STAT 700

APPLIED STATISTICS I

Fall, 2005

Lecture Notes

Joshua M. Tebbs

Department of Statistics

The University of South Carolina

Contents

1 Looking at Data: Distributions

TERMINOLOGY: **Statistics** is the development and application of theory and methods to the collection (design), analysis, and interpretation of observed information from planned (or unplanned) experiments and other studies.

TERMINOLOGY: **Biometry** is the development and application of statistical methods for biological experiments (which are often planned).

SCOPE OF APPLICATION: Statistical thinking can be used in all disciplines! Consider the following examples:

- In a reliability study, tribologists aim to determine the main cause of failure in an engine assembly. Failures reported in the field have had an effect on customer confidence (for safety concerns).
- In a marketing project, store managers in Aiken, SC want to know which brand of coffee is most liked among the 18-24 year-old population.
- In an agricultural study in North Carolina, researchers want to know which of three fertilizer compounds produces the highest yield.
- In a clinical trial, physicians on a Drug and Safety Monitoring Board want to determine which of two drugs is more effective for treating HIV in the early stages of the disease.
- A chemical reaction that produces a product may have higher or lower yield depending on the temperature and stirring rate in the vessel where the reaction takes place. An engineer would like to understand how yield is affected by the temperature and stirring rate.
- The Human Relations Department at a major corporation wants to determine employee opinions about a prospective pension plan adjustment.

- In a public health study conducted in Sweden, researchers want to know whether or not smoking (i) causes lung cancer and/or (ii) is strongly linked to a particular social class.

TERMINOLOGY: In a statistical problem, the **population** is the entire group of individuals that we want to make some statement about. A **sample** is a part of the population that we actually observe.

TERMINOLOGY: A **variable** is a characteristic (e.g., temperature, age, CD4 count, growth, etc.) that we would like to measure on individuals. The actual measurements recorded on individuals in the sample are called **data**.

TWO TYPES OF VARIABLES: **Quantitative** variables have measurements (data) on a numerical scale. **Categorical** variables have measurements (data) where the values simply indicate group membership.

TERMINOLOGY: The **distribution** of a variable tells us what values the variable takes and how often it takes these values.

Example 1.1. Which of the following variables are quantitative in nature? Which are categorical?

- IKEA-Atlanta daily sales (measured in $1000's)
- store location (Baltimore, Atlanta, Houston, Detroit, etc.)
- CD4 cell count
- yield (bushels/acre)
- payment times (in days)
- payment times (late/not late)
- age

- advertising medium (ratio/TV/internet)
- number of cigarettes smoked per day
- smoking status (yes/no).

TERMINOLOGY: An **experiment** is a planned study where individuals are subjected to **treatments**. In an **observational study**, individuals are not "treated;" instead; we simply observe their information (data).

TERMINOLOGY: The process of generalizing the results in our sample to that of the entire population is known as **statistical inference**. We'll study this more formally later in the course.

Example 1.2. Salmonella bacteria are widespread in human and animal populations; in particular, some serotypes can cause disease in swine. A food scientist wants to see how withholding feed from pigs prior to slaughter can reduce the number and size of gastrointestinal tract lacerations during the actual slaughtering process. This is an important issue since pigs infected with salmonellosis may contaminate the food supply.

- **Individuals** = pigs
- **Population** = all market-bound pigs, say
- **Sample** = 45 pigs from 3 farms (15 per farm) assigned to three treatments:
  - Treatment 1: no food withheld prior to transport,
  - Treatment 2: food withheld 12 hours prior to transport, and
  - Treatment 3: food withheld 24 hours prior to transport.
- Data were measured on many variables, including body temperature prior to slaughter, weight prior to slaughter, treatment assignment, the farm from which each pig originated, number of lacerations recorded, and size of laceration (cm).
- **Boxplots** of the lacerations lengths (by treatment) are in Figure 1.1.

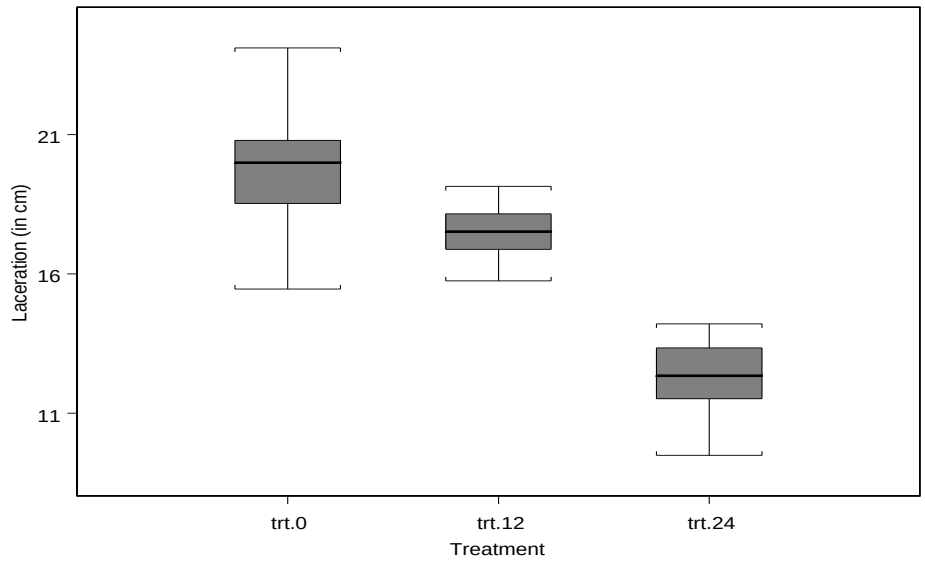


Figure 1.1: *Salmonella experiment. Laceration length for three treatments.*

SOME PARTICULAR QUESTIONS OF INTEREST:

