ball in either the first selection or second selection or both selections. One way is with the formula for $P(A \cup B)$ in Equation (3.3). Another method is to recognize that the event "at least one white" can be partitioned into three mutually exclusive events: First draw white and second draw red; first draw red and second draw white; and both draws white. The probability of "at least one white" is seen to be the sum of the probabilities of the events comprising this partitioning.

3.2 Random Variables and Probability Distributions

A *random variable*, abbreviated as r.v., is a function that associates events with real numbers. For example, if we toss a coin 10 times, we can define an r.v. X to be the number of heads observed in these 10 tosses. This r.v. has *possible values* $x = 0, 1, 2, \ldots, 10$. Observing 7 heads among the 10 tosses is an event, and "7" is the number that this r.v. X associates with it.

A closely related concept is the r.v.'s *probability distribution*, which indicates how the total probability, 1, is distributed or allocated to the possible values of the r.v. It is usual to denote an r.v. with a capital letter and a possible value of this r.v. with the corresponding lowercase letter.

Table 3.1 Probability of intersection events, conditional events, union events in the setting of a box containing six white and four red billiard balls. We select two balls from the box. The A event is “the first ball is white” and the B event is “the second ball is white”. See Figure 3.1 for an illustration of this distribution.

Event	Position		Probability		Probability of event
	1	2	1	2	
A	W	?	$\frac{6}{10}$	1	$\frac{6}{10}$
B	?	W	1	$\frac{6}{10}$	$\frac{6}{10}$
$B \cap A$	W	W	$\frac{6}{10}$	$\frac{5}{9}$	$\frac{30}{90}$
$\bar{B} \cap \bar{A}$	R	R	$\frac{4}{10}$	$\frac{3}{9}$	$\frac{12}{90}$
$B \mid A$	$[W]$	W	$\frac{(\frac{6}{10})}{(\frac{6}{10})}$	$\frac{5}{9}$	$\frac{5}{9}$
$B \cup A$	W	R	$\frac{6}{10}$	$\frac{4}{9}$	$P(WR) + P(RW) + P(WW) = P(A) + P(B) - P(B \cap A) = P(B \cup A)$ $\frac{24}{90} + \frac{24}{90} + \frac{30}{90} = \frac{6}{10} + \frac{6}{10} - \frac{30}{90} = \frac{78}{90}$
	R	W	$\frac{4}{10}$	$\frac{6}{9}$	
	W	W	$\frac{6}{10}$	$\frac{5}{9}$	

3.2.1 Discrete Versus Continuous Probability Distributions

There are essentially two distinct types of probability distribution of a quantitative variable: discrete and continuous. (Random variables are also classified as discrete or continuous according to the classification of their probability distributions.) It is important to distinguish between the two types because they differ in their methods of display and calculation.

The key distinction between these two types relates to the spacings between adjacent possible values of the data. For discrete data, the distance separating consecutive possible values of the variable does not depend on a measurement device; indeed it may be completely arbitrary. For continuous data, the distances may (theoretically) assume all possible values in some interval.

For example, the number of times an archer hits a target in 10 attempts is a discrete variable because the answer is a count of the number of occurrences. It is impossible for there to be 3.5 hits. A discrete variable need not be integer-valued.

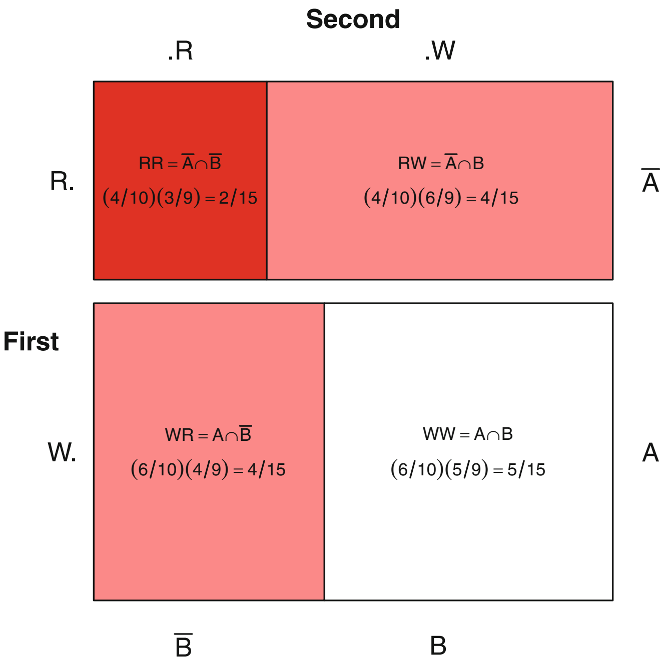


Fig. 3.1 Mosaic plot corresponding to Table 3.1. The area of each panel is equal to the probability of the event identified in that panel. The bottom row representing the event A = "W is selected first" consists of the two panels WR and WW. The bottom row has height .6 = $P(A)$. The right-hand column represents the event B = "W is selected second" consists of the two panels RW and WW. The event "$B \cap A$" is the white region WW in the lower right corner. The event WW has height .6 and width 5/9, hence area $.6 \times 5/9 = 1/3$. The event $B \mid A$ is also the white area WW, but now thought of as the proportion of the A area that is also B. The probability of $B \mid A$ is the ratio of the area of $B \mid A$ to the A area $(1/3)/.6 = 5/9$. The event $B \mid \bar{A}$ is the pink region RW in the upper right corner. The probability of $B \mid \bar{A}$ is the ratio of the pink area RW to the $\bar{A}$ area $(4/15)/.4 = 2/3$. The event $\bar{B} \cap \bar{A}$ is the red region RR in the upper left corner. The event RR has height .4 and width 3/9, hence area $.4 \times 3/9 = 2/15$.

The proportion of hits in 10 attempts is also discrete. It is impossible for this proportion to be .35. It is possible for a discrete variable to have a *countably infinite* number of possible values. An example would be the number of attempts needed for the archer to achieve her ninth hit. This variable can assume any positive integer value; it is possible but unlikely that the archer will need 100 attempts.

On the other hand, the archer's height in inches is a continuous variable because it can be anything between perhaps 3 feet and 8 feet (90–240 cm). While as a practical matter it would be difficult to measure height to within $\frac{1}{4}$-inch (6 mm) accuracy, it is not theoretically impossible for someone to be $68\frac{3}{4}$ inches (174.6 cm) tall.

In summary, it is possible to make a list of the possible values of a discrete random variable, but this is not true for a continuous random variable.

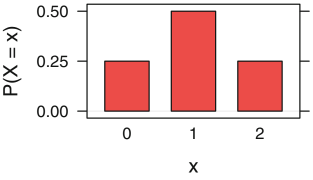
For completeness, we also point out that it is possible for data to be a mixture of discrete and continuous types. Let Y = the total measurable daily precipitation measured at Philadelphia International Airport. On some fraction of all days, roughly 70% of them, there is no precipitation. So $P(Y = 0) \approx .7$. But considering only those dates with measurable precipitation, Y is continuous, i.e., the distribution of $(Y \mid Y > 0)$ is continuous.

3.2.2 Displaying Probability Distributions—Discrete Distributions

The display of a probability distribution varies according to whether the r.v. is discrete or continuous. We can make an ordered list of the possible values of a discrete r.v. For example, if X denotes the number of heads in two tosses of a fair coin, then X has three possible values {0,1,2}. We will see later that for this coin, the probabilities are as given in Table 3.2.

Table 3.2 The total probability 1.0 has been *distributed* to the three possible values: 0, 1, 2.

x	$P(X = x)$
0	.25
1	.50
2	.25



Sometimes we choose to study several interdependent random variables at the same time. In such instances, we require their bivariate or multivariate probability distribution.

In Table 3.3 we consider an example of a discrete bivariate and conditional distribution. Here p.m.f. stands for *probability mass function.*

Here X and Y are dependent r.v.'s because, e.g., $f(1, 0) = .10$, which differs from $f(1) \times g(0) = .60 \times .15 = .09$. Alternatively, $f(1 \mid 0) = \frac{2}{3}$, which differs from $f(1) = .6$. In general, if U and V are discrete random variables, then U and V are independent r.v.'s if

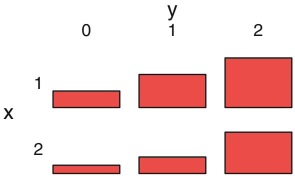
$$P\Big((U = u) \cap (V = v)\Big) = P(U = u) \times P(V = v)$$

for all possible values u of U and v of V, i.e., the distribution of U doesn't depend on the value of V.

Table 3.3 Example of Discrete Bivariate and Conditional Distributions. The top panel shows the probabilities of each of the six events in the distribution. The area of the six events adds up to 1. The center panel shows conditioning of x on y. Within each column, the area adds up to 1. The bottom panel shows conditioning of y on x. Within each row, the area adds up to 1.

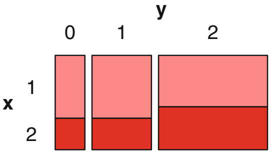
Joint p.m.f. $f(x, y)$

	y			
x	0	1	2	$y(x) = x$-margin
1	.10	.20	.30	.60
2	.05	.10	.25	.40
$g(y) = y$-margin	.15	.30	.55	1.00



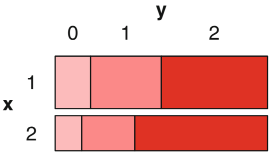
Conditional p.m.f. $f(x \mid y)$

	y		
x	0	1	2
1	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{6}{11}$
2	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{5}{11}$
all	1	1	1



Conditional p.m.f. $g(y \mid x)$

	y			
x	0	1	2	all
1	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	1
2	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{5}{8}$	1



The cumulative distribution $\mathcal{F}$ of a discrete random variable is calculated as

$$\mathcal{F}(x) = P(X \le x) = \sum_{t \le x} f(t)$$

where the sum is taken over all possible values t of X that are less than or equal to x.

3.2.3 *Displaying Probability Distributions—Continuous Distributions*

The probability distribution of a continuous random variable cannot be described in the manner of Table 3.2 or 3.3 (listing its possible values alongside their associated probabilities) because a continuous r.v. has an *uncountably infinite* number of possible values. Instead the probability distribution of a continuous r.v. X is described by its probability density function (p.d.f.), say $f(x)$. This function has the properties that

1. $f(x) \geq 0$
2. the probability that X lies in any interval is given by the area under $f(x)$ above this interval.

In the p.d.f. in Figure 3.2, the shaded area under the density and above the horizontal axis represents the probability that the random variable lies between 2 and 4.

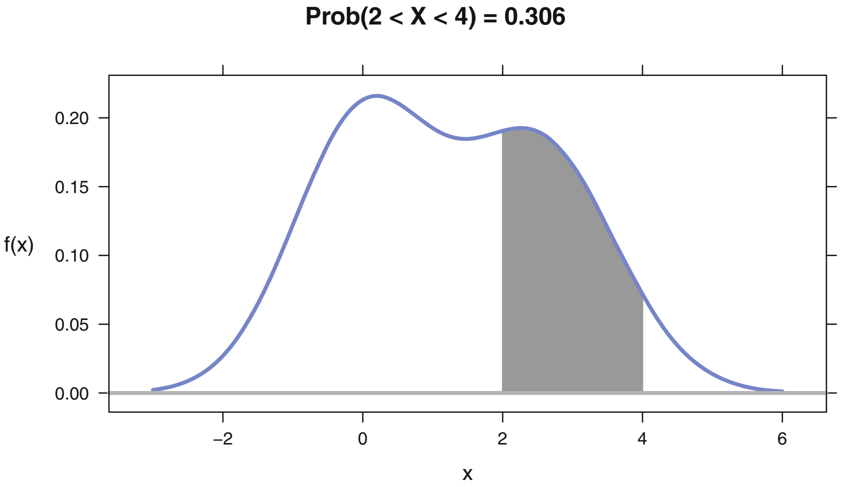


Fig. 3.2 $P(2 < X < 4)$ equals the area under the density between 2 and 4.

The cumulative distribution $\mathcal{F}$ of a continuous random variable is calculated as

$$\mathcal{F}(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$$

Continuous r.v.'s U and V are also independent if the distribution of U doesn't depend on the value of V or, equivalently, if the distribution of V doesn't depend on the value of U. In this case, we can express the independence condition as

$$P\Big((U \le u) \cap (V \le v)\Big) = P(U \le u) \times P(V \le v) \tag{3.4}$$

for all u and v.

Appendix J catalogs frequently encountered probability distributions, illustrates their density functions, and includes function names in R for calculations with the distributions.

3.3 Concepts That Are Used When Discussing Distributions

Understanding the distribution of observations is critical to interpreting data. In this section we introduce several concepts that are used to describe distributions: mean, variance, median, symmetry, correlation; and types of graphs that are used to display these concepts: histogram, stem-and-leaf, density, scatterplot.

3.3.1 Expectation and Variance of Random Variables

The expectation of an r.v. X, denoted $E(X)$, is its expected or long-run average value; alternatively it is the mean of the probability distribution of X and so we write $E(X) = \mu$. If X is discrete with p.m.f. $p(x)$, then $E(X) = \sum x\, p(x)$. If X is continuous, then $E(X) = \int x\, f(x)\, dx$, where the range of integration extends over the set of real numbers that X may assume. The variance of X is defined by $\sigma^2 = \text{var}(X) = E(X - \mu)^2 = E(X^2) - \mu^2$. The square root σ of the variance is called the *standard deviation*, abbreviated s.d. It is a more useful measure of variability than the variance because it is measured in the same units as X, rather than in artificial squared units.

If $x_1, x_2, \ldots, x_n$ is a random sample of n items selected from some population, the sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3.5}$$

estimates the population mean μ, and the sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3.6}$$

estimates the population variance σ^2. In addition, the sample standard deviation $s = \sqrt{s^2}$ estimates the population standard deviation σ. Please see Section G.12 for a discussion on the importance of using the two-pass algorithm based on the definition in Equation 3.6, and not the alternative one-pass algorithm based on Equation 3.7,

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i^2 - n\bar{x}^2 \right) \tag{3.7}$$

when doing arithmetic by computer. The short explanation is that you will always get the right answer with Equation 3.6 and may sometimes get a very wrong answer with Equation 3.7.

It can be shown that if a_1 and a_2 are constants and x_1 and x_2 are any two random variables, then

$$E(a_1 x_1 \pm a_2 x_2) = a_1 E(x_1) \pm a_2 E(x_2) \tag{3.8}$$

If, in addition, x_1 and x_2 are uncorrelated random variables, then

$$\text{var}(a_1 x_1 \pm a_2 x_2) = a_1^2 \,\text{var}(x_1) + a_2^2 \,\text{var}(x_2) \tag{3.9}$$

When x_1 and x_2 are correlated, then the variance of the sum is given by

$$\text{var}(a_1 x_1 \pm a_2 x_2) = a_1^2 \,\text{var}(x_1) + a_2^2 \,\text{var}(x_2) \pm 2a_1 a_2 \,\text{cov}(x_1, x_2) \tag{3.10}$$

These three formulas (Equations 3.8, 3.9, and 3.10) generalize to the multivariate situation in Equations 3.16 and 3.17.

3.3.2 Median of Random Variables

The median of an r.v. X, denoted median(X) $=$ η, is the middle value of the distribution. The population median is defined as the value η such that

$$\int_{-\infty}^{\eta} f(x)\, dx = .5 \qquad \text{for continuous distributions} \tag{3.11}$$

or

$$\sum_{x \le \eta} p(x) \ge .5 \text{ and } \sum_{x < \eta} p(x) \le .5 \qquad \text{for discrete distributions.} \tag{3.12}$$

We show an example of the median of a distribution in Figure 3.6.

The order statistics $X_{(i)}$ are the values of the observed X_i ordered from smallest to largest. The middle order statistic $\overset{\perp}{X}$ is called the sample median and is defined as

$$\overset{\perp}{X} = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{odd } n \\[2ex] \left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n+1}{2}\right)} \right) / 2 & \text{even } n \end{cases} \tag{3.13}$$

The notation $\overset{\perp}{X}$ for the sample median used here is intended to be self-descriptive, with an overbar split in the middle into two equal halves. We believe the notation is due to Tukey. There is no standard notation for the median.

3.3.3 Symmetric and Skewed Distributions

Symmetry and skewness are classifications applicable to both continuous and discrete distributions. The mean of a symmetric distribution coincides with its median. A continuous distribution example is the normal distribution having a density function such as that plotted in Figure 3.13. A symmetric distribution has equivalent behavior on either side of its mean. In particular, its *tails*, the values of the density function away from the center, are mirror images.

A skewed distribution is one that is not symmetric. Unimodal distributions (ones having a single point where the probability mass is higher than at adjacent points) that are skewed are further classified as being positively or negatively skewed. A positively skewed distribution has a long, thin tail on its right side and a short, fat tail on its left side. Its mean exceeds its median. A negatively skewed distribution has a long, thin tail on its left side and a short, fat tail on its right side. Its median exceeds its mean. Note that the left/right naming convention for skewed distributions is based on the side containing the long, thin tail. We illustrate a negatively skewed, symmetric, and positively skewed distribution in Figure 3.3. We show boxplots of negatively skewed, symmetric, and positively skewed data in Figure 3.7.