- How should we assign pigs to one of the three treatments?
- What are the **sources of variation**? That is, what systematic components might affect laceration size or number of lacerations?
- Why would one want to use animals from three farms?
- Why might body temperature or prior weight be of interest?

GENERAL REMARKS:

- In agricultural, medical, and other biological applications, the most common objective is to **compare** two or more treatments. In light of this, we will often talk about statistical inference in the context of comparing treatments in an experimental setting. For example, in the salmonella experiment, one goal is to compare the three withholding times (0 hours, 12 hours, and 24 hours).

- Since populations are usually large, the sample we observe is just one of many possible samples that are possible to observe. That is, samples may be similar, but they are by no means identical. Because of this, *there will always be a degree of uncertainty about the decisions that we make concerning the population of interest!!*
- A main objective of this course is to learn how to design controlled experiments and how to analyze data from these experiments. We would like to make conclusions based on the data we observe, and, of course, we would like our conclusions to apply for the entire population of interest.

1.1 Displaying distributions with graphs

IMPORTANT: Presenting data effectively is an important part of any statistical analysis. How we display data depends on the type of variable(s) or data that we are dealing with.

- **Categorical**: pie charts, bar charts, tables
- **Quantitative**: stemplots, boxplots, histograms, timeplots

UNDERLYING THEMES: Remember that the data we collect are often best viewed as a sample from a larger population of individuals. In this light, we have two primary goals in this section:

- learn how to summarize and to display the **sample** information, and
- start thinking about how we might use this information to learn about the underlying **population**.

1.1.1 Categorical data distributions

Example 1.3. HIV infection has spread rapidly in Asia since 1989, partly because blood and plasma donations are not screened regularly before transfusion. To study this,

researchers collected data from a sample of 1390 individuals from villages in rural eastern China between 1990 and 1994 (these individuals were likely to donate plasma for financial reasons). One of the variables studied was **education level**. This was measured as a categorical variable with three categories (levels): illiterate, primary, and secondary.

TABLES: I think the easiest way to portray the distribution of a categorical variable is to use a **table** of counts and/or percents. A table for the education data collected in Example 1.3 is given in Table 1.1.

Table 1.1: *Education level for plasma donors in rural eastern China between* 1990-1994.

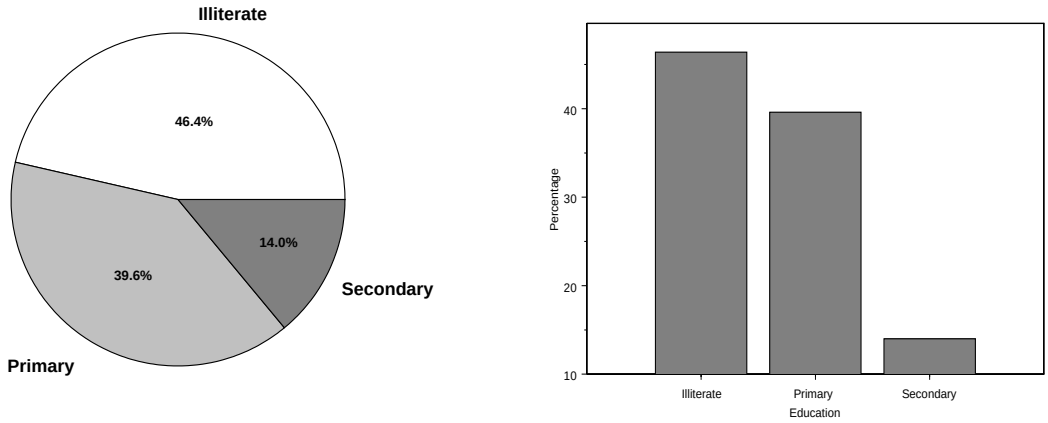
Education level	Count	Percentage
Illiterate	645	46.4
Primary	550	39.6
Secondary	195	14.0
Total	1390	100.0

REMARK: Including percentages in the table (in addition to the raw counts) is helpful for interpretation. Most of us can understand percentages easily. Furthermore, it puts numbers like "645" into perspective; e.g., 645 out of what?

INFERENCE: These data are from a sample of rural villagers in eastern China. From these data, what might we be able to say about the entire population of individuals?

PIE CHARTS AND BAR CHARTS: **Pie charts** and **bar charts** are appropriate for categorical data but are more visual in nature. A pie chart for the data collected in Example 1.3 is given in Figure 1.2(a). A bar chart is given in Figure 1.2(b).