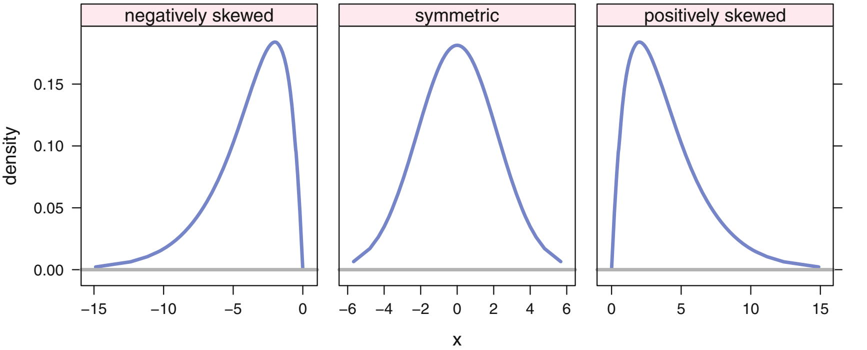


Fig. 3.3 Negatively skewed, symmetric, and positively skewed distributions.

The χ^2 distribution described in Section J.1.3 is an example of a continuous positively skewed distribution. The (discrete) binomial distribution to be described in Section 3.4.1 is negatively skewed, symmetric, or positively skewed according to whether its parameter p is less than, equal to, or greater than 0.5.

The skewness terminology often comes into play because many statistics procedures work best when underlying distributions are symmetric, and tactics that move the distribution toward symmetry (for example, with data transformations such as the power transformations described in Section 4.8) are frequently used in the analysis of skewed distributions.

Each of the densities in Figure 3.3 has a single mode. Some densities have more than one mode. Figure 3.2 is an example of a bimodal density, with one mode between 0 and 1 and another mode between 2 and 3. Multimodal distributions, ones having more than two modes, are occasionally encountered. Sometimes bimodality and multimodality arise as a result of interpreting samples coming from two or more populations with different locations as having arisen from a single population. Therefore, bimodality or multimodality may suggest a need for disaggregation of samples.

3.3.4 Displays of Univariate Data

It is difficult to gain an understanding of data presented as a table of numbers. Summary statistics such as those presented in the preceding sections are helpful for this purpose but may fail to capture some important features. In this section we present three displays (Histogram, Stem-and-leaf, and Boxplots) for univariate data that are basic tools for studying both the distributional shape and unusual data values. We illustrate these displays with the variable `male.life.exp` (1990 male life expectancy) in each of 40 countries, part of the datafile `data(tv)` to be examined in more detail in Section 4.6. We summarize the variable in Table 3.4 as a frequency table, a partitioning of the data into *k* evenly spaced nonoverlapping categories, and a tally of the number or proportion of items in each category.

3.3.4.1 Histogram

The construction of a histogram begins with the frequency table. Usually the number of categories is between 6 and 12—the use of fewer than 6 categories tends to undersummarize the data while the use of more than 12 categories tends to oversummarize the data. For `male.life.exp` we chose 6 age-range categories that encompass the ages from all 40 countries.

The corresponding histogram in Figure 3.4 is a graph consisting of rectangles with width covering the breadth of the classes and heights equal to the class frequencies. This plot is also called a *relative frequency* histogram, particularly when the vertical axis is labeled to show the *proportion* of countries in each category, for example $\frac{6}{40} = 0.15$ in the first category for ages 50–54. We show both axis labelings in Figure 3.4 with the proportion axis on the right.

Table 3.4 Frequency Distribution of Male Life Expectancy

```
> data(tv)

> tmp <- as.matrix(table(cut(tv$male.life.exp, breaks=seq(49.5,79.5,5))))

> dimnames(tmp) <-
+     list("Male Life Expectancy"=
+               c("50--54","55--59","60--64","65--69","70--74","75--79"),
+          " "="Frequency")

> tmp

Male Life Expectancy Frequency
              50--54         6
              55--59         4
              60--64         9
              65--69        11
              70--74         7
              75--79         3
```

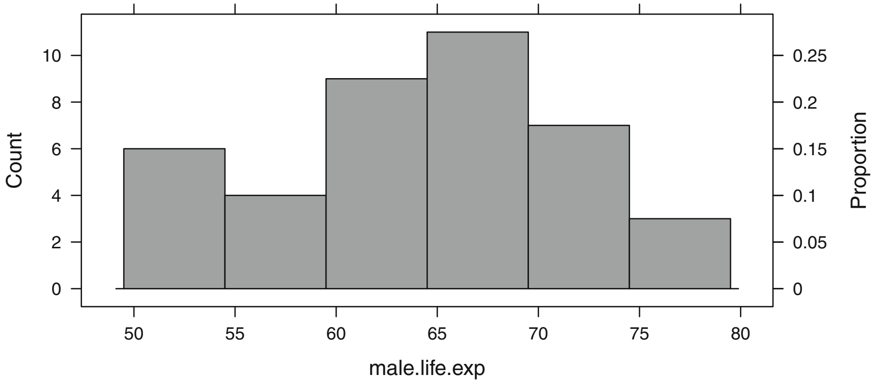


Fig. 3.4 Life Expectancy for Males. The count axis is on the left and the proportion axis is on the right.

Figure 3.4 is an example of a bimodal distribution, one having two peaks. In this example, the lower peak may correspond to economically poorer countries and the upper peak to wealthier countries, with relatively few countries falling between these extremes. In general, bimodal distributions sometimes suggest an amalgamation of samples from two separate populations that perhaps should be investigated separately. An advantage of histograms is that they can be constructed from huge datasets with no more effort than from small data sets. A disadvantage is that the data used to construct a histogram cannot be recovered from the histogram itself.

3.3.4.2 Stem-and-Leaf Display

Stem-and-leaf displays, designed by John Tukey, resemble histograms in that they portray the shape of a distribution. The stem-and-leaf display is usually preferable because it is possible to recover the data used to construct a stem-and-leaf display (at least to some degree of precision). Unlike histograms, stem-and-leaf displays are limited to data sets of not more than a few hundred observations in order that the display fits entirely on one page or one computer monitor.

A stem-and-leaf display for male life expectancy is in Table 3.5. This is a table, not a figure, because stem-and-leaf is a text-based graphic display.

Table 3.5 Stem-and-Leaf Display of Male Life Expectancy

```
> stem(tv$male.life.exp)

  The decimal point is 1 digit(s) to the right of the |

  5 | 002234
  5 | 6799
  6 | 012223344
  6 | 66777888899
  7 | 1223334
  7 | 556
```

The column of numbers in this display to the left of the vertical bars represent the tens digit of each of the life expectancies. This column is the *stem*. The numbers to the right of the vertical bars, one digit for each country, are the leaves, the unit digits of the life expectancies for the 40 countries. The stem-and-leaf display, following Tukey, rounds down, to maintain the same digit as appears in the data table. A 90° counterclockwise rotation of the stem and leaves gives a picture that closely resembles Figure 3.4. The legend locating the decimal point tells the reader that "`5 | 0`" in the display stands for 50, rather than .05 or 500.

Stem-and-leaf displays can accommodate measurements containing more than two significant digits. This is accomplished either by suppressing the values of trailing digits or by allowing more than a single digit for each leaf. For example, suppose in a different problem the measurement is 564. This can be represented as "`5 | 6`", with the stem indicating the hundreds, rounding the units digit down to a multiple of 10, and with a legend locating the decimal point 2 places to the right of the vertical bar. Alternatively, it can be represented with a stem indicating the hundreds and with two-digit leaves as "`5 | 64,`", again locating the decimal point two places to the right of the vertical bar, and with the "`,`" indicating that the leaf is two digits wide. Or, another option, as "`56 | 4`" with a stem of 56 tens (representing 560) and with a single-digit leaf of 4.

3.3.4.3 Boxplots

Boxplots, also known as box-and-whisker plots, are among the many inventions of John Tukey. Their main use is as a compact, simultaneous display to compare several related data sets. Many examples of side-by-side boxplots appear in this book. Boxplots may be arranged along either a vertical or horizontal scale. This book contains examples illustrating both options.

Boxplots make use of the sample first quartile Q_1, median $\overset{\perp}{x} = Q_2$, and third quartile Q_3. The statistics $Q_1, \overset{\perp}{x}, Q_3$ divide the sample into four equal parts. Q_1 is the median of the sample values that are less than or equal to $\overset{\perp}{x}$ and Q_3 is the median of the sample values that are greater than or equal to $\overset{\perp}{x}$. Approximately 25% of the sample lies within each of the four intervals (all finite intervals are closed, so double counting is possible)

$$(-\infty, Q_1], \qquad [Q_1, \overset{\perp}{x}], \qquad [\overset{\perp}{x}, Q_3], \qquad [Q_3, \infty)$$

A rectangle (box) is drawn so that when placed against a numerical scale its edges occur at Q_1 and Q_3. A line is drawn, parallel to the edges, through the inside of the box at the median $\overset{\perp}{x}$. Lines perpendicular to the edges of the box extend outward from the midpoints of the edges. These lines are sometimes called "whiskers". The lower whisker extends to the lowest sample item not more than $1.5 \times \text{IQR}$ below Q_1. The upper whisker extends to the largest sample item not more than $1.5 \times \text{IQR}$ above Q_3. Points outside the range of the whiskers are plotted as filled-in circles. Such points are deemed extreme or outlying values ("outliers"). In general, outliers should be carefully scrutinized. Sometimes they are due to transcription errors and are not legitimately part of the data under consideration (in which case you should attempt to correct the data). Other times, they are the critical data points that provide the key to an explanation of the study. One example of a critically important outlier is the Gulf of Mexico oil spill. On most days very little oil is released into the ocean. If we ignored the large spill detected on 20 April 2010, we would be missing the important information. In astronomy, ”transient” events are very important. That is how supernovas are detected (Table 3.6).

Table 3.6 shows the quartiles for the `male.life.exp` variable. Figure 3.5 shows the boxplot for the `male.life.exp` variable.

Table 3.6 Quartiles of Life Expectancy for Males

```
> quantile(tv$male.life.exp)
   0%   25%   50%   75%  100%
50.00 59.75 66.00 69.50 76.00
```

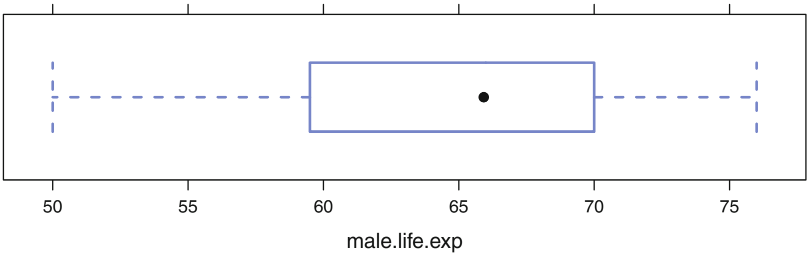


Fig. 3.5 Boxplot of Life Expectancy for Males

See the illustration in Figure 3.6 for the quartiles of a continuous distribution. The interquartile range

$$\text{IQR} = Q_3 - Q_1$$

is a measure of dispersion of the central portion of a distribution. When X is normally distributed $X \sim N(\mu, \sigma^2)$, we have $\text{IQR} = 1.34898\sigma$.

Figure 3.7 contains parallel boxplots depicting three samples on a common scale, illustrating the distinctions between boxplots for negatively skewed, symmetric, and positively skewed distributions. This parallels the density presentations in Figure 3.3. Asymmetry is nicely displayed in this figure.

Several more elaborate versions of the boxplot exist. For example, adding a *notch* to the sides of a box provides information on the variability of the sample median. For details, see Hoaglin et al. (1983).

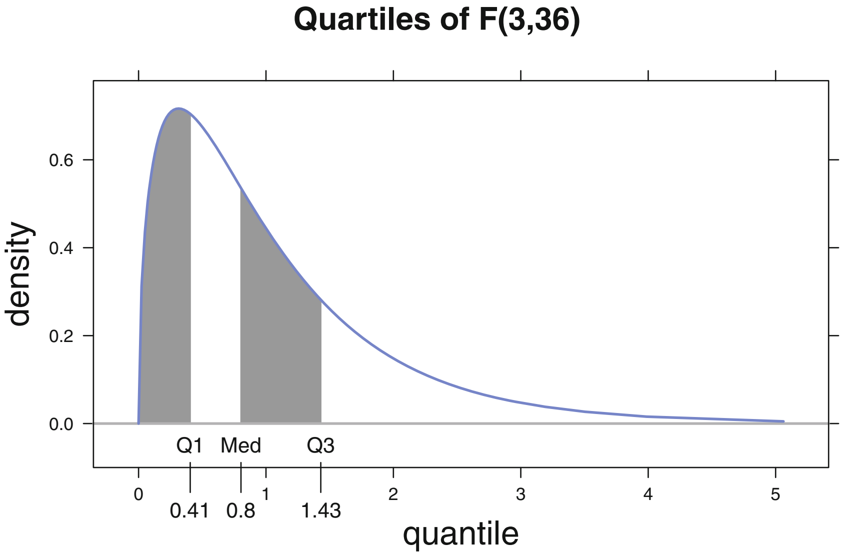


Fig. 3.6 Illustration of median and quartiles for a continuous distribution.

Boxplots are generally unsuccessful in conveying the existence of multiple modes. For such data, histograms and stem-and-leaf displays are often preferred choices.

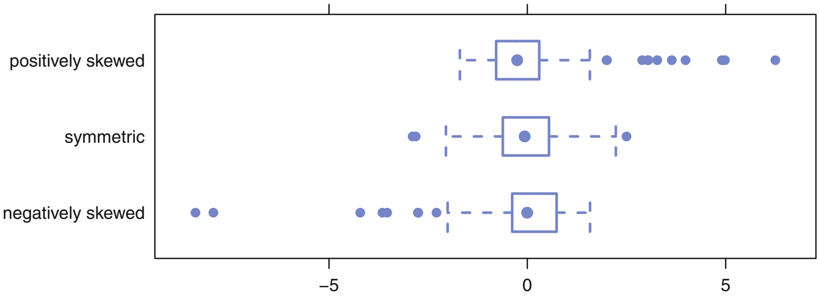


Fig. 3.7 Boxplots illustrating negatively skewed, symmetric, and positively skewed distributions.

3.3.5 Multivariate Distributions—Covariance and Correlation

In Section 3.2.2 we give an example of a discrete multivariate (actually bivariate) probability distribution. We now touch on the notion of the continuous multivariate distribution of a continuous random vector $X = (X_1, X_2, \ldots, X_p)'$. For example, variable X_1 could be height and variable X_2 weight, all measured on the same set of people. The mean or expectation of X is $\mu = (\mu_1, \mu_2, \ldots, \mu_p)'$, the vector of means of the univariate distribution of the $X_i's$. The variance–covariance matrix of X, say V, also called the covariance matrix or dispersion matrix, is the symmetric $p \times p$ matrix having the variances of the $X_i's$ on its main diagonal, and the covariances of different $X_i's$ elsewhere. The covariance of X_i and X_j is

$$V_{ij} = \sigma_{ij} = \text{cov}(X_i, X_j) = E\Big((X_i - \mu_i)(X_j - \mu_j)\Big)$$

is the element in the row i column j position of V. If we denote the standard deviations of X_i and X_j by σ_i and σ_j, respectively, then the *correlation* between X_i and X_j is

$$\rho_{ij} = \frac{\text{cov}(X_i, X_j)}{\sigma_i \sigma_j} = \frac{V_{ij}}{\sqrt{V_{ii} V_{jj}}} \tag{3.14}$$

This is a rescaling of the covariance, interpreted as a measure of the strength of the (straight line) linear relationship between X_i and X_j. It can be shown that $-1 \le \rho_{ij} \le 1$. If this correlation is close to ±1, X_i and X_j are closely linearly associated; the association is direct if $\rho_{ij} > 0$ and inverse if $\rho_{ij} < 0$. If $\rho_{ij} = 0$, then X_i and X_j are said to be *uncorrelated*, i.e., the X's are not linearly related. It is easy to construct an example of correlated variables for any specified correlation. Figure 3.8 gives a static view of a sequence of related variables with specified correlation coefficient. A dynamic illustration of the effect of the correlation coefficient can be constructed by plotting a sequence of panels similar to those in Figure 3.8 and cycling through them. We do so in a **shiny** app in the **HH** package with the statement

```
shiny::runApp(system.file("shiny/bivariateNormalScatterplot",
                          package="HH"))
```

at the R prompt. See Figure E.3 for a screenshot. In both the static and dynamic illustrations the formula is very simple. Define x and e as independent realizations from the $N(0, 1)$ distribution. Then

$$y = \rho x + (1 - \rho^2)^{1/2} e \tag{3.15}$$

has correlation ρ with x.

Matrix algebra plays an important role in the study of multivariate distributions. For example, in matrix notation, the covariance matrix is

$$V = E\Big((X - \mu)(X - \mu)'\Big)$$

and the correlation matrix P (uppercase ρ) is given by

$$\mathrm{P} = \Big(\mathrm{diag}(V)\Big)^{-\frac{1}{2}} V \Big(\mathrm{diag}(V)\Big)^{-\frac{1}{2}}$$

When the individual x_i are normally distributed, their joint distribution is called the multivariate normal and is notated $x \sim N(\mu, V)$. The bivariate ($p = 2$) normal distribution with means $\mu_i = 0$, variances $\sigma_i^2 = 1$, and correlation $\rho = .7$ $\left[\text{hence } V = \begin{pmatrix} 1.0 & 0.7 \\ 0.7 & 1.0 \end{pmatrix}\right]$ is plotted as a three-dimensional object in Figure 3.9. This is actually one panel of the set of rotated views of the density shown in Figure 3.10. A rotating version (see the **shiny** screenshot in Figure E.2) of the bivariate normal density example runs in R with the statement

```
shiny::runApp(system.file("shiny/bivariateNormal",
                          package="HH"))
```

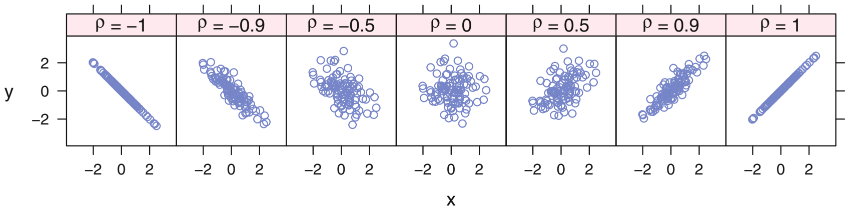


Fig. 3.8 Bivariate Normal distribution—scatterplot at various correlations. The distributions in the panels are related. The x-variable in all panels is the same. The y is generated from a common e-variable by the formula $y = \rho x + (1 - \rho^2)^{1/2} e$ for a sequence of values for ρ. The x- and e-variables were independently generated from the N(0,1) distribution. We provide a **shiny** app `bivariateNormalScatterplot` for a dynamic version of this set of panels. See Figure E.3 for a screenshot.

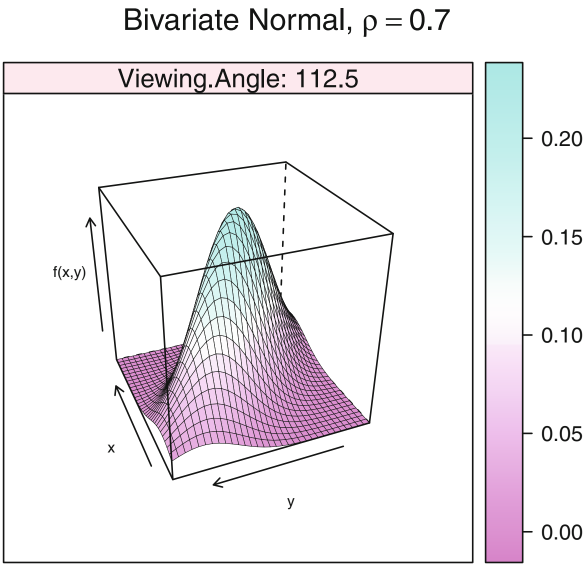


Fig. 3.9 Bivariate Normal density with $(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \rho) = (0, 1, 0, 1, .7)$ in 3D space with viewing angle = 112.5°. A set of eight viewing angles is shown in Figure 3.10

If X and Y are random vectors with

$$Y = B + CX$$

for some vector B and some matrix C, then

$$E(Y) = B + C\,E(X) \tag{3.16}$$

and

$$\text{var}(Y) = C\ \text{var}(X)\,C' \tag{3.17}$$

If, moreover, X has a multivariate normal distribution, then so does Y. In other words, linear functions of normal r.v.'s are normal. Equations 3.16 and 3.17 generalize the scalar versions in Equations 3.8, 3.9, and 3.10.

It follows from Equation 3.17 that if X_1, X_2, X_3, X_4 are univariate random variables, then

$$\text{var}(X_1 + X_2) = \text{var}(X_1) + \text{var}(X_2) + 2\,\text{cov}(X_1, X_2)$$

and

$$\text{cov}(X_1 + X_3, X_2 + X_4) = \text{cov}(X_1, X_2) + \text{cov}(X_1, X_4) + \text{cov}(X_3, X_2) + \text{cov}(X_3, X_4)$$

If Y has a k-dimensional multivariate normal distribution with mean μ and covariance matrix V, then

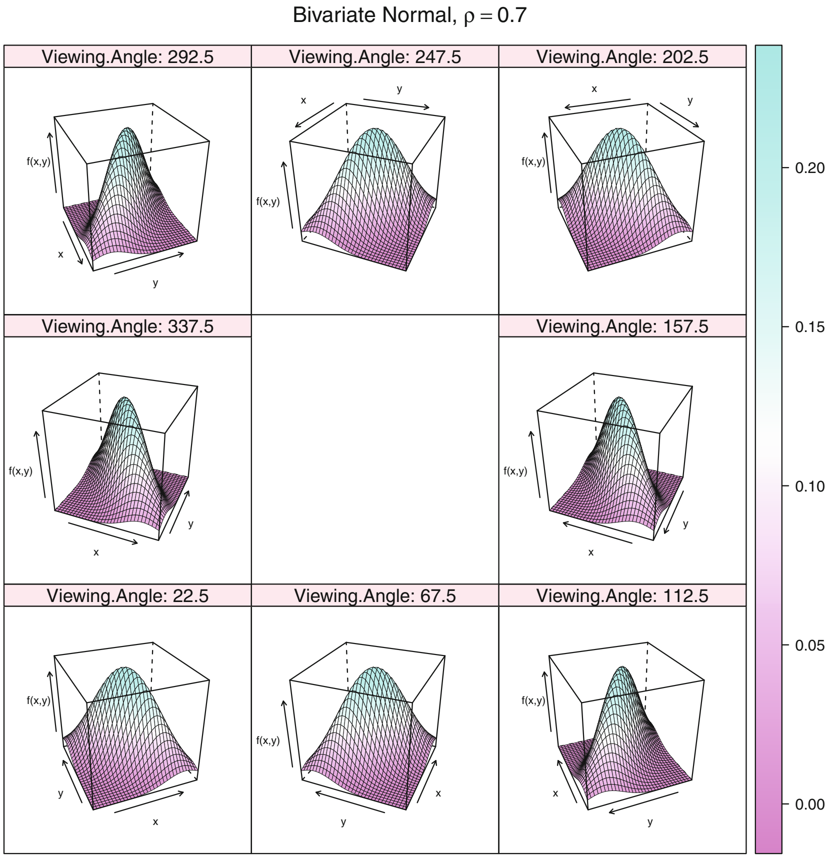


Fig. 3.10 Bivariate Normal density in 3D space with various viewpoints. Figure 3.9 shows a higher resolution view of the 112.5° panel. The reader can view an interactive version of this plot with the **shiny** app `shiny::runApp(system.file("shiny/bivariateNormal", package="HH"))`. See Figure E.2 for a screenshot.

$$Q = (Y - \mu)' V^{-1} (Y - \mu)$$

has a χ^2 distribution with k degrees of freedom (See Appendix J).

3.4 Three Probability Distributions

In this section we introduce three probability distributions, the (discrete) binomial distribution and the (continuous) Normal and t distributions, that frequently arise in practice. Details of how to perform probability-related calculations for these and other frequently encountered distributions are discussed in Appendix J.

3.4.1 The Binomial Distribution

The binomial distribution is perhaps the most commonly encountered discrete distribution in statistics. Consider a sequence of n independent trials, or mini-experiments, each of which can result in one of just two possible outcomes. For convenience these outcomes are labeled *success* and *failure* although in context the success outcome may not connote a favorable event. Further assume that the probability of success, p, is the same for each trial. Let X denote the number of successes observed in the n trials. Then X has a binomial distribution with parameters n and p. This distribution has mean $\mu = np$ and standard deviation $\sigma = \sqrt{np(1-p)}$. We show an illustration of the discrete density for the binomial with $n = 15$ and $p = .4$ in Section J.3.2. In Figure 3.11 we show the discrete density for the binomial with $n = 15$ and $p = .4$, underlaid with the normal approximation with $\mu = np = 15 \times .4 = 6$ and $\sigma = \sqrt{np(1-p)} = \sqrt{15 \times .4 \times .6} = \sqrt{3.6} = 1.897$.

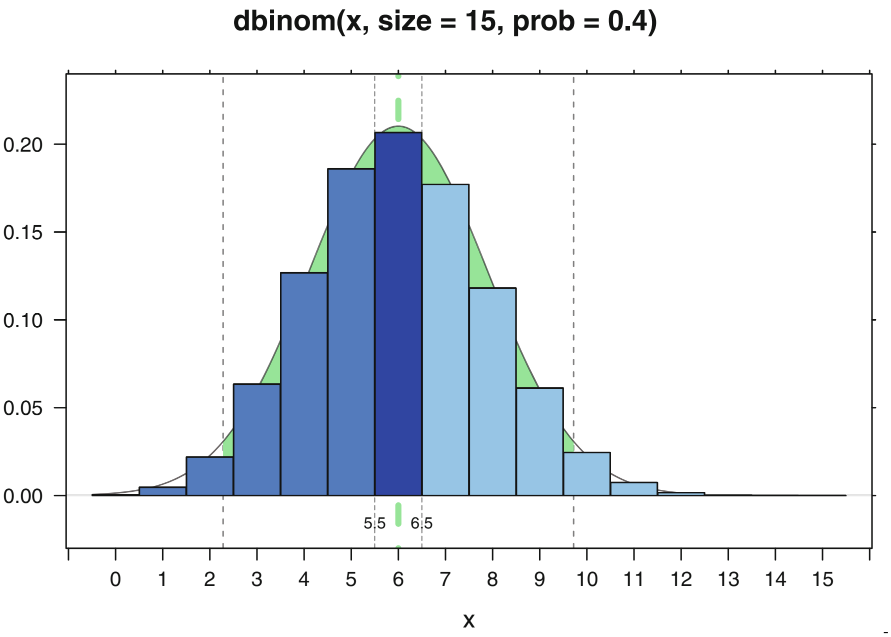
The above scenario is widely applicable. If one randomly samples with replacement from a population with a proportion p of successes, then the number of successes in the sample is binomially distributed. Even if the sampling is *without* replacement, the number of successes is approximately binomial if the population size is much greater than the sample size; in this case the first two assumptions above are only mildly violated. Applications include the number of voters favoring a candidate in a political poll, the number of patients in a population that suffer from a particular illness, and the number of defective items in one day's output from an assembly line.

However, it is not unusual for one or more of the binomial assumptions to be violated. For example, suppose we sample *without* replacement from a population of successes and failures and the population size is not much greater than the sample size, say less than 20 times as large as the sample. Then the trials are not independent and the *success* probability is not constant from trial to trial. (In this situation the correct distribution to use for X is the *hypergeometric* distribution. See Appendix J.)

Similarly, the binomial model is unlikely to apply to the number of hits by the archer in Section 3.2.1 because her shots (trials) may not be independent and may not have the same probability of a hit.

Usually in practice, we need to calculate not just $P(X = x)$, the probability of achieving *exactly* x successes, but probabilities of an interval of successes such as $P(X \le x)$, the probability of *at most* x successes, or $P(a \le X \le b)$, the probability of observing between a and b successes inclusive.

A table of binomial probabilities can be used when n and p appear in the table. Otherwise, as illustrated in Appendix J, R functions can easily be used to produce accurate results.



```
> pbinom(size=15, prob=.4, q=6)
[1] 0.6098

> pnorm(q=6.5, mean=15*.4, sd=sqrt(15*.4*(1-.4)))
[1] 0.6039

> dbinom(size=15, prob=.4, x=6)
[1] 0.2066

> diff(pnorm(q=c(5.5, 6.5), mean=15*.4, sd=sqrt(15*.4*(1-.4))))
[1] 0.2079
```

Fig. 3.11 We show the discrete density for the binomial with $n = 15$ and $p = .4$, underlaid with the normal approximation with $\mu = np = 15 \times .4 = 6$ and $\sigma = \sqrt{np(1-p)} = \sqrt{15 \times .4 \times .6} = \sqrt{3.6} = 1.897$. The dark bar at $x = 6$ has probability $P(x = 6) = .2066$ from the binomial and $P(5.5 < x < 6.5) = .2079$ from the normal. The dark bar at $x = 6$ and all bars to its left together have probability $P(x \le 6) = .6098$ from the binomial and $P(x < 6.5) = .6039$ from the normal approximation. The normal approximations are calculated with the correction for continuity (the interval [6-.5, 6+.5] is the full width of the dark bar at $x = 6$).

3.4.2 The Normal Distribution

Many natural phenomena follow the normal distribution, whose probability density function is the familiar “bell-shaped” curve, symmetric about the mean μ. In addition, a celebrated theoretical result called the Central Limit Theorem says that the sampling distributions of sample means (see Section 3.5), sample proportions, and sample totals each are approximately normally distributed if the sample size is "sufficiently large." Since this theorem applies to almost all possible probability distributions from which a sample might be selected, including discrete distributions, the theorem brings the normal distribution into play in a wide variety of circumstances.

If X has a normal distribution with mean μ and standard deviation σ, and we define the standardization of X as $Z = \frac{X-\mu}{\sigma}$, then Z is normally distributed with mean 0 and standard deviation 1, *i.e.,* the *standard normal* distribution. We write $X \sim N(\mu, \sigma^2)$ to indicate that X has a normal distribution with mean μ and variance σ^2 (or we could say with standard deviation σ). In this notation, the standard normal distribution is $N(0, 1)$. The density function $\phi(z)$ and cumulative distribution function $\Phi(z)$ are defined in Section J.1.9 and illustrated in Figure 3.12.

The normal distribution is "bell-shaped" and symmetrically distributed about μ, which is also this distribution's median and mode. Almost all of the probability is concentrated in the interval $\mu \pm 3\sigma$. We use z_α to be the solution to the equation $P(Z > z_\alpha) = \alpha$. This is the value on the horizontal axis that has area α under the curve and to its right. For example, $z_{.05} = 1.645$. Figure 3.13 shows the normal density function for a $N(100, 25)$ distribution. If X has this distribution, the left shaded area in Figure 3.13 represents 95% of the area under the density function. That is,

$$P(Z < 1.645) = P\left(\frac{X-\mu}{\sigma} < 1.645\right) = P(X < 108.225) = .95$$

after substituting $\mu = 100$ and $\sigma = 5$. The right shaded area is

$$\alpha = .05 = P\Big((X-\mu)/\sigma \geq \Phi^{-1}(1-\alpha) = 1.645\Big)$$

A dynamic version of any call to the `NTplot` function is available as a **shiny** app in the **HH** package with the argument `shiny=TRUE` included as an additional argument, for example

```
NTplot(shiny=TRUE)
```

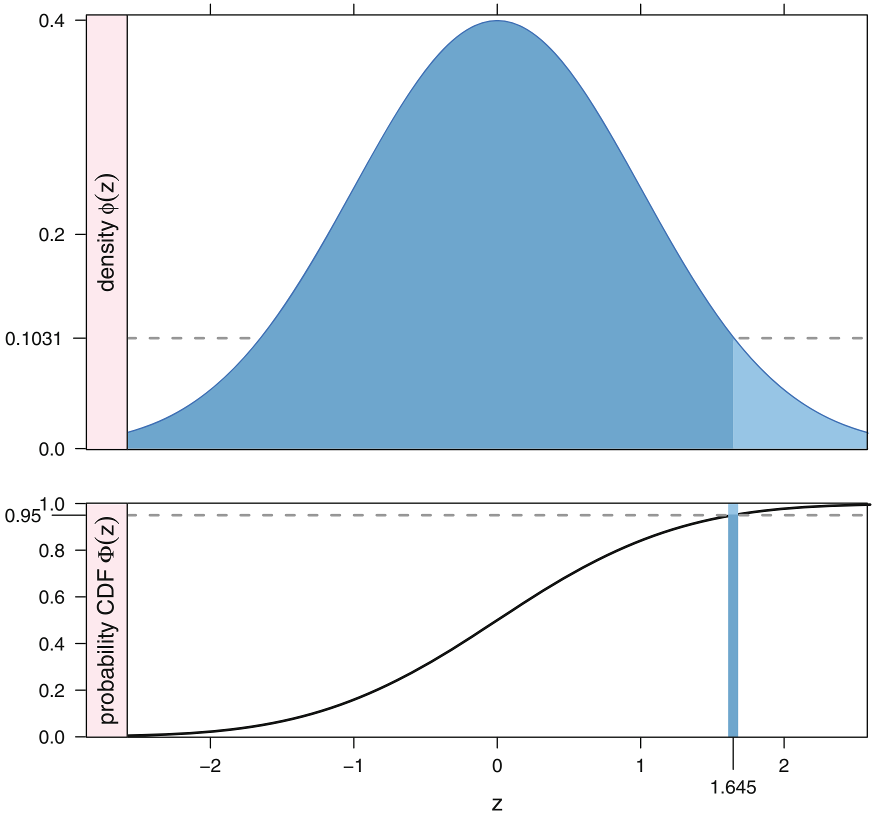
A dynamic version of Figure 3.13 is initialized with the call

```
NTplot(mean0=100, mean1=NA, xbar=NA, xlim=c(75, 125),
       sd=5, digits=6, zaxis=TRUE, cex.z=0.6,
       cex.prob=.9, shiny=TRUE)
```

A screenshot of a dynamic `NTplot` example is in Figure E.1.