REMARK: Unfortunately, pie charts and bar charts are of limited use because they can only be used effectively with a single categorical variable. On the other hand, tables can be modified easily to examine the distribution of two or more categorical variables (as we will see later).



(a) Pie chart. (b) Bar chart.

Figure 1.2: *Education level for plasma donors in rural eastern China between* 1990-1994.

1.1.2 Quantitative data distributions

GRAPHICAL DISPLAYS: Different graphical displays can also be used with quantitative data. We will examine **histograms**, **stem plots**, **time plots**, and **box plots**.

Example 1.4. Monitoring the shelf life of a product from production to consumption by the consumer is essential to assure quality. A sample of 25 cans of a certain carbonated beverage were used in an industrial experiment that examined the beverage's shelf life, measured in days. The data collected in the experiment are given in Table 1.2.

Table 1.2: *Beverage shelf life data.*

262	188	234	203	212	212	301	225	241	211	231	227	217
252	206	281	251	219	268	231	279	243	241	290	249	

GOALS:

- The goal of a graphical display is to provide a visual impression of the characteristics of the data from a sample. The hope is that the characteristics of the **sample** are a likely indication of the characteristics of the population from which it was drawn.

- In particular, we will be always be interested in the
  - **center** of the distribution of data
  - **spread** (variation) in the distribution of data
  - **shape**: is the distribution symmetric or skewed?
  - the presence of **outliers**.

TERMINOLOGY: A **frequency table** is used to summarize data in a tabular form. Included are two things:

- **class intervals**: intervals of real numbers
- **frequencies**: how many observations fall in each interval.

Table 1.3: *Frequency table for the shelf life data in Example* 1.4.

Class Interval	Frequency
[175, 200)	1
[200, 225)	7
[225, 250)	9
[250, 275)	4
[275, 300)	3
[300, 325)	1

NOTE: The number of intervals should be large enough that not all observations fall in one or two intervals, but small enough so that we don't have each observation belonging to its own interval.

HISTOGRAMS: To construct a **histogram**, all you do is plot the frequencies on the vertical axis and the class intervals on the horizontal axis. The histogram for the shelf life data in Example 1.4 is given in Figure 1.3.

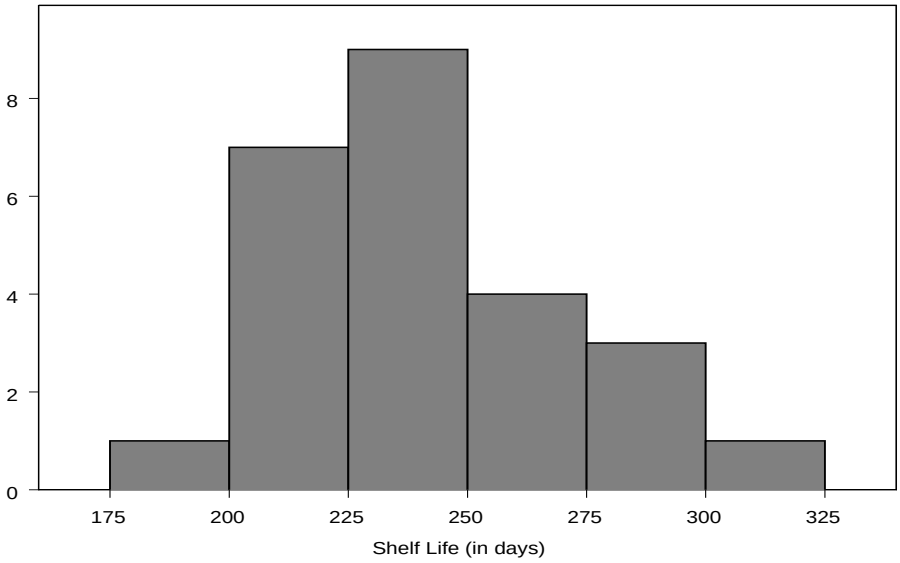


Figure 1.3: *Histogram for the shelf life data in Example* 1.4.

INTERPRETATION: We see that the distribution of the shelf life data is approximately **symmetric**. The center is around 240 days, and variation among the shelf lives is pretty apparent. There are no gross outliers in the data set.

INFERENCE:

- We can use the distribution to estimate what percentage of cans have shelf life in a certain range of interest. Suppose that the experimenter believed "most shelf lives should be larger than 250 days." From the distribution, we see that this probably is **not** true if these data are representative of the population of shelf lives.
- We can also associate the percentage of lives in a certain interval as being proportional to the **area** under the histogram in that interval. For example, are more cans likely to have shelf lives of 200-225 days or 300-325 days? We can estimate these percentages by looking at the data graphically.

TERMINOLOGY: If a distribution is not symmetric, it is said to be **skewed**.

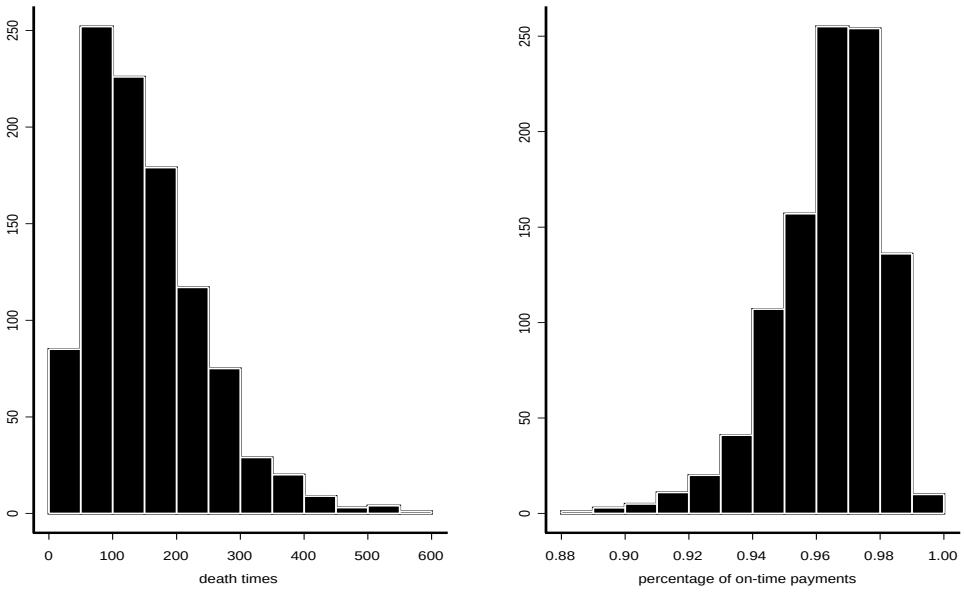


Figure 1.4: *Left: Death times for rats treated with a toxin. Right: Percentage of monthly on-time payments.*

SKEWED DISTRIBUTIONS: Skewed distributions occur naturally in many applications. Not all distributions are symmetric or approximately symmetric! In Figure 1.4, the left distribution is **skewed right**; the right distribution is **skewed left**.

STEM PLOTS: These plots provide a quick picture of the distribution while retaining the numerical values themselves. The idea is to separate each data value into a **stem** and a **leaf**. Stems are usually placed on the left, with leaves on the right (in ascending order). Stem plots work well with small-to-moderate-sized data sets, say, 15 to 50 observations. The stem plot for the shelf life data in Example 1.4 appears in Table 1.4. In this plot, the units digit is the leaf; the tens and hundreds digits form the stem.

LONGITUDINAL DATA: In many applications, especially in business, data are observed **over time**. Data that are observed over time are sometimes called **longitudinal data**. More often, longitudinal data are quantitative (but they need not be). Examples of longitudinal data include monthly sales, daily temperatures, hourly stock prices, etc.

Table 1.4: *Stem plot for the shelf life data in Example* 1.4.

18	8
19	
20	3 6
21	1 2 2 7 9
22	5 7
23	1 1 4
24	1 1 3 9
25	1 2
26	2 8
27	9
28	1
29	0
30	1

TIME PLOTS: If it is the longitudinal aspect of the data that you wish to examine graphically, you need to use a graphical display which exploits this aspect. **Time plots** are designed to do this (histograms and stem plots, in general, do not do this). To construct a time plot, simply plot the individual values (on the vertical axis) versus time (on the horizontal). Usually, individual values are then connected with lines.

Example 1.5. The Foster Brewing Company is largely responsible for the development of packaged beers in Australia. In fact, canned beer had been developed in the USA just prior to the Second World War and was first produced in Australia in the early 1950s. Developments since have included improved engineering techniques has led to larger vessels and improved productivity. The data available online at

`http://www.maths.soton.ac.uk/teaching/units/math6011/mwhdata/Beer2.htm`

are the monthly beer sales data in Australia from January 1991 to September 1996. A time plot of the data appears in Figure 1.5.

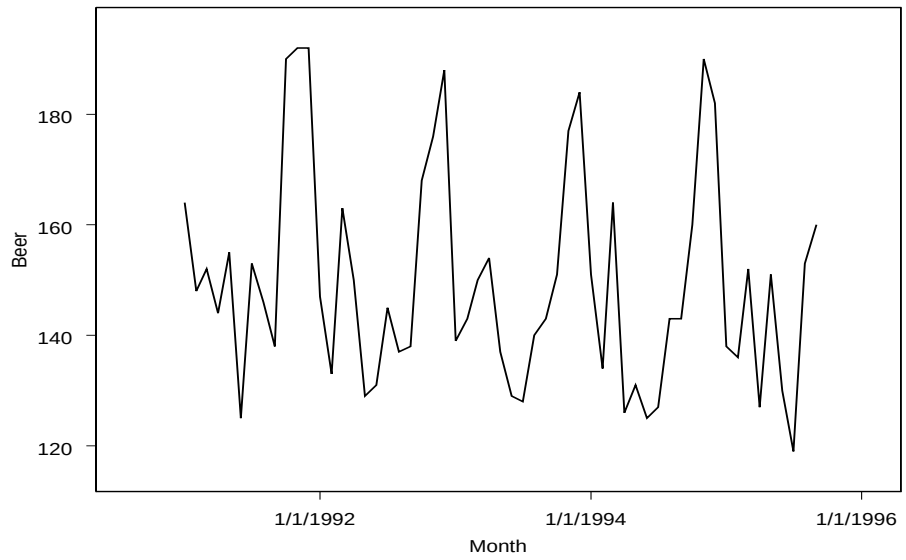


Figure 1.5: *Australian beer sales from January* 1991 *to September* 1996.