3.4.3 The (Student's) t Distribution

The t distribution is similar to the standard normal distribution in that its density is a bell-shaped curve symmetric about 0. However, as we see in Figure 3.14, where

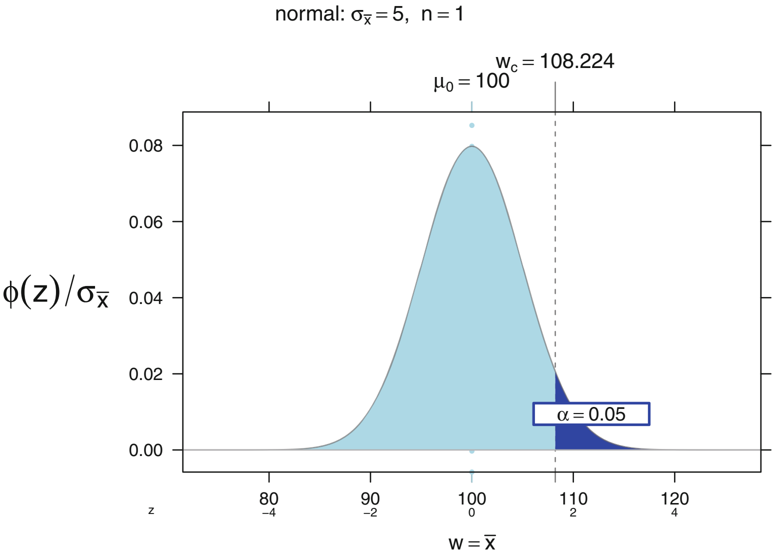


```
>   dnorm(1.645, m=0, s=1)
[1] 0.1031

>   pnorm(1.645, m=0, s=1)
[1] 0.95

>   qnorm(0.95, m=0, s=1)
[1] 1.645
```

Fig. 3.12 The standard normal density $N(0, 1)$ is shown in the top panel. The darker colored area is $\Phi(1.645) = P(Z \leq 1.645) = .95$. The lighter colored area is $1 - \Phi(1.645) = P(Z > 1.645) = .05$. The height of the density function in the top panel at $z = 1.645$ is $\phi(1.645) = .1031$. The cumulative distribution is shown in the bottom panel. The height of the darker line segment (below the curve) at $z = 1.645$ is $P(Z \leq 1.645) = .95$. The height of the lighter line segment (above the curve) at $z = 1.645$ is $P(Z > 1.645) = .05$.

we compare several t distributions to the normal distribution, the probability density function for the t is lower in the center and "heavier" in the tails. If the mean of a sample of size n is standardized with a sample standard deviation s rather than with a population standard deviation σ, then the resulting standardization, $\frac{\bar{X}-\mu}{s/\sqrt{n}}$, has a Student's t distribution with *degrees of freedom* parameter $n-1$. The t distribution is used for inference on population means and regression coefficients.

That $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ has a t distribution rests on the fact that $\bar{X}$ and s are independent random variables when sampling from a normal population.

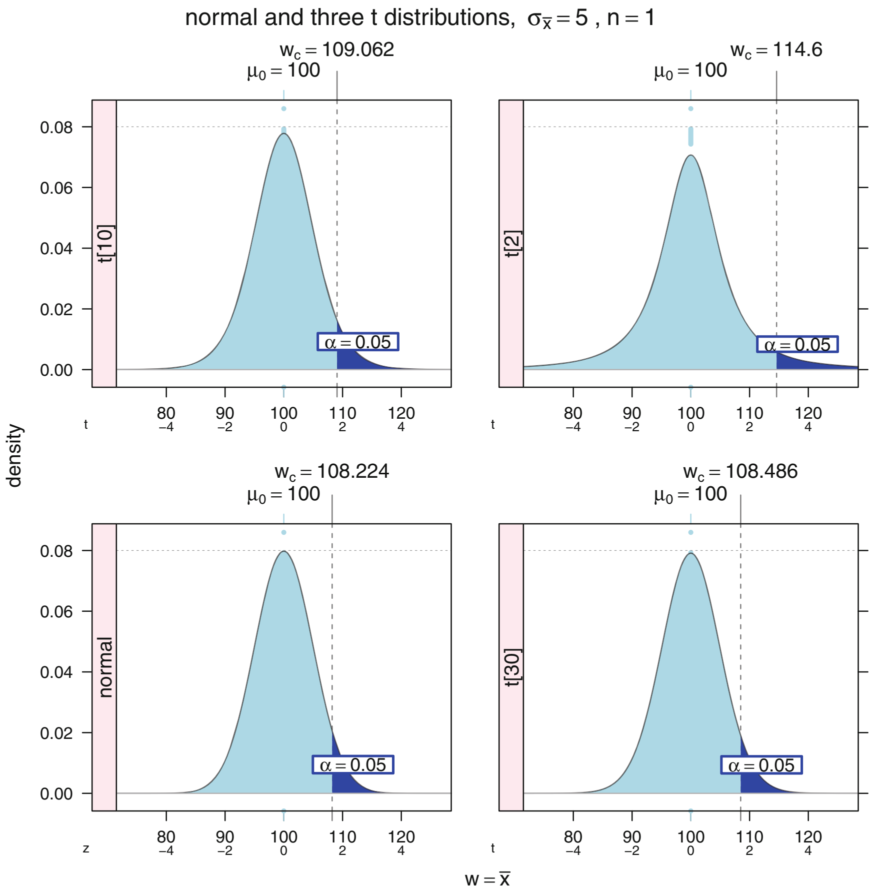


	μ_0	$w_{crit.R}$
$\bar{x}$	100	108.2
z	0	1.645

	Probability
α	0.05

Fig. 3.13 A normal curve centered on the assumed true mean $\mu = 100$. We assume $\sigma = 5$ and $\alpha = .05$. The left lightly shaded area is $.95 = P\left(z = (X-\mu)/\sigma \leq \Phi^{-1}(1-\alpha) = 1.645\right)$. The right darkly shaded area is $\alpha = .05 = P\left(z = (X-\mu)/\sigma \geq \Phi^{-1}(1-\alpha) = 1.645\right)$. The plot shows both the $\bar{x}$ scale and (in smaller font) the z scale. The table below the plot shows μ_0 and the right critical value $\bar{x}_{\text{crit.R}}$ in both scales. The critical value in the z scale is directly from the normal table.

As the sample size n and hence the degrees of freedom get large, the sample standard deviation s increasingly approximates σ so that $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ increasingly approximates $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$. In other words, as the degrees of freedom increases, a t distribution increasingly resembles a standard normal distribution.



	normal	t_{30}	t_{10}	t_2
$w_{crit.R}$	108.224	108.486	109.062	114.6
t	1.64486	1.69726	1.81246	2.91998

	Probability
α	0.05

Fig. 3.14 These panels are similar to Figure 3.13, the first panel is identical to Figure 3.13. The remaining panels show *t*-distributions with 30 df, 10 df, and 2 df. Each panel has less area in the center and more area in the tails. Use the reference line at $y = .08$ to see the drop in central area, use the thickness of the tails at $\bar{x} = 120$ to see the increase in the probability in the tails. Use the location of the critical value $w_c = \bar{x}_c$ on the graph and in the table below the graph to see that the critical value for the $\alpha = .05$ test is moving away from the null hypothesis value μ_0 as the df gets larger.

3.5 Sampling Distributions

In Chapter 1 we learn that knowledge about characteristics of populations can be gleaned from analogous characteristics of random samples from these populations. Also recall that population characteristics are called parameters and sample characteristics are called statistics. In the next two sections we discuss the two main techniques for using statistics to infer about parameters: estimation, and hypothesis testing. Implementation of these techniques requires that we use knowledge about the likely values of statistics. Such information about statistics is contained in their *sampling distribution*. The sampling distribution of a statistic depends on our assumed knowledge of the distribution of values in the population to which we are inferring. The term *standard error* is used to refer to the standard deviation of a sampling distribution.

Consider first the mean $\bar{X}$ of a sample of n items randomly selected from a normal population, $N(\mu, \sigma^2)$. It can be shown that the sampling distribution of $\bar{X}$ is also normally distributed with this same mean but with a much smaller variance:

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

We illustrate this phenomenon in Figures 3.15 and 3.16. Figure 3.15 shows the individual observations and their means. Figure 3.16 shows the distribution of the means.

In the more likely situation where σ^2 is unknown, analogous probability statements are made with reference to the Student's t distribution.

Next suppose that the population is not necessarily normal. Then under fairly general conditions, a statistical theory result called the Central Limit Theorem states that $\bar{X}$ has "approximately" a $N(\mu, \sigma^2/n)$ distribution if the sample size n is "sufficiently large". Thus, the inferential statements concerning μ made in the normal distribution case are also approximately valid if the population is not normal.

What is meant here by "approximately" and "sufficiently large"? We mean that the closer the population is to a normal population, the smaller the sample size needs to be for the approximation to be acceptably accurate. Unless the population is multimodal or severely skewed, a sample size of 30 to 50 is usually sufficient for the approximation to hold.

Another application of the Central Limit Theorem implies that the sampling distribution of the proportion $\hat{p} = X/n$ of successes in n binomial trials is approximately normally distributed with mean $\mu = np$ and variance $\sigma^2 = npq$, where $q = 1 - p$. This result is used for inferences concerning the proportion of successes in a dichotomous population where the binomial assumptions apply.

If S^2 is the variance of a random sample of size n from a normal population having variance σ^2, then the sampling distribution of $(n-1)S^2/\sigma^2$ is χ^2 with $n-1$ degrees of freedom. We use this result for inferences concerning the population standard deviation σ.

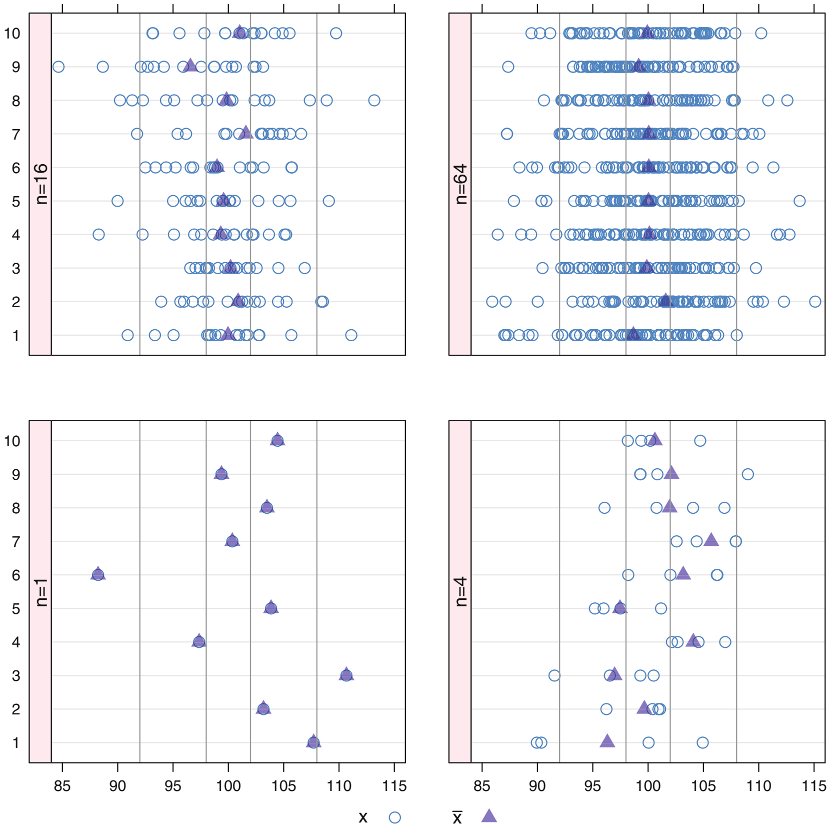
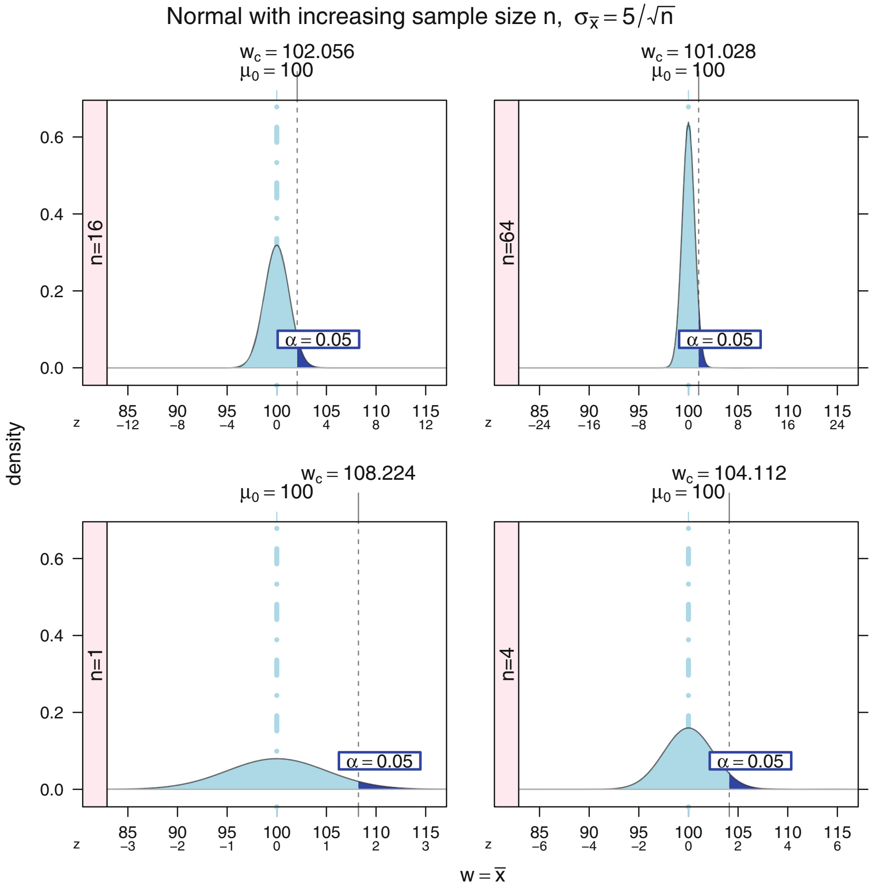


Fig. 3.15 Each panel shows 10 sets of n observations from the $N(\mu = 100,\ \sigma^2 = 5^2)$ distribution. The number n (from the set $\{1, 4, 16, 64\}$) differs by panel. The open circles show each individual observation. The semi-transparent triangle overlay shows the mean of each set of observations. As n gets larger, the set of 10 means are closer together. In the $n = 1$ panel, the means are identical to the individual observations and they occupy the full width of the panel. More precisely the variance of the means in the $n = 1$ panel is $\sigma^2 = 5^2$. In the $n = 4$ panel, the means are the average of 4 observations and they spread over only the central half of the panel with $\sigma^2_{\bar{x}} = 5^2/4$. In the $n = 16$ panel, the means are the average of 16 observations and they spread over only the central quarter of the panel with $\sigma^2_{\bar{x}} = 5^2/16$. In the $n = 64$ panel, the means are the average of 64 observations and they spread over only the central eighth of the panel with $\sigma^2_{\bar{x}} = 5^2/64$.



	n = 1	n = 4	n = 16	n = 64
$w_{crit.R}$	108.224	104.112	102.056	101.028
z	1.64486	1.64484	1.64488	1.6448
$\sigma_{\bar{x}}$	5	2.5	1.25	0.625

	Probability
α	0.05

Fig. 3.16 These panels are also similar to Figure 3.13, with both the $\bar{x}$-scale and the z-scale shown in each panel of the graph. Again the panel with $n = 1$ is identical to Figure 3.13. The remaining panels show the sampling distribution of $\bar{x}$ as n increases. Each time the sample size goes up by a multiple of 4, the distance on the $\bar{x}$-scale from the critical value $w_c = \bar{x}_c$ to μ_o is halved, and the height of the density is doubled. On the z-scale, the distance from $w_c = \bar{x}_c$ to μ_o is always exactly $z_\alpha = 1.645$.

3.6 Estimation

A fundamental task of statistical analysis is inference of the characteristics of a large population from a sample of n items or individuals selected at random from the population. Sampling is commonly undertaken because it is

a. cheaper and

b. less prone to error

than examining the entire population. Estimation is one of the two broad categories of statistical techniques used for this purpose. The other is hypothesis testing, discussed in Section 3.7.

An *estimator* is a formula that can be evaluated with numbers from the sample. When the sample values are plugged into the formula, the result becomes an *estimate*. An estimator is a particular example of a statistic.

3.6.1 Statistical Models

A key component of statistical analysis involves proposing a statistical model. A statistical model is a relatively simple approximation to account for complex phenomena that generate data. A statistical model consists of one or more equations involving both random variables and parameters. The random variables have stated or assumed distributions. The parameters are unknown fixed quantities. The random components of statistical models account for the inherent variability in most observed phenomena. Subsequent chapters of this book contain numerous examples of statistical models.

The term *estimation* is used to describe the process of determining specific values for the parameters by fitting the model to the data. This is followed by determinations of the quality of the fit, often via hypothesis testing or evaluation of an index of goodness-of-fit.

Model equations are often of the form

$$\texttt{data} = \texttt{model} + \texttt{residual}$$

where `model` is an equation that explains most of the variation in the data, and `residual`, or lack-of-fit, represents the portion of the data that is not accounted for by the model. A good-quality model is one where `model` accounts for most of the variability in the data, that is, the data are well fitted by the model.

A proposed model provides a framework for the statistical analysis. Experienced analysts know how to match models to data and the method of data collection. They are also prepared to work with a wide variety of models, some of which are discussed in subsequent chapters of this book. Statistical analysis then proceeds by estimating the model and then providing figures and tables to support a discussion of the model fit.

3.6.2 Point and Interval Estimators

There are essentially two types of estimation: point estimation and interval estimation.

A typical example begins with a sample of n observations collected from a normal distribution with unknown mean μ and unknown standard deviation σ. We calculate the sample statistics

$$\bar{x} = \left(\sum_{i=1}^{n} x_i \right) / n$$

$$s^2 = \left(\sum_{i=1}^{n} (x - \bar{x})^2 \right) / (n - 1)$$

Then $\bar{x}$ is a point estimator for μ. Define the standard error of the mean $s_{\bar{x}}$ as $s_{\bar{x}} = s/\sqrt{n}$. We then have

$$\bar{x} \pm t_{\alpha/2,\nu}\, s_{\bar{x}} = \left(\bar{x} - t_{\alpha/2,\nu}\, s_{\bar{x}},\ \bar{x} + t_{\alpha/2,\nu}\, s_{\bar{x}}\right)$$

as a two-sided $100(1 - \alpha)\%$ confidence interval for μ.

For specificity, let us look in Figure 3.17 at the situation with $n = 25$, $\bar{x} = 8.5$, $\nu = 24$, $s^2 = 4$, $\alpha = .05$. From the t-table, the critical value $t_{\alpha/2,24} = 2.064$. We get $s_{\bar{x}} = s/\sqrt{n} = 2/\sqrt{25} = .4$ as the standard error of the mean.

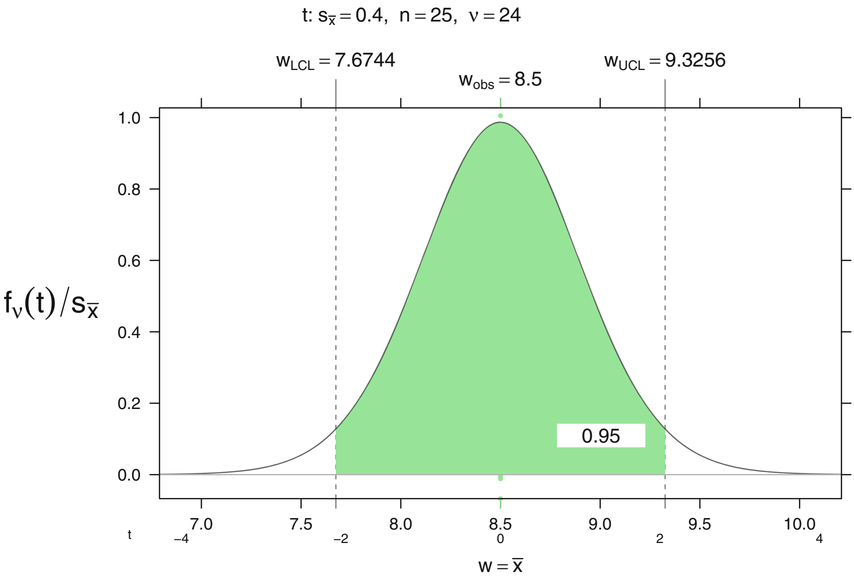
Point estimators are single numbers calculated from the sample, in this example $\hat{\mu} = 8.5$. Interval estimators are intervals within which the parameter is expected to fall, with a certain degree of confidence, in this example $95\%\,\text{CI}(\mu) = 8.5 \pm 2.064 \times 0.4 = (7.6744, 9.3256)$. Interval estimators are generally more useful than point estimators because they indicate the precision of the estimate. Often, as here, interval estimators are of the form:

$$\text{point estimate} \pm \text{constant} \times \text{standard error}$$

where "standard error" is the observed standard deviation of the statistic used as the point estimate. The constant is a percentile of the standardized sampling distribution of the point estimator. We summarize the calculations in Table 3.7.

3.6.3 Criteria for Point Estimators

There are a number of criteria for what constitutes "good" point estimators. Here is a heuristic description of some of these.



	w_{obs}	w_{LCL}	w_{UCL}
$\bar{x}$	8.5	7.674	9.326
t	0	−2.064	2.064

	Left	Confidence	Right
Probability	0.025	0.95	0.025

Fig. 3.17 Confidence interval plot for the t distribution with $n = 25$, $\bar{x} = 8.5$, $\nu = 24$, $s^2 = 4$, $\alpha = .05$. We calculate $t_{\alpha/2,24} = 2.064$ and the two-sided 95% confidence interval (7.674, 9.326). The algebra and R notation for the estimators are shown in Table 3.7.

Table 3.7 Algebra and R notation for the example in Figure 3.17.

$\bar{x}$	`> xbar <- 8.5`
s	`> s <- sqrt(4)`
n	`> n <- 25`

$s_{\bar{x}}$	`> s.xbar <- s/sqrt(n)`
	`> s.xbar`
	`[1] 0.4`

$t_{\alpha/2,24}$	`> qt(.975, df=24)`
	`[1] 2.063899`

$\bar{x} \pm t_{\alpha/2,24}\ s_{\bar{x}}$	`8.5 + c(-1,1) * 2.064 * 0.4`
	`[1] 7.6744 9.3256`

unbiasedness: The expected value of the sampling distribution of the estimator is the parameter being estimated. The bias is defined as:

$$\text{bias} = \text{expected value of sampling distribution} - \text{parameter}$$

Unbiasedness is not too crucial if the bias is small and if the bias decreases with increasing n. The sample mean $\bar{x}$ is an unbiased estimator of the population mean μ and the sample variance s^2 is an unbiased estimate of the population variance

σ^2. The sample standard deviation s is a biased estimator of the population standard deviation σ. However, the bias of s decreases toward zero as the sample size increases; we say that s is an *asymptotically unbiased* estimator of σ.

small variance: Higher precision. For example, for estimating the mean μ of a normal population, the variance $s_{\bar{x}} = s/\sqrt{n}$ of the sample mean $\bar{x}$ is less than the variance $s_{\overset{\perp}{x}} = \sqrt{\frac{\pi}{2}}\, s/\sqrt{n}$ of the sample median $\overset{\perp}{x}$.

consistency: The quality of the estimator improves as n increases.

sufficiency: the estimator fully uses all the sample information. Example: If X is distributed as continuous uniform on $[0,a]$, how would you estimate a? Since the population mean is $a/2$, you might think that $2\bar{x}$ is a "good" estimator for a. The largest item in the sample of size n, denoted $x_{(n)}$, is a better and *sufficient* estimator of a. This estimator cannot overestimate a while $2\bar{x}$ can either underestimate or overestimate a. If $x_{(n)}$ exceeds $2\bar{x}$, then it must be closer to a than is $2\bar{x}$.

3.6.4 Confidence Interval Estimation

A confidence interval estimate of a parameter is an interval that has a certain probability, called its *confidence coefficient*, of containing the parameter. The confidence coefficient is usually denoted $1-\alpha$ or as a percentage, $100(1-\alpha)\%$. Common values for the confidence coefficient are 95% and 99%, corresponding to $\alpha = .05$ or .01, respectively. Figure 3.17 illustrates a 95% confidence interval for the mean of a normal distribution.

If we construct a 95% confidence interval (CI), what is the meaning of 95%? It is easy to incorrectly believe that 95% is the probability that the CI contains the parameter. This is false because the statement *"CI contains the parameter"* is not an event, but rather a situation that is certainly either true or false. The correct interpretation refers to the *process used to construct the CI:* If, hypothetically, many people were to use this same formula to construct this CI, plugging in the results of their individual random samples, about 95% of the CI's of these many people would contain the parameter and about 5% of the CI's would exclude the parameter.

It is important to appreciate the tradeoff between three quantities:

- confidence coefficient (the closer to 1 the better)
- interval width (the narrower the better)
- sample size (the smaller the better)

In practice it is impossible to optimize all three quantities simultaneously. There is an interrelationship among the three so that specification of two of them uniquely determines the third. A common practical problem is to seek the sample size required to attain a given interval width and confidence. Examples of such formulas appear in Section 5.6.

3.6.5 Example—Confidence Interval on the Mean μ of a Population Having Known Standard Deviation

The interpretation of the confidence coefficient may be further clarified by the following illustration of the construction of a $100(1-\alpha)\%$ confidence interval on an unknown mean μ of a normal population having known standard deviation σ, using a random sample of size n from this population. If $\bar{X}$ denotes the sample mean, then $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ has a standard normal distribution. Let $z_{\frac{\alpha}{2}}$ denote the $100(1-\frac{\alpha}{2})^{\text{th}}$ percentile of this distribution. Then

$$P\left(-z_{\frac{\alpha}{2}} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

After a bit of algebraic rearrangement, this becomes

$$P\left(\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

The endpoints of the interval $\left(\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$ are random variables, so the probability statement refers to the probability that the interval contains the parameter, not the probability that the parameter is contained in the interval.

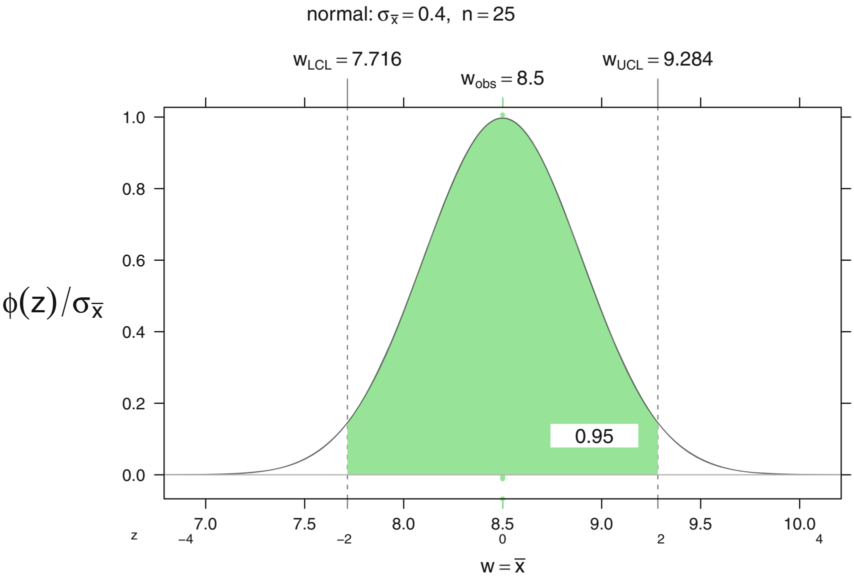
In practice, we replace the random variable $\bar{X}$ with $\bar{x}$, the realized value from the sample, and wind up with the $100(1-\alpha)\%$ confidence interval for μ:

$$\left(\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \tag{3.18}$$

Figure 3.18 shows an example from the situation with known variance σ^2.

3.6.6 Example—One-Sided Confidence Intervals

One-sided confidence intervals correspond to one-sided tests of hypotheses. Such intervals have infinite width and therefore are much less commonly used in practice than two-sided confidence intervals, which have finite width. The rationale for using



	w_{obs}	w_{LCL}	w_{UCL}
$\bar{x}$	8.5	7.716	9.284
z	0	−1.96	1.96

	Left	Confidence	Right
Probability	0.025	0.95	0.025

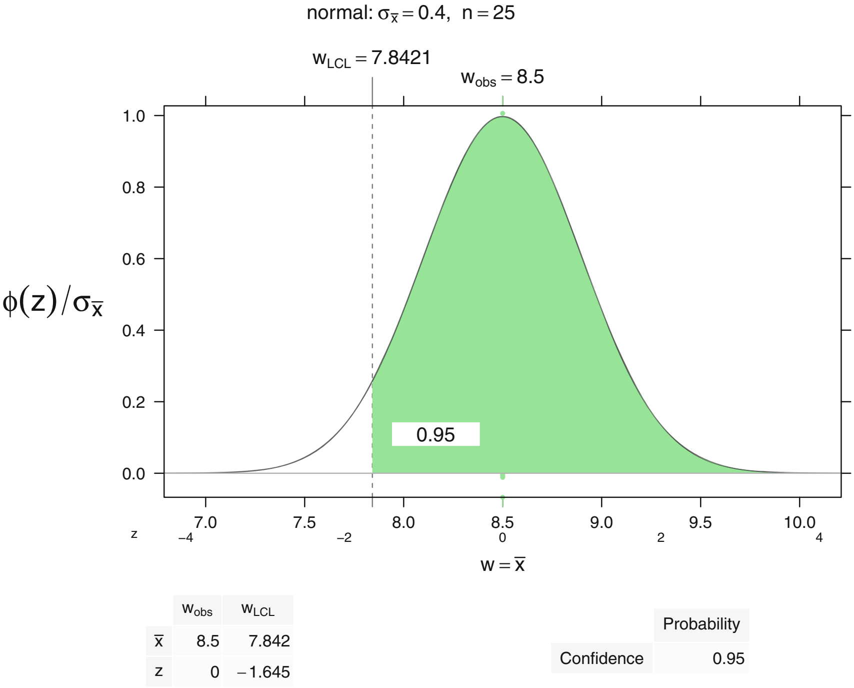
Fig. 3.18 Confidence interval plot for the normal distribution with $n = 25$, $\bar{x} = 8.5$, $\sigma^2 = 4$, $\alpha = .05$. We calculate $z_{\alpha/2} = 1.96$ and the two-sided 95% confidence interval (7.716, 9.284). Compare this to the t-based confidence interval in Figure 3.17 and note that the width of the interval is narrower here because we have more information, that is, because we know the variance, we don't have to estimate the variance.

one-sided intervals matches that for one-sided tests—sometimes the analyst believes the value of a parameter is at least or at most some value rather than on either side. One-sided confidence intervals on the mean of a population having known standard deviation are shown in Table 5.1. Other examples of one-sided confidence intervals appear in Tables 5.2 and 5.3. Figure 3.19 shows a one-sided example from the situation with known variance σ^2.

3.7 Hypothesis Testing

The statistician sets up two competing hypotheses, the null hypothesis H_0 and the alternative hypothesis H_1, for example in Figure 3.21 in Section 3.8,

$H_0\colon \mu = 32$ vs $H_1\colon \mu \neq 32$. The task is to decide whether the sample evidence better supports H_0 (decision to "retain H_0") or H_1 (decision to "reject H_0").



	w_{obs}	w_{LCL}
$\bar{x}$	8.5	7.842
z	0	−1.645

	Probability
Confidence	0.95

Fig. 3.19 One-sided confidence interval plot for the normal distribution with $n = 25$, $\bar{x} = 8.5$, $\sigma^2 = 4$, $\alpha = .05$. We are confident that the true mean is larger than the calculated value. We calculate $z_\alpha = 1.645$ and the one-sided 95% confidence interval $(7.842, \infty)$.

There are two types of errors: the Type I error of *rejecting* H_0 when H_0 is true, and the Type II error of *retaining* H_0 when H_1 is true. In the classical hypothesis setup, the statistician prespecifies α—the maximum probability of committing a Type I error. Subject to this constraint, we select a testing procedure that gives good control over β—the probability of committing a Type II error. This probability is a function of the unknown parameter being tested. A plot of the probability against the parameter is called an *operating characteristic curve* (O.C. curve) of the test.

The *power* of a hypothesis test is the probability of correctly rejecting a false null hypothesis, equivalently, the probability $1-\beta$. A *power curve* is a plot of the probability of rejecting H_0 against the true value of the parameter. It contains information identical to that conveyed by an O.C. curve. It is a convention in various scientific fields whether the power or the O.C. curve is used. We illustrate both in Figure 3.20. Statisticians can determine the sample size needed to conduct a test that has a high probability of detecting a departure from H_0 by studying O.C. or power curves for a variety of proposed sample sizes. Examination of these curves displays the tradeoffs between Type I error, Type II error, and sample size. See Figure 3.20 for a static example. The reader can explore these options dynamically with the `shiny=TRUE` argument to the `NTplot` function. Figure E.1 shows a screenshot of our **shiny** app duplicating Figure 3.20. Further discussion of Operating Characteristic and power curves is in Section 3.9.