INTERPRETING TIME PLOTS: When I look at time plots, I usually look for two things in particular:

- Increasing or decreasing **trends**. Is there a general shift over time upward or downward? Is it slight or notably apparent?
- Evidence of **seasonal effects**. Are there repeated patterns at regular intervals? If there is, what is most likely to have produced this pattern?

USEFULNESS: Analyzing longitudinal data is important for **forecasting** or **prediction**. For example, can we forecast the next two years of beer sales? Why might this information be important?

DECOMPOSING TIME SERIES: For a given set of longitudinal data, it is often useful to decompose the data into its **systematic** parts (e.g., trends, seasonal effects) and into its **random** parts (the part that is left over afterwards). See pages 21-23 MM.

1.2 Displaying distributions with numbers

NOTE: In Section 1.1, the main goal was to describe distributions of data graphically. We now wish to do this numerically. For the remainder of the course, we will adopt the following notation to describe a sample of data.

$$n = \text{number of observations in sample} \qquad x = \text{variable of interest}$$

$$x_1, x_2, ..., x_n = \text{the } n \text{ data values in our sample}$$

We now examine numerical summaries of data. We will quantify the notion of **center** and **variation** (i.e., spread).

PREVAILING THEME: Our goal is to numerically summarize the distribution of the sample and get an idea of these same notions for the population from which the sample was drawn.

1.2.1 Measures of center

TERMINOLOGY: With a sample of observations $x_1, x_2, ..., x_n$, the **sample mean** is defined as

$$\begin{aligned} \overline{x} &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i. \end{aligned}$$

That is, the sample mean is just the arithmetic average of the n values $x_1, x_2, ..., x_n$. The symbol $\overline{x}$ is pronounced "x-bar," and is common notation. Physically, we can envision $\overline{x}$ as the balancing point on the histogram for the data.

SIGMA-NOTATION: The symbol

$$\sum \qquad \text{or} \qquad \sum_{i=1}^{n}$$

denotes **sum**. In particular, the sum of the data $x_1, x_2, ..., x_n$ can be expressed either as

$$x_1 + x_2 + \cdots + x_n \qquad \text{or} \qquad \sum_{i=1}^{n} x_i.$$

We will be using **sigma-notation** throughout the course. The symbol i denotes the **index of summation**. This is used to tell us which values to add up. The n on top of the summation symbol is the index of the final quantity added.

TERMINOLOGY: With a sample of observations $x_1, x_2, ..., x_n$, the **sample median**, denoted by M, is the middle ordered value (when the data are ordered low to high). If the sample size n is an odd, then the median will be uniquely defined; if n is even, then the median is the **average** of the middle two ordered observations.

Example 1.6. With our beverage shelf life data from Example 1.4, the sum of the data is

$$\sum_{i=1}^{25} x_i = x_1 + x_2 + \cdots + x_{25} = 262 + 188 + \cdots + 249 = 5974,$$

and the sample mean of the 25 shelf lives is given by

$$\overline{x} = \frac{1}{25}\sum_{i=1}^{25} x_i = \frac{1}{25}(5974) = 238.96 \text{ days}.$$

From the stemplot in Table 1.4, we can see that the median is

$$M = 234 \text{ days}.$$

Note that the median is the 13th ordered value; 12 values fall below M and 12 values fall above M. Also, note that the mean and median are fairly "close," but not equal.

COMPARING THE MEAN AND MEDIAN: The mean is a measure that can be heavily influenced by **outliers**. Unusually high data observations will tend to increase the mean, while unusually low data observations will tend to decrease the mean. One or two outliers will generally not affect the median! Sometimes we say that the median is generally **robust** to outliers.

MORAL: If there are distinct outliers in your data set, then perhaps the median should be used as a measure of center instead of the mean.

OBSERVATION: If the mean and median are close, this is usually (but not always) an indication that the distribution is **approximately symmetric**. In fact,

- if a data distribution is **perfectly symmetric**, the median and mean will be equal.
- if a data distribution is **skewed right**, the mean will be greater than the median.
- if a data distribution is **skewed left**, the median will be greater than the mean.

1.2.2 The five number summary, box plots, and percentiles

BOX PLOT: A **box plot** is a graphical display that uses the **five number summary**:

- **min**, the smallest observation.
- Q_1, **the first quartile**; i.e., the observation such that approximately 25 percent of the observations are smaller than Q_1.
- **median**, M, the middle ordered observation; i.e., the observation such that roughly half of the observations fall above the median, and roughly half fall below.
- Q_3, **the third quartile**; i.e., the observation such that approximately 75 percent of the observations are smaller than Q_3.
- **max**, the largest observation.

NOTE: To compute the five number summary, one first has to **order** the data from low to high. The values of Q_1, the median, and Q_3 may not be uniquely determined! In practice, computing packages give these values by default or by request.

Example 1.7. For the beverage shelf life data in Example 1.4, the five number summary is

$$\text{min} = 188 \qquad Q_1 = 217 \qquad \text{median} = 234 \qquad Q_3 = 252 \qquad \text{max} = 301.$$

The box plot for the shelf life data uses these values and is given in Figure 1.6.

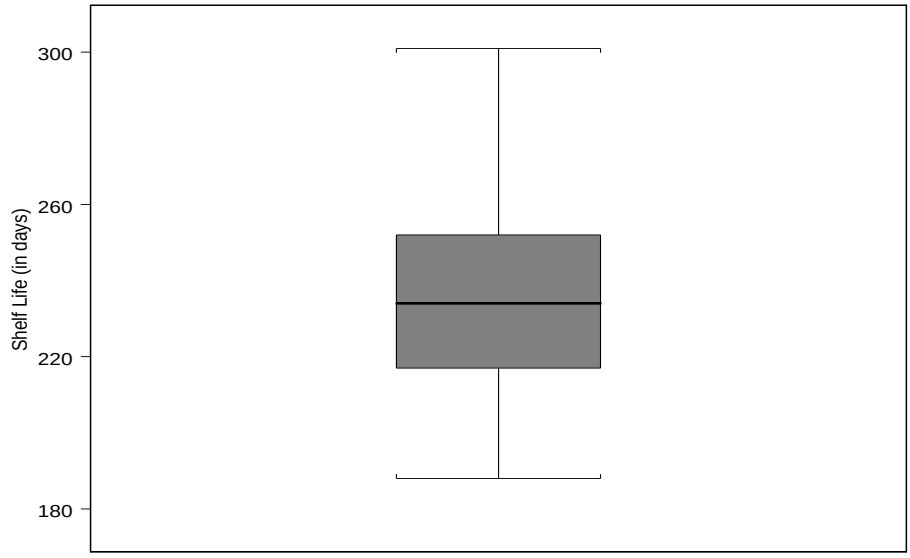


Figure 1.6: *Box plot for shelf life data in Example* 1.4.

TERMINOLOGY: The p**th percentile** of a data distribution is the value at which p percent of the observations fall at or below it. Here, p is a number between 0 and 100. We have already seen three "special" percentiles! Q_1 is the 25th percentile, M is the 50th percentile, and Q_3 is the 75th percentile.

1.2.3 Measures of variation

OBSERVATION: Two data sets could have the same mean, but values may be spread about the mean value differently. For example, consider the two data sets:

24	25	26	27	28
6	16	26	36	46

Both data sets have $\overline{x} = 26$. However, the second data set has values that are much more spread out about 26. The first data set has values that are much more compact about 26. That is, **variation** in the data is different for the two data sets.

RANGE: An easy way to assess the variation in a data set is to compute the **range**, which we denote by R. The range is the largest value minus the smallest value; i.e.,

$$R = x_{max} - x_{min}.$$

For example, the range for the first data set above is $28 - 24 = 4$, while the range for the second is $46 - 6 = 40$.

DRAWBACKS: The range is very sensitive to outliers since it only uses the extreme observations. Additionally, it ignores the middle $(n-2)$ observations, which is potentially a lot of information!

INTERQUARTILE RANGE: The **interquartile range**, IQR, measures the spread in the center half of the data; it is the difference between the first and third quartiles; i.e.,

$$IQR = Q_3 - Q_1.$$

NOTE: This measure of spread is more resistant to outliers since it does not use the extreme observations. Because of this, the IQR can be very useful for describing the spread in skewed distributions.

INFORMAL OUTLIER RULE: The $1.5 \times IQR$ **outlier rule** says to classify an observation as an outlier if it falls $1.5 \times IQR$ above the third quartile or $1.5 \times IQR$ below the first quartile.

Example 1.8. For the shelf life data in Example 1.4 and Example 1.7, we see that

$$IQR = Q_3 - Q_1 = 252 - 217 = 35.$$

Thus, observations above

$$Q_3 + 1.5 \times IQR = 252 + 1.5(35) = 304.5$$

or below

$$Q_3 - 1.5 \times IQR = 217 - 1.5(35) = 164.5$$

would be classified as outliers. Thus, none of the shelf life observations would be classified as outliers using this rule.

TERMINOLOGY: The **sample variance** of $x_1, x_2, ..., x_n$ is denoted by s^2 and is given by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right].$$

RATIONALE: We are trying to measure how far observations deviate from the sample mean $\overline{x}$. A natural quantity to look at is each observation's **deviation from the mean**, i.e., $x_i - \overline{x}$. However, one can show that

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0;$$

that is, the positive deviations and negative deviations "cancel each other out" when you add them!

REMEDY: Devise a measure that maintains the magnitude of each deviation but ignores their signs. Squaring each deviation achieves this goal. The quantity

$$\sum_{i=1}^{n}(x_i - \overline{x})^2$$

is called the **sum of squared deviations**. Dividing $\sum_{i=1}^{n}(x_i - \overline{x})^2$ by $(n-1)$ leaves (approximately) an average of the n squared deviations. This is the sample variance.