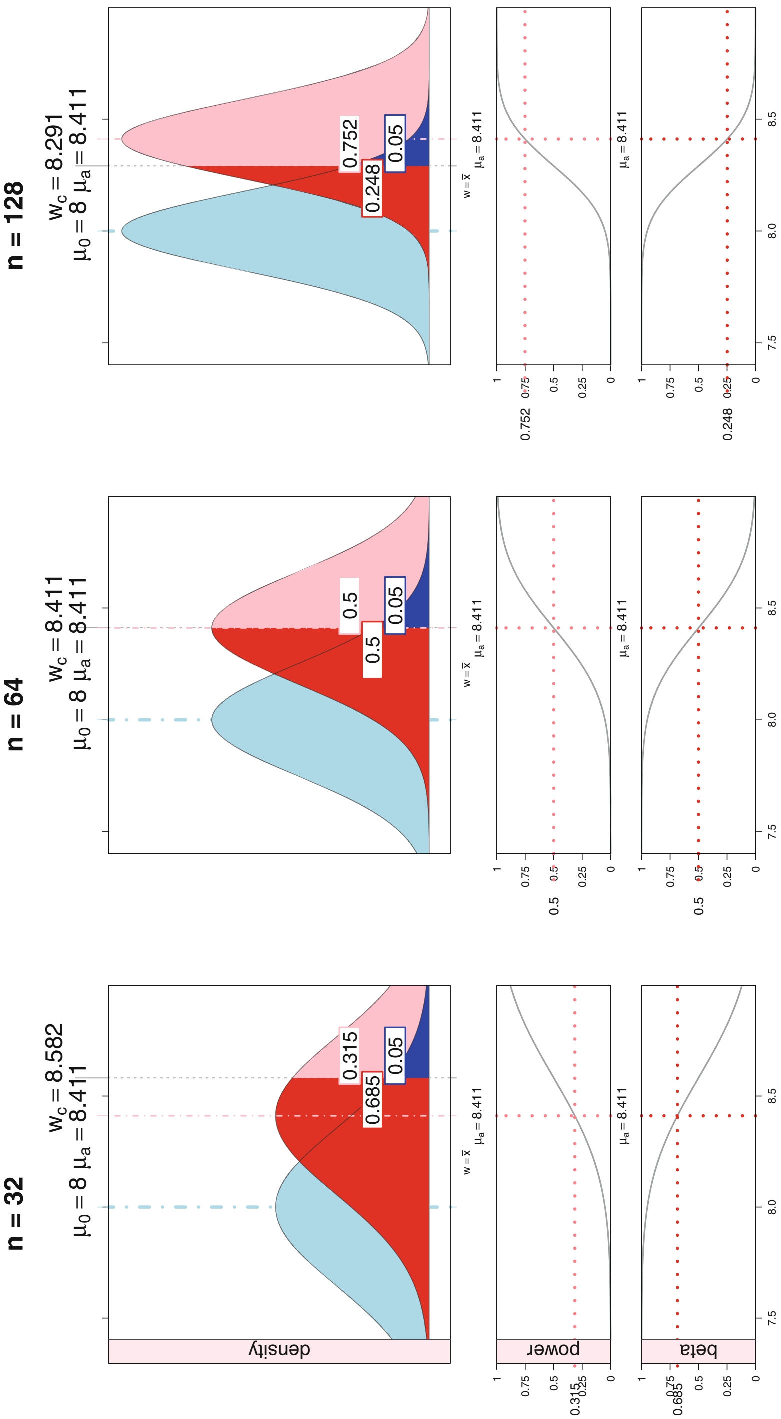

n = 32
$\mu_0 = 8$ $\mu_a = 8.411$
$w_c = 8.582$
0.315
0.685
0.05
density
power
beta
0.315
0.685
n = 64
$\mu_0 = 8$ $\mu_a = 8.411$
$w_c = 8.411$
0.5
0.5
0.05
n = 128
$\mu_0 = 8$ $\mu_a = 8.411$
$w_c = 8.291$
0.752
0.248
0.05
0.752
0.248
$w = \bar{x}$
$\mu_a = 8.411$
1
0.75
0.5
0.25
0
7.5
8.0
8.5

Fig. 3.20 (continued)

The three sets of panels show the same null hypothesis ($\mu_0 = 8$) and alternative hypothesis ($\mu_1 = 8.411$) with three different sample sizes ($n = 32, 64, 128$) and their corresponding powers (.315, .500, .752). Each set contains a normal plot in the top panel, the corresponding power curve in the middle panel, and the beta curve (operating characteristic curve) in the bottom panel.

The pink area in each top panel shows the power, the probability that an observed $\bar{x}$ will be to the right of the critical value $\bar{x}_C = 8.411$ when the true mean is $\mu_1 = 8.411$. The gray curve in each middle panel is the power curve, showing the power for all possible values of the alternate mean μ_1. The crosshairs in the middle panel are at $\mu_1 = 8.411$ and power($\mu_1 = 8.411$). The red area in the top panels shows $\beta = 1 -$ power, the probability of the Type II Error. The gray curve in each bottom panel is the beta curve (the Operating Characteristic curve) showing the β for all possible values of the alternate mean μ_1. The crosshairs in the bottom panel are at $\mu_1 = 8.411$ and beta($\mu_1 = 8.411$). As we increase the sample size (move from the left set of panels toward the right set of panels), the density functions get taller and thinner while maintaining a constant area of 1, the power and beta curves get steeper, and the power increases (hence beta decreases) for any specified value of μ_1. The reader can duplicate these panels by running the R code in file `HHscriptnames(3)`. The reader can set up a dynamic version of this plot from the same code with `NTplot(tmp64, shiny=TRUE)` and then clicking the animate icon for the n-slider. See Figure E.1 for a screenshot. The screenshot initially doesn't show the Power and Beta curves. They can be included by checking the Power and Beta checkboxes on the Display Options tab.

Do not confuse the decision to retain H_0 with the statement that H_0 is true. We might be committing a Type II error. Similarly, the decision to reject H_0 is not the same as saying that H_0 is false because we might be committing a Type I error.

Commonly selected values of α are .05 or .01. The choice is sometimes governed by what is traditional in a research area.

With the prespecification of α, the statistician maintains better control over Type I error than Type II error. When we have a choice, the names H_0 and H_1 should be assigned such that the hypothesis with the more serious error is called H_0 and its more serious error is the Type I error. The hypothesis with the less serious error is called H_1 and its less serious error is the Type II error. In many applications, H_0 is essentially the statement that the status quo is better, while H_1 is the statement that an innovation is better. The Type I error of incorrectly deciding in favor of an innovation is typically more serious than the error of incorrectly maintaining the status quo because innovation is usually costly. As a result, classical testing puts the burden of proof on the innovation H_1; H_0 is retained unless there is compelling evidence not to do so.

The preceding rules for deciding which hypothesis is H_0 are based on the fact that classical hypothesis testing places more control over Type I error at the cost of reduced control over Type II error. The logic for this approach is seen by comparing in Table 3.8 the definitions of these two errors in the hypothesis testing context with the potential errors in a U.S. courtroom.

Table 3.8 Comparison of Hypothesis Testing with the Decision Options in a Court of Law

Hypothesis Testing			Court of Law		
	True situation			True situation	
Decision	H_0 true	H_0 false	Decision	Innocent	Guilty
Reject H_0	Type I error	correct	Convict	greater error	correct
Retain H_0	correct	Type II error	Acquit	correct	lesser error

In the United States, the error of convicting an innocent defendant is viewed as far more serious than the error of acquitting a guilty defendant. Accordingly, the U.S. legal system places the burden of proof on the prosecution to establish guilt beyond a reasonable doubt. If sufficient evidence is not presented to the court, the defendant is acquitted. Similarly, in hypothesis testing, the burden is placed on the analyst to provide convincing evidence that H_0 is false; in the absence of such evidence, H_0 is accepted. Continuing the analogy, in the hypothesis testing framework, the way to reduce the probability of committing a Type II error without compromising control of Type I error is to seek an increased sample size. In the legal framework, courts can best reduce the probability of acquitting guilty defendants by obtaining as much relevant evidence as possible.

Table 3.8 also demonstrates that if we modify a hypothesis testing procedure to less readily reject a null hypothesis, this results in both greater control of Type I error and reduced control of Type II error.

Tests of hypotheses are conducted by determining what sample results would be likely if H_0 is true. If then a sufficiently unlikely sample statistic is observed, doubt is cast on the truth of H_0; i.e., H_0 is rejected.

Most tests are constructed by calculating a test statistic from a random sample. This is compared to a critical value, or values. If the test statistic is on one side of the critical value(s), H_0 is retained; if on the other side, H_0 is rejected. If the value of the test statistic leads to rejection of H_0, the test statistic is said to be (statistically) *significant*.

A criticism of classical hypothesis testing is the requirement that α be prespecified. One way around this is to calculate the p-value of the test.

> The p-value is the probability of observing, in hypothetical repeated samples from the null distribution (that is, when H_0 is true), a value of the test statistic at least as extreme in the direction of H_1 as the test statistic calculated from the present sample.

For most testing procedures, calculating the p-value requires the use of the computer. We reject H_0 (that is, we make the decision to act as if H_0 does not describe the world) if $\alpha > p$-value; we retain H_0 (that is, we make the decision to act as if H_0 does describe the world) otherwise. Then the analyst needs only to know how α compares with the p-value, and does not have to commit to a particular value of α. Most software provides p-values as part of the output rather than requesting α as part of the input.

Another criticism of classical hypothesis testing is that if H_0 is barely false, it is always possible to reject H_0 simply by taking a large enough sample size. For example, if we test $H_0\colon \mu = 32$, where μ is the mean amount of soda a bottling plant puts into 32-ounce (0.946 liter) bottles, and if in reality, $\mu = 32.001$ ounces, H_0 can be rejected even though as a practical matter it makes no sense to act as though anything is wrong with the filling mechanism. This would be an instance of a statistically significant result that is not of practical significance. Because of this criticism, many statisticians are much more comfortable using CIs than tests.

In practice, a very small p-value may be regarded as sufficiently strong evidence against H_0 to convince us to act as though H_0 is false (that is, as though H_0 does not describe the world). However, even in this situation and especially if the sample size is large, we should be mindful of the possibility that one is making a Type I error. Also, we should always be alert to the possibility that an underlying assumption about the population is incorrect; if so, the p-value calculation may be distorted.

3.8 Examples of Statistical Tests

Suppose in the example of the previous section, the standard deviation of fill volume is known to be 0.3 ounces, and that a sample of 100 bottles yields a mean of 31.94 ounces. If the alternative hypothesis is $H_1\colon \mu \neq 32$, then we should reject H_0 if $\bar{x}$ is sufficiently above or below 32. We illustrate this example in Figure 3.21. In this example, in order to maintain Type I error probability at $\alpha = .01$, we should reject H_0 if

$$\begin{aligned} \bar{x} &< 32 - z_{.005}\ \sigma\ / \sqrt{n} \\ &= 32 - 2.576\,(0.3)/\ 10 = 31.923 \end{aligned}$$

or

$$\begin{aligned} \bar{x} &> 32 + z_{.005}\ \sigma\ / \sqrt{n} \\ &= 32 + 2.576\,(0.3)/\ 10 = 32.077 \end{aligned}$$

Since $\bar{x}$ meets neither condition, we should retain H_0 when testing at $\alpha = .01$. This is an example of a "two-tailed" (or "two-sided") test because we reject H_0 if $\bar{x}$ lies sufficiently far on either tail of the Z distribution with the null hypothesized mean.

At this point we might ask whether a larger choice of α would have led to the "retain H_0" decision. This is answered by finding the p-value, here equal to $2P(Z > |z_{\text{calc}}|)$ for $z_{\text{calc}} = (\bar{x} - \mu_0)/(\sigma/\sqrt{n}) = -2$. Thus p-value$= 2P(Z > 2.00) = 0.046$. Then any choice of $\alpha \leq 0.046$ requires retention of H_0; i.e., the decision to act as if the filling machine is in control.

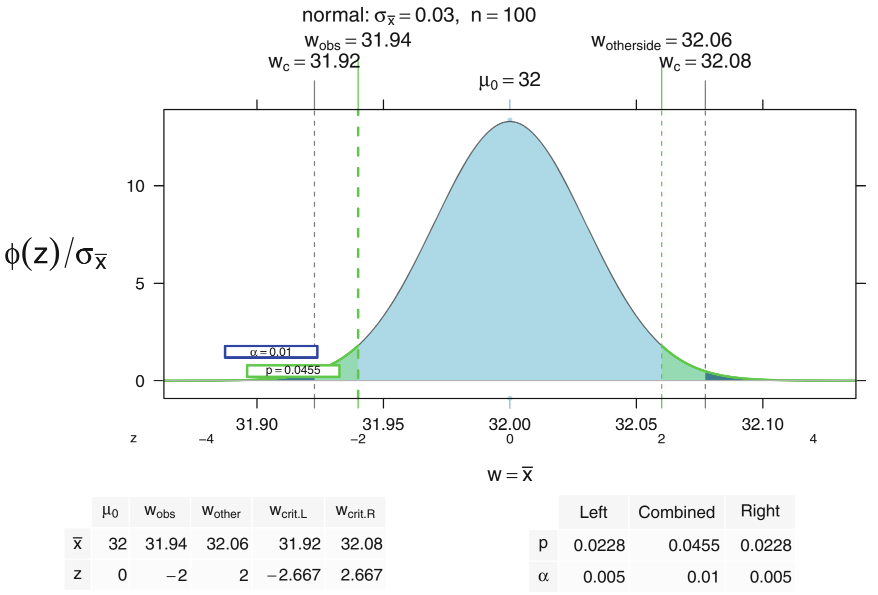
A two-tailed test can be conducted as follows. Reject the null hypothesis at level α if the null hypothesized value of the parameter lies outside the $100(1 - \alpha)\%$ confidence interval for the parameter.

Sometimes analysts prefer to conduct a "one-tailed" (or "one-sided") test where the alternative hypothesis statement is a one-sided inequality. Suppose in the soda bottling example it was felt that the error of incorrectly claiming bottles are being underfilled is much more serious than an error of incorrectly claiming bottles are being overfilled. We illustrate the one-tailed test in Figure 3.22. Then we might test $H_0\colon \mu \geq 32$ vs $H_1\colon \mu < 32$, because this way the more serious error is the better controlled Type I error. Now H_0 will be rejected only when $\bar{x}$ is sufficiently below 32. If once again we take $\alpha = .01$, we reject H_0 if

$$\begin{aligned} \bar{x} &< 32 - z_{.01}\ \sigma\ / \sqrt{n} \\ &= 32 - 2.326\,(0.3)/\ 10 = 31.93 \end{aligned}$$

As with the two-tailed test, H_0 is retained.

Note that, if instead we had observed $\bar{x} = 31.925$ ounces, we would have rejected H_0 with the one-tailed alternative but retained it with the two-tailed alternative. The explanation for this distinction is that the portion of the left side of the parameter space where H_1 is true is larger under the one-tailed setup than under the analogous two-tailed setup.



	μ$_0$	w$_{obs}$	w$_{other}$	w$_{crit.L}$	w$_{crit.R}$
x̄	32	31.94	32.06	31.92	32.08
z	0	−2	2	−2.667	2.667

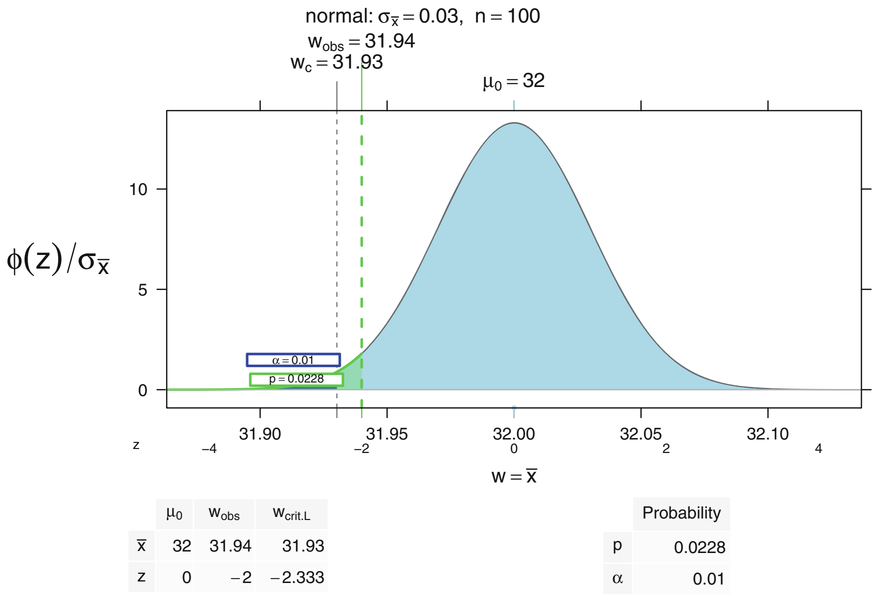
	Left	Combined	Right
p	0.0228	0.0455	0.0228
α	0.005	0.01	0.005

Fig. 3.21 Test whether the bottle production is within bounds. The figure shows a two-sided rejection region—anything in the deep blue region outside the critical bounds ($\bar{x}_{\text{critLeft}}$ = 31.92, $\bar{x}_{\text{critRight}}$ = 32.08). The observed value $\bar{x}$ = 31.94 is within the central light-blue do-not-reject region. The p-value is the green shaded area outside the bounds ($\bar{x}$ = 31.94, $\bar{x}_{\text{otherside}}$ = 32.06) where $\bar{x}_{\text{otherside}} = \mu_0 + (\mu_0 - \bar{x}) = 32.06$ is the value equally far from the null value $\mu_0 = 32$ in the other direction.