TERMINOLOGY: The **sample standard deviation** is the positive square root of the variance; i.e.,

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

FACTS ABOUT THE STANDARD DEVIATION AND VARIANCE:

1. The larger the value of s (s^2), the more variation in the data $x_1, x_2, ..., x_n$.
2. $s \geq 0$ ($s^2 \geq 0$).
3. If $s = 0$ ($s^2 = 0$), then $x_1 = x_2 = \cdots = x_n$. That is, all the data values are equal (there is zero spread).

4. s and s^2, in general, are heavily influenced by outliers.

5. s is measured in the **original units** of the data; s^2 is measured in (units)^2.

6. The quantities s and s^2 are used mostly when the distribution of the sample data $x_1, x_2, ..., x_n$ is approximately **symmetric** (as opposed to skewed distributions).

7. Computing s and s^2 can be very tedious if n is large! Fortunately, many hand-held calculators can compute the standard deviation and variance easily, and, of course, many software packages can as well.

Example 1.9. Keeping plants healthy requires an understanding of the organisms and agents that cause disease as well as an understanding of how plants grow and are affected by disease. In a phytopathology experiment studying disease transmission in insects, x denotes the number of insects per plot. A sample of $n = 5$ plots is observed yielding $x_1 = 5$, $x_2 = 7$, $x_3 = 4$, $x_4 = 9$ and $x_5 = 5$. Calculating the sum and the **uncorrected sum of squares**, we get

$$\sum_{i=1}^{5} x_i = 5 + 7 + 4 + 9 + 5 = 30$$

$$\sum_{i=1}^{5} x_i^2 = 5^2 + 7^2 + 4^2 + 9^2 + 5^2 = 196.$$

The sample mean is $\overline{x} = \frac{1}{5}(30) = 6$ insects/plot. The sum of squared deviations is equal to

$$\begin{aligned} \sum_{i=1}^{5}(x_i - \overline{x})^2 &= \sum_{i=1}^{5} x_i^2 - \frac{1}{5}\left(\sum_{i=1}^{5} x_i\right)^2 \\ &= 196 - \frac{1}{5}(30)^2 \\ &= 196 - 180 = 16. \end{aligned}$$

Thus, the sample variance is $s^2 = \frac{1}{4}\sum_{i=1}^{5}(x_i - \overline{x})^2 = 16/4 = 4$, and the sample standard deviation is $s = \sqrt{4} = 2$ insects/plot.

Example 1.10. With our shelf life data from Example 1.4, we now compute the variance and standard deviation. Recall in Example 1.6 we computed $\overline{x} = 238.96$ days. First, we

compute

$$\sum_{i=1}^{25} x_i = 262 + 188 + \cdots + 249 = 5974$$

$$\sum_{i=1}^{25} x_i^2 = 262^2 + 188^2 + \cdots + 249^2 = 1447768.$$

Thus,

$$\begin{aligned}\sum_{i=1}^{25}(x_i - \overline{x})^2 &= \sum_{i=1}^{25} x_i^2 - \frac{1}{25}\left(\sum_{i=1}^{25} x_i\right)^2 \\ &= 1447768 - \frac{1}{25}(5974)^2 \\ &= 20220.96,\end{aligned}$$

and the sample variance is $s^2 = \frac{1}{24}\sum_{i=1}^{25}(x_i - \overline{x})^2 = 20220.96/24 = 842.54$. The sample standard deviation is $s = \sqrt{842.54} \approx 29.03$ days. The unit of measurement associated with the variance is days^2, which has no physical meaning! This is one of the advantages of using the standard deviation as a measure of variation.

1.2.4 The empirical rule

The **empirical rule**, or the **68-95-99.7 rule** says that if a histogram of observations is **approximately symmetric**, then

- approximately 68 percent of the observations will be within **one** standard deviation of the mean,
- approximately 95 percent of the observations will be within **two** standard deviations of the mean, and
- approximately 99.7 percent (or almost all) of the observations will be within **three** standard deviations of the mean.

We will see the justification for this in the next section when we start discussing the **normal distribution**.

Example 1.11. Returning to our shelf life data from Example 1.4, recall that the histogram of measurements was approximately symmetric (see Figure 1.3). Hence, the data should follow the empirical rule! Forming our intervals one, two, and three standard deviations from the mean, we have

- $(\overline{x} - s, \overline{x} + s) = (238.96 - 29.03, 238.96 + 29.03)$ or $(209.93, 267.99)$,
- $(\overline{x} - 2s, \overline{x} + 2s) = (238.96 - 2 \times 29.03, 238.96 + 2 \times 29.03)$ or $(180.91, 297.02)$,
- $(\overline{x} - 3s, \overline{x} + 3s) = (238.96 - 3 \times 29.03), 238.96 + 3 \times 29.03)$ or $(151.88, 326.04)$.

Checking the Empirical Rule with the shelf life data, we see 17/25 or 68 percent fall in $(209.93, 267.99)$, 24/25 or 96 percent fall in $(180.91, 297.02)$, and all the data fall in the interval $(151.88, 326.04)$. Thus, for the shelf life data, we see a close agreement with the empirical rule.