3.9 Power and Operating Characteristic (O.C.) (Beta) Curves

These two types of curves are used to assess the degree of Type II error control of a proposed test. The O.C. curve is a plot of the probability of retaining H_0 under the condition of a specified value of the parameter vs the specified value of the parameter being tested, and the power curve is a plot of the probability of rejecting H_0 vs the parameter being tested. These two plots give equivalent information, and the choice of which to use is a matter of taste or tradition in one's discipline.

Power and O.C. curves are used to display the menu of competing choices of sample size, α, and Type II error probability. One desires that all three of these quantities be as small as possible, but fixing any two of them uniquely determines the third. Analysts commonly use one of these curves to assess the needed sample size to achieve desired control over the two errors. If the required sample size is infeasibly large, the analyst can see what combinations of diminished control over the two errors are possible with the maximum attainable sample size. Note that β = P(Type II error) is a function of the true value of the unknown parameter being tested and that α is the *maximum* probability of committing a Type I error.



	μ_0	w_{obs}	$w_{crit.L}$
$\bar{x}$	32	31.94	31.93
z	0	−2	−2.333

	Probability
p	0.0228
α	0.01

Fig. 3.22 Test whether the bottle production is within bounds. The figure shows a one-sided rejection region—anything in the deep blue region below the limit $\bar{x}_c = 31.93$. The observed value $\bar{x} = 31.94$ is in the right light-blue do-not-reject region. The p-value is the green shaded area to the left of $\bar{x} = 31.94$).

In the case discussed above, $\beta = P(\text{Type II error}|\mu_a)$ is a function of the true (and unknown) value μ_a of the parameter.

We illustrate the formulation of an O.C. curve and its construction using R. The `pnorm` function calculates the normal c.d.f. Φ. The `qnorm` function calculated the inverse normal c.d.f. Φ^{-1}.

Consider a situation where we have a normal population with unknown mean μ and known s.d. $\sigma = 2.0$. Suppose we wish to test $H_0\colon \mu \le 8$ vs $H_1\colon \mu > 8$, using $\alpha = .05$ and a sample of $n = 64$ items. Here we retain H_0 if

$$\begin{aligned}\bar{X} &\le \mu_0 + \Phi^{-1}(.95)\ \sigma/\sqrt{n}\\ &= 8 \ \ + \quad 1.645 \quad 2/8\\ &= 8.411\end{aligned}$$

i.e., H_0 is retained if $\bar{X} \le 8.411$. Since the true μ is unknown, the probability that H_0 is retained is a function of this μ:

$$\begin{aligned}P(\bar{X} \le 8.411 \mid \mu) &= P\left[\tfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \tfrac{8.411-\mu}{(2/8)}\right]\\ &= P\left[Z \le 4(8.411 - \mu)\right]\\ &= \Phi(33.644 - 4\mu)\end{aligned}$$

where Z is $N(0, 1)$. The power curve for this problem is the plot of $1 - \Phi(33.644 - 4\mu)$ vs μ. Figure 3.23 shows the normal plot under both the null and alternative hypotheses for several values of μ_1, and the associated power plot and beta (Operating Characteristic) plots. The power and beta curves in all three columns of Figure 3.23 are identical. The crosshairs identify the location on the curves of the power and probability of the Type II error for the specified value μ_a of the alternative.

For most distributions, tests of hypotheses, calculation of Type II error probabilities, and construction of O.C. and power curves involves the use of a *noncentral* probability distribution. Noncentral distributions are discussed in Section J.2 in Appendix J. Noncentrality is not an issue for tests using the normal distribution, as the normal does not have a noncentral form.

We illustrate a noncentral alternative distribution in Figure 3.24.

The `power.t.test` function is essentially the same as the right panel in Figure 3.24, the only difference is that `power.t.test` assumes $\mu_0 = 0$.

```
PowerT <- power.t.test(n=12, sd=2, delta=1.4,
                       type="one.sample",
                       alternative="one.sided")
NTplot(PowerT, beta=TRUE, power=TRUE)
```

3.10 Efficiency

Efficiency is a measure of value (usually information in Statistics) per unit cost. We wish to maximize efficiency. We want small sample sizes because each observation has a cost, and fewer observations cost less than more observations. We want larger sample sizes because that gives us a better estimate of the precision of our study. A larger sample size increases the degrees of freedom for the error term. When we look at a table of t- or F- or χ^2-values we see that the critical value of the test statistics for a specified significance level is smaller as the sample size increases. We can see this in many of the figures in this chapter. Figure 3.20 shows that the critical value for a normal test goes down as the sample size goes up. Figure 3.14 shows that the critical value is smaller as the degrees of freedom increase. Choosing the right sample size is therefore important. It needs to be large enough that there is information about the population, and small enough that the client is willing to pay for the observations.

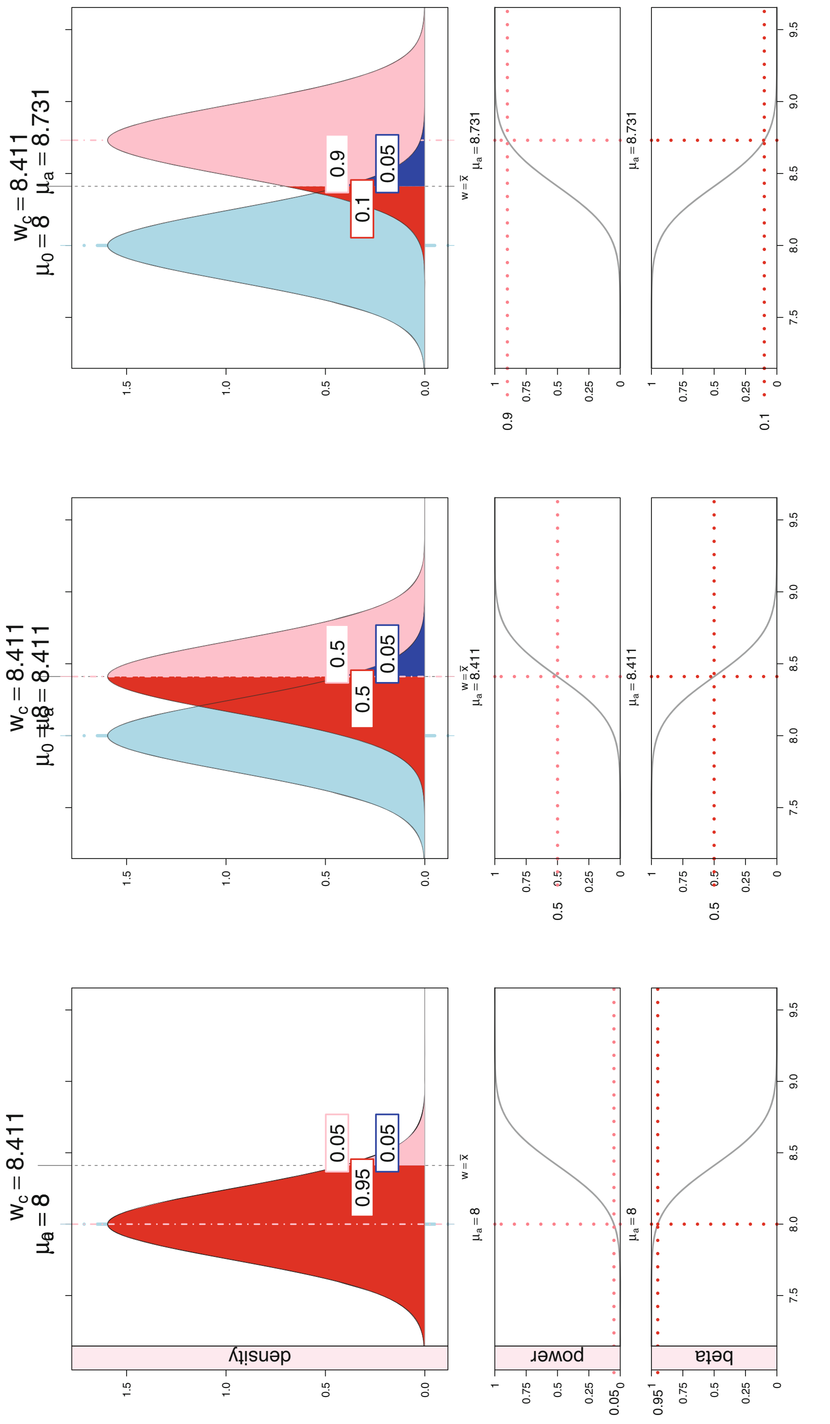

$w_c = 8.411$
$\mu_a = 8$
density
power
beta
0.05
0.95
0.05
$w = \bar{x}$
$\mu_a = 8$
$\mu_a = 8$
0.05
0.95
$w_c = 8.411$
$\mu_0 = 8$
$\mu_a = 8.411$
0.5
0.5
0.05
$w = \bar{x}$
$\mu_a = 8.411$
$\mu_a = 8.411$
0.5
0.5
$w_c = 8.411$
$\mu_0 = 8$
$\mu_a = 8.731$
0.9
0.1
0.05
$w = \bar{x}$
$\mu_a = 8.731$
$\mu_a = 8.731$
0.9
0.1
0.0
0.5
1.0
1.5
0
0.25
0.5
0.75
1
7.5
8.0
8.5
9.0
9.5

Fig. 3.23 (continued)

The three sets of panels show the same null hypothesis ($\mu_0 = 8$) with a sequence of alternative hypothesis values ($\mu_1 = 8$, $\mu_1 = 8.411 = \mu_c$, $\mu_1 = 8.7314$) and their corresponding powers (.05, .5, .90). Each set contains a normal plot in the top panel, the corresponding power curve in the middle panel, and the beta curve in the bottom panel.

Start on the right with $\mu_1 = 8.7314$, power = .90, and $\beta = .10$. In this rightmost set, the μ_1 and μ_0 are far enough apart that there is less overprinting in the top axis. The pink area in the top panel shows the power, the probability that an observed $\bar{x}$ will be to the right of the critical value $\bar{x}_C = \mu_1 = 8.411$ when the true mean is $\mu_1 = 8.7314$. The gray curve in the middle panel is the power curve, showing the power for all possible values of the alternate mean μ_1. The crosshairs in the middle panel on the right are at $\mu_1 = 8.7314$ and power($\mu_1 = 8.7314$) = .9. The red area in the top panel shows $\beta = 1 -$ power, the probability of the Type II Error. The gray curve in the bottom panel is the beta curve (the Operating Characteristic curve) showing the β for all possible values of the alternate mean μ_1. The crosshairs in the bottom panel on the right are at $\mu_1 = 8.7314$ and beta($\mu_1 = 8.7314$) = .1.

Move to the center set of panels. Now the $\mu_0 = \bar{x}_c = 8.411$ and the power is exactly 1/2. In the top axis, we see that the labels μ_0 and μ_1 are close together and partially obscure each other. In the bottom panel, the $\beta = 1 - 1/2 = 1/2$.

Move to the left panels where $\mu_0 = \mu_1 = 8$ and note that the power is equal to $\alpha = .05$ and $\beta = 1 - \alpha = .95$. The value and labels μ_0 and μ_1 are now identical and the coloring for the alternative hypothesis regions completely masks the coloring for the null hypothesis regions.

The reader can duplicate these panels by running the R code in file `HHscriptnames(3)`. The user can set up a dynamic version of this plot from the same code with `NTplot(tmp8411, shiny=TRUE)` and then clicking the animate icon for the mu[a]-slider. The user can also get there from the example in Figure E.1. On the Display Options tab, check Power and Beta. On the Normal and t tab, click the ▶ button on mu[a].

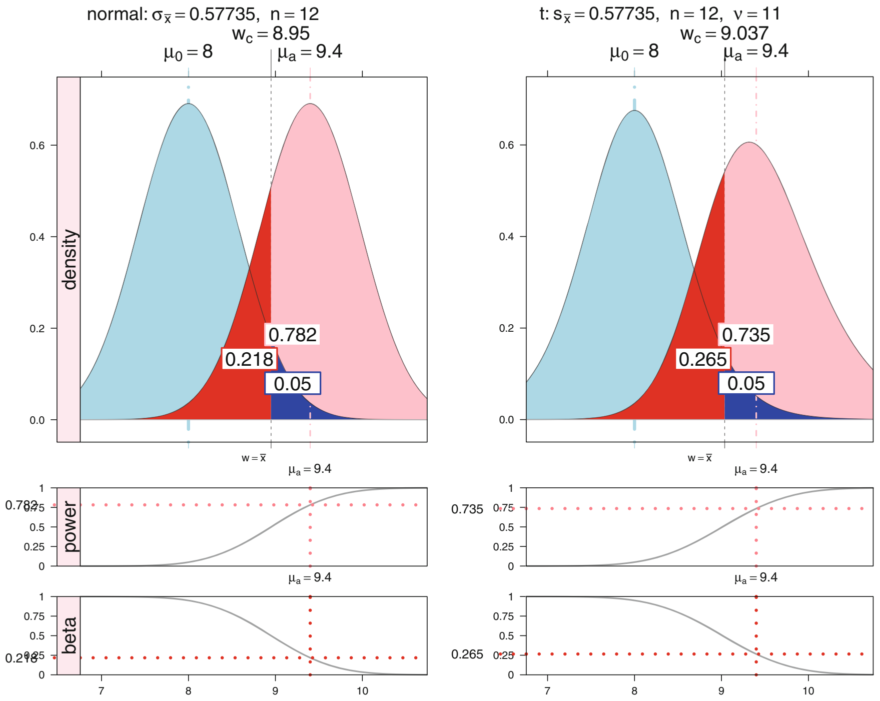


Fig. 3.24 The t-test of the same null hypothesis ($\mu_0 = 8$) as Figure 3.23 and alternative hypothesis values ($\mu_1 = 9.4$). On the left, under the assumption of known variance which implies that the density curve for the alternative is also normal with the same variance as under the null, the power is .782. On the right, under the assumption of unknown variance which requires that s must be estimated from the data, the alternative distribution has a noncentral t distribution. The null has a smaller central peak value and larger critical value. The alternative is no longer symmetric and has an even smaller peak value. See further discussion of the noncentral t distribution in Section J.2.2.

3.11 Sampling

Whenever we wish to learn the characteristics of a large *population* or *universe* that is unwieldy or expensive to completely examine, we may instead select a *sample* from the population. If the sample has been selected by a *random* mechanism, it is usually possible to infer population characteristics from the analogous characteristics in the sample. Much of the remainder of this volume deals with methods for conducting such inferences. In this section we discuss methods for selecting random samples. Only rarely is it practical to sample the entire population; such a sample is called a *census* of the population.

Here are some examples of situations where we would learn about a population by choosing a random sample from it.

- A factory wishes to know if the proportion of today's output that is defective is sufficiently small that the output may be shipped for sale rather than scrapped. Examining the entire output stream is likely to be impractical and expensive, and clearly impossible if examining an item results in its destruction. Instead, a quality-control worker may suggest a random sample of the output, with a size of sample that is sufficient to accurately estimate the proportion of defectives without being excessively costly. [Formula (5.17) may be used for determining the sample size in this situation.]
- A candidate for statewide political office wants to assess whether more than half of the electorate will vote for her. An accurate estimate of the proportion favoring her would greatly influence her future campaign strategy. She obviously must contract for a sample because her campaign cannot afford to contact all potential voters. A complication in this situation is that the population of voters and their opinions are apt to be somewhat different on election day from what they are at the time the sample is selected.
- A timber company wishes to estimate the average height of the trees in a forest under its control. Such measurements are expensive to obtain because they involve sighting a tree's top at a fixed ground distance from the tree. Therefore, a census of the forest would be prohibitively expensive and some type of random sample of trees is preferred.

If an arbitrary sample (essentially any procedure that isn't based on a specified probability distribution) is used, there is no guarantee that it will truly represent the population. To ensure that the sample adequately reflects the population, a randomization mechanism must be used. The techniques for inferring from sample to population discussed in the following chapters rest on the assumption that samples are randomly selected. If this assumption is unjustified, the probability-based statements that accompany the inferences will be incorrect.

For a given sample size n the analyst seeks to maximize the likely precision of the inference from sample to population while minimizing the cost of selecting and using the sample information. The most straightforward random sampling plan is termed *simple random sampling*. Sometimes, however, a different sampling plan can afford greater precision, or lower cost, or be easier to administer. We discuss simple random sampling and several commonly used alternatives.

3.11.1 Simple Random Sampling

A simple random sample of size n from a population of size N is one selected according to a mechanism guaranteeing that each of the $\binom{N}{n}$ potential samples have the same probability, $1/\binom{N}{n}$, of being the sample actually selected.

If, as is usually the case, the population is already identified with a numbering from 1 to N, or if it is easy to set up such a numbering, then statistical software can be used to select n distinct integers in the range 1 to N so that all potential selections are equally likely to occur.

Such a sample is easily produced in R with the statement `sample(N, n)`. If the population is not numbered but exists as a character vector `x` [where $n \leq$ `length(x)`], then `sample(x, n)` produces the required sample from `x`.

3.11.2 Stratified Random Sampling

Sometimes the population of interest is meaningfully partitioned into groups, called *strata* in the sampling literature. For example, in a school situation the strata could be individual classrooms. In addition to making inferences about the entire population, it is also desired to learn about each *stratum* (the singular of *strata*). When this is the case, we may wish to select a random sample within each stratum. Then sample estimates are available for each stratum, and these can be combined into estimates for the entire population.

Suppose there are k strata and the number of population items in stratum i is $N_i, i = 1, \ldots, k$, where $\sum_{i=1}^{k} N_i = N$. The analyst then needs to decide how many of the n total sample items should be selected from stratum i. One popular possibility, called *proportional allocation*, stipulates sampling $n_i = \left(\frac{N_i}{N}\right) n$ items from the i^{th} stratum. Since n_i need not be an integer, it is customary to round this calculation to the nearest integer. The mean estimated from the stratified random sample is $\bar{x}_{\text{ST}} = \frac{1}{N} \sum_i N_i \bar{x}_i$, i.e., a weighted average of the stratum sample means using the relative strata sizes as weights.

As an example, suppose it is desired to estimate the average annual malpractice premium paid by physicians licensed to practice in Pennsylvania. Since the risk of malpractice differs across medical specialties, it is likely also to be of interest to determine such estimates for each medical specialty. A physician considering relocation to Pennsylvania from elsewhere will be more interested in the estimated premium for her own medical specialty than the average premium of all Pennsylvania physicians. Accordingly, an investigator first decides the size n of a statewide sample she can afford. Then she obtains a directory of Pennsylvania physicians classified according to specialty and notes the number N_i of Pennsylvania physicians in each specialty $i, i = 1, \ldots, k$, where k is the number of distinct medical specialties. (Such a directory may be available for purchase from the American Medical Association.) Then a sample of approximately $n_i = \left(\frac{N_i}{N}\right) n$ physicians is selected from among the Pennsylvania practitioners of specialty i.

Stratified sampling has the virtue of avoiding an undersampling of any stratum and so guarantees some minimum degree of precision for estimates from each stratum. When the population exhibits minimal variability within strata but considerable variability between units in different strata, estimates based on stratified random sampling are likely to be more precise than ones based on simple random samples of comparable total size. This fact will be demonstrated in Section 3.11.5.

3.11.3 Cluster Random Sampling

This technique is designed to control the cost of sampling in exchange for some decrease in precision of estimation. It is most frequently used when it is necessary to make personal contact with the *sampling units* (entity that is to be sampled), and the sampling units are physically dispersed to the extent that traveling from one unit to another is an appreciable cost.

As with stratified sampling, cluster sampling involves two stages. Assume that the population is partitioned into c clusters. A cluster is typically formed from geographically contiguous units so that sampling units within the same cluster are much closer to one another than two units in different clusters. In stage 1 the analyst selects c_0 of these clusters, where c_0 is considerably less than c. Then in stage 2 the analyst randomly samples n_i items from each selected cluster i, where $\sum_{i=1}^{c_0} n_i = n$. The samples within each cluster can be simple random samples, stratified random samples, etc. As in the case of stratified random sampling, we must decide on a rule for allocating the total sample size n to the clusters.

If T_i is the total for all observations in cluster i, then the mean estimated from the cluster random sample is $\bar{y}_{\text{CRS}} = \left(\sum_i T_i\right)/\left(\sum_i N_i\right)$, where both sums extend from 1 to c_0.

Cluster random sampling saves costs because it involves much less travel from one cluster to another than other sampling methods. But precision is sacrificed because this method prevents a large part of the population from appearing in the sample. In contrast to stratified sampling of strata, cluster sampling of clusters is most efficient when the variation within clusters is large compared to the variation between clusters.

When it is required to personally interview persons sampled from a city's population of eligible voters, a good strategy would be to identify voting districts as clusters and use cluster sampling. If, instead, we wanted to interview city residents as to their product preferences, an analyst might prefer to use zip codes as clusters because geography-based marketing strategies are more likely to be segmented by zip code than by voting district.

3.11.4 Systematic Random Sampling

This method may be considered when simplicity of the sampling design and administration is of prime importance.

Order the population from 1 to N and initially assume that N is an integral multiple of n, say $N = mn$. Then randomly select an integer i, $1 \leq i \leq m$. Then sample population item i and every m^{th} item thereafter. For example, if $N = 120$, $n = 20$, $m = 6$, we might randomly sample items $4, 10, 16, \ldots, 118$.

Suppose instead that $N = mn + l$, $1 \leq l < n$. The analyst may then seek to move toward the N proportional to n situation. Suppose we modify the preceding illustration to $N = 132$. A possibility is to accept a larger $n = 22$. Another option that maintains $n = 20$ is to randomly remove $l = 12$ observations from sampling consideration and then proceed as before with the mn remaining observations.

This method should not be used if the population displays a periodic characteristic with the same period as m. For example, if we wish to randomly sample 20 houses in a subdivision consisting of 120 houses where each block has exactly 6 houses, then the preceding plan would either contain, or avoid, sampling houses on the end of blocks. Such houses tend to be on larger lots than ones in the middle of blocks and the plan would either include them exclusively or miss them entirely.

3.11.5 Standard Errors of Sample Means

In this section we provide standard errors for the means of random samples selected by various methods. Then according to the Central Limit Theorem, an approximate large-sample $100(1-\alpha)\%$ confidence interval for the population mean is of the form

$$\text{sample mean} \pm \text{standard error} \cdot z_{(1-\frac{\alpha}{2})}$$

For a simple random sample, the standard error is

$$s_{\text{SRS}} = \sqrt{\frac{s^2}{n}\left(\frac{N-n}{N-1}\right)}$$

For a stratified random sample with sample variance s_i^2 from stratum i, the standard error is

$$s_{\text{ST}} = \frac{1}{N}\sqrt{\sum_i N_i^2 \left(\frac{N_i - n_i}{N_i - 1}\right)\frac{s_i^2}{n_i}}$$

If the $\{s_i^2\}$ tend to be smaller than s, then s_{ST} will tend to be smaller than s_{SRS} with the conclusion that stratification was worthwhile.

To present the standard error for the mean of a cluster random sample, define $\bar{N} = N/c$ to be the average cluster size. The standard error is

$$s_{\text{CRS}} = \sqrt{\left(\frac{c - c_0}{c_0 c \bar{N}^2}\right) \frac{\sum_i (T_i - \bar{y}_{\text{CRS}} N_i)^2}{c_0 - 1}}$$

The summation extends from 1 to c_0, where as before, c_0 is the number of clusters that were sampled.

3.11.6 Sources of Bias in Samples

Sampling error is the discrepancy between the estimate and parameter being estimated. This error decreases as the sample size increases. Nonsampling errors are more serious than sampling errors because they can't be minimized by increasing the sample size. Continuing the example discussed in Section 3.11.2, we discuss two such sources of bias in the context of randomly sampling physicians who practice in Pennsylvania. *Selection bias* occurs when it is impossible to sample some members of the population. *Nonresponse bias* occurs if responses are not obtained from some members of the sample.

In order to randomly sample from the population consisting of all physicians licensed to practice medicine in Pennsylvania, we must obtain a list or computer file of such physicians. Even if we could obtain a list of physicians licensed to practice, there is no way to know which physicians on such a list are in fact practicing medicine (as opposed to performing medical research or administrative tasks). Therefore, use of such a list would introduce selection bias. A better approach might be to obtain a list of the Pennsylvania membership of the American Medical Association (AMA). This list does indicate the nature of the physician's practice, if any, so nonpractitioners on the list can be ignored. However, not all physicians practicing in Pennsylvania are AMA members; such membership is not legally required in order to practice medicine. Thus some selection bias would still be present with this approach. Selection bias would be eliminated if the client can be persuaded to amend the target population to AMA members practicing in Pennsylvania.

Next suppose that this amendment is accepted and that a random sample of n practicing physicians is selected from the list. How should the physicians be contacted? Since physicians are busy individuals; visiting them in person or contacting them by telephone is unlikely to yield a response. Ignoring nonrespondents is likely to result in nonresponse bias because busier physicians are less likely to respond, and busyness may be associated with the survey questions.

Mail contact of the sampled physicians is preferred for several reasons. Since a written questionnaire can be answered at the physician's convenience, the physician is more likely to respond. Second, the questionnaire can be placed under a cover letter that encourages participation, written by a person respected by the respondents. Third, it is possible to keep track of who does not initially respond so that such individuals can be contacted again. This is accomplished by asking respondents to mail in a signed postcard indicating that they have participated, and to return the anonymous questionnaire in an envelope mailed separately.

Even this elaborate mail questionnaire approach does not eliminate the possibility of nonresponse bias. The extent of any remaining bias can be judged by comparing characteristics of the sampled physicians with those of the physician population reported in the AMA membership directory.

3.12 Exercises

3.1. Refer to the discrete bivariate distribution considered in Table 3.3.

a. Let $Z = X + 1$. Find the distribution of Z.
b. Find $E(2X + 1)$ and $2E(X) + 1$. Then find $E(X^2)$ and $[E(X)]^2$.
c. Find $P(X < Y)$.
d. Let X_1 and X_2 be independent and identically distributed as X. Make a table of the joint distribution of X_1 and X_2, and use this to find $P(X_1 < X_2 + 1)$.

3.2. How large a random sample is required for there to be a 92% probability of sampling at least one defective from a lot of 100,000 items which contains 100 defectives? (Hints: What is the random variable here? Consider the event that is the complement of "at least one defective".)

3.3. Suppose X is binomial(50, .10), and Y is binomial(20, .25). Draw the distribution functions of X and Y. Which one has a bigger mean? Which one has a bigger standard deviation?

3.4. If X, Y are each standard normal random variables, and they are independent of one another, what is the distribution of $Z = 3X + 2Y$?

3.5. Suppose that Y is a 2×1 random vector such that

$$W = \begin{pmatrix} 80 \\ 40 \end{pmatrix} + \begin{pmatrix} 10 & 7 \\ 7 & 5 \end{pmatrix} Y$$

has a bivariate normal distribution with mean $\begin{pmatrix} 60 \\ 70 \end{pmatrix}$ and covariance matrix $\begin{pmatrix} 100 & 40 \\ 40 & 50 \end{pmatrix}$. Find the probability distribution of Y, including its mean vector and covariance matrix.

3.6. In class #1, 32 out of 40 students earned fewer than 70 points on the final exam. In class #2, 40 out of 50 students earned fewer than 75 points on the same exam. Restate the given class information in terms of percentiles. Is it possible to tell which class had a higher average score?

3.7. Somebody tells you that a 95% confidence interval for the mean number of customers per day is (74.2, 78.5), and that this indicates that 95% is the probability that the mean is between 74.2 and 78.5. Criticize this statement and replace it with one correct sentence.

3.8. Acme, Inc. thinks it has a new way of manufacturing a key product. It is trying to choose between A = "new way is better than old way" or B = "old way is better than new way". Acme plans to reach its tentative conclusion by sampling some of the product produced the new way and conducting a statistical test. The new way is much more expensive than the old way. Which statement, A or B, should be the null hypothesis? Justify your answer.

3.9. The probability that a project succeeds in New York is .4, the probability that it succeeds in Chicago is .5, and the probability that it succeeds in at least one of these cities is .6. Find the probability that this project succeeds in Chicago given that it succeeds in New York.

3.10. You are considering two projects, A and B. With A you estimate a payoff of $60,000 with probability .6 and $30,000 with probability .4. With B you estimate a payoff of $80,000 with probability .5 or $30,000 with probability .5. Answer the following questions after performing appropriate calculations.

a. Which project is better in terms of expected payoff?

b. Which project is better in terms of variability of payoff?

3.11. If X has a mean of 15 and a standard deviation of 4, and if $Y = 5 - 3X$, what are the mean and standard deviation of Y?

3.12. State the two ways in which a data analyst can modify a statistical test in order to decrease its Type II error probability.

3.13. An analyst makes three independent inferences. For each of these inferences, the probability is .05 that it is *in*correct. Find the probability that *all three* inferences are *correct*.

3.14. Let A = "a McDonald's franchise in Kansas is profitable" and let B = "the Philadelphia Eagles will have a winning season next year". If $P(A) = .8$ and $P(B) = .6$, find the probability that *either A or B* occurs.

3.15. Use statistical software commands to do this problem. A new medicine has probability .70 of curing gout. If a random sample of 10 people with gout are to be given this medicine, what is the probability that among the 10 people in the sample, between 5 and 8 people will be cured?

3.16. Use statistical software commands to do this problem. The daily output of a production line is normally distributed with a mean of 163 units and a standard deviation of 4 units.

a. Find the probability that a particular day's output will be 160 units or less.

b. The production manager wants to tell her supervisor, "80% of the time our production is at least x units". What number should she use for x?

3.17. Find the expected value and standard deviation of a random variable U if its probability distribution is as follows:

u	$P(U = u)$
1	.6
2	.3
3	.1

3.18. A random variable W has probability density function $f(w) = 2 - 2w$, $0 < w < 1$, and $f(w) = 0$ for all other values of w.

a. Verify that $f(w)$ is indeed a probability density function.

b. Find the corresponding cumulative distribution function, $\mathcal{F}(w)$.

c. Find the expectation of W.

d. Find the standard deviation of W.

e. Find the median of this distribution, i.e., the number w_m such that $P(W < w_m) = .5$.

3.19. Use a statistical software command to approximate the value of $z_{.08}$.

3.20. State the two things that a data analyst can do in order to make a confidence interval *narrower*.

3.21. A data analyst tentatively decides on values for α and n for a statistical test. Before performing the test she investigates its Type II error control and finds this to be unsatisfactory. What two options does she have to improve Type II error control?

3.22. In the discussion of *sufficiency* of a point estimator in Section 3.6.3, we indicated that $2\bar{x}$ is not a good estimator of a from a sample of n items from a continuous uniform distribution on $[0, a]$. Can you suggest a better estimator of a and explain why it is better than $2\bar{x}$?

3.23. The dataset `data(salary)`, from Forbes Magazine (1993), contains the ages and salaries of the chief executives of the 60 most highly ranked firms among *Forbes Magazine*'s "Best small firms in 1993." Consider the variable `age`.

a. Produce a boxplot and a stem-and-leaf plot for `age`.

b. Construct a 95% confidence interval for the mean age. What assumptions were made in your construction?

c. Test $H_0: \mu \le 50$ against $H_1: \mu > 50$, reporting and interpreting the p-value for this test.

d. Approximate the power of this test for the alternative $\mu_1 = 53$ by using the normal distribution as an approximation for the test statistic in part c, assuming $\alpha = .05$.

3.24. The dataset `data(cereals)` contains various nutritional measurements for 77 breakfast cereals. We are concerned here with the variable `carbo` (carbohydrates) measured in grams per serving. Be aware that the cereal Quaker Oatmeal shows a negative value for carbohydrates, probably indicating a missing value for that observation. Be sure that you inform your data analysis package of this anomaly and that the package does something sensible with that information. Elimination of the observation is one possible response to missingness.

a. Produce boxplots and stem-and-leaf plot for `carbo`. Do these plots suggest that this variable comes from a normal population?

b. Construct at 99% confidence interval for the mean carbohydrate content.

c. Test $H_0: \mu \ge 16$ against $H_1: \mu < 16$, reporting and interpreting the p-value for this test.

d. Approximate the probability of committing a Type II error for the alternative $\mu_1 = 15$. Use the normal distribution to approximate the test statistic in part c, assuming $\alpha = .05$.

3.25. The sampling bias in the December 1969 U.S. Draft Lottery, with data in file `data(draft70mn)`, is described in Exercise 4.1. Suppose you had been the administrator of that lottery. Explain how you would have performed the sampling without incurring such bias.

3.26. Royalties paid to authors of novels have sometimes been based on the number of words contained in the novel. Recommend to an old-fashioned author how to estimate the number of words in a handwritten manuscript she is planning to give to her publisher.

3.27. Samples are taken from two strata. Suppose the variance of the two samples combined is $s^2 = 7.6$ and the following within-stratum information is known:

Stratum	N_i	n_i	s_i^2
1	100	30	1.2
2	120	40	1.4

Observe that there is far less variability within the two strata than between the two strata. Calculate s_{SRS} and s_{ST} to verify that for estimating the common population mean in this situation, $\bar{x}_{\text{SRS}}$ is much preferred to $\bar{x}_{\text{ST}}$.

3.28. The organization of a candidate for a city political office wishes to poll the electorate. For this purpose, discuss the relative advantages and disadvantages of personal interview polling vs telephone polling.

3.29. Explain how it is possible for a census to yield less accurate results than a random sample from the same population.

3.30. A student claims that a random sample of n items from a population of N items is one selected so that each item in the population has the same probability $\frac{n}{N}$ of appearing in the sample. Demonstrate that this definition is inadequate.

3.31. A four-drawer file cabinet contains several thousand sheets of paper, each containing a statement of the dollar amount due to be paid to your company. The sheets are arranged in the order that the debt was incurred. You are asked to spend not more than one hour to estimate the average dollar amount on all sheets in the file cabinet. Propose a plan for accomplishing this.