A NOTE ON COMPUTING STATISTICS BY HAND. Fortunately, we have calculators and computing to perform calculations involved with getting means and standard deviations. Except where explicitly instructed to perform calculations by hand, please feel free to use SAS, Minitab, or any other computer package that suits you. The goal of this course is not to be inundated with hand calculations (although we will be doing some!).

1.2.5 Linear transformations

REMARK: In many data analysis settings, we often would like to work with data on different **scales**. Converting data to a different scale is called a **transformation**.

LINEAR TRANSFORMATIONS: A **linear transformation** changes the original variable x into the new variable u in a way given by the following formula:

$$u = a + bx,$$

where b is nonzero and a is any number.

EFFECTS OF LINEAR TRANSFORMATIONS: Suppose that we have data $x_1, x_2, ..., x_n$ measured on some original scale with sample mean $\overline{x}$ and sample variance s_x^2. Consider transforming the data by

$$u_i = a + bx_i,$$

for $i = 1, 2, ..., n$, so that we have the "new" data

$$\begin{aligned} u_1 &= a + bx_1 \\ u_2 &= a + bx_2 \\ &\vdots \\ u_n &= a + bx_n \end{aligned}$$

Then, the mean and variance of the transformed data $u_1, u_2, ..., u_n$ are given by

$$\overline{u} = a + b\overline{x} \quad \text{and} \quad s_u^2 = b^2 s_x^2.$$

Furthermore, the standard deviation of $u_1, u_2, ..., u_n$ is given by

$$s_u = |b| s_x,$$

where $|b|$ denotes the absolute value of b. It is interesting to note that the constant a does not affect the variance or the standard deviation (why?).

Example 1.12. A common linear transformation characterizes the relationship between the Celsius and Fahrenheit temperature scales. Letting x denote the Celsius temperature, we can transform to the Fahrenheit scale u by

$$u = 32 + 1.8x.$$

You will note that this is a linear transformation with $a = 32$ and $b = 1.8$. From the web, I found the high daily temperatures recorded in June for Baghdad, Iraq (the last twenty years worth of data; i.e., 600 observations!). I plotted the data (in Celcius) in Figure 1.7. I then transformed the data to Fahrenheit using the formula above and plotted the Fahrenheit data as well. You should note the shape of the distribution is unaffected by the linear transformation. This is true in general; that is, *linear transformations do not change the shape of a distribution.*

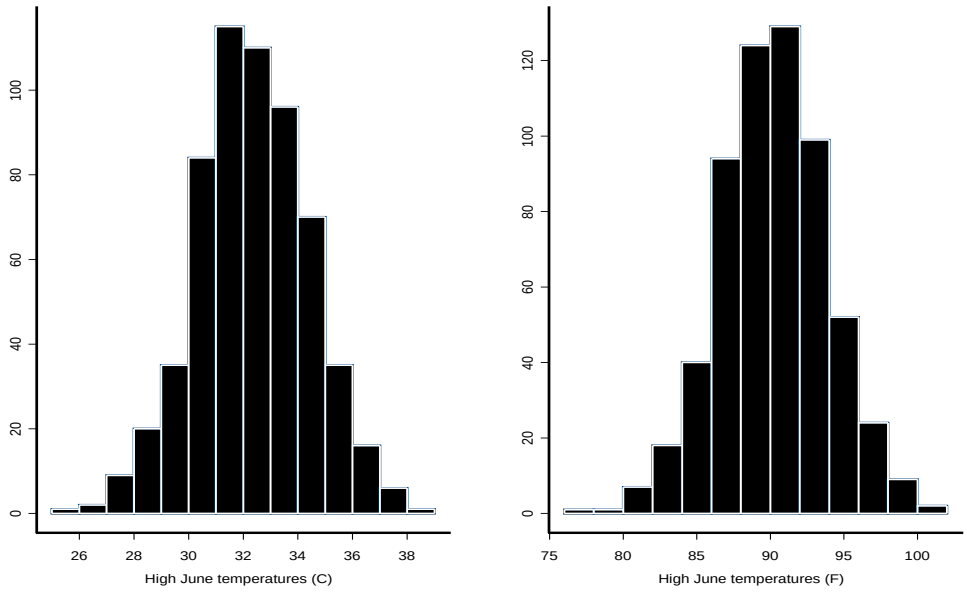


Figure 1.7: *High June temperatures in Baghdad from* 1984-2004.

MEAN, VARIANCE, AND STANDARD DEVIATION COMPARISONS: I asked the software package R to compute the mean, variance, and standard deviation of the 600 Celsius observations. I also asked for the 5-Number Summary.

```
> mean(x)
[1] 32.341
> var(x)
[1] 4.116
> stdev(x)
[1] 2.029
> summary(x)
 Min.  1st Qu.  Median  3rd Qu.  Max.
 25.1     31.0    32.4    33.6   38.4
```

We see that $\overline{x} = 32.341$, $s_x^2 = 4.116$, and $s_x = 2.029$. What are the mean, variance, and standard deviation of the data measured in Fahrenheit? I also asked R to give us these:

```
> mean(u)
[1] 90.214
> var(u)
[1] 13.335
> stdev(u)
[1] 3.652
> summary(u)
 Min.  1st Qu.  Median  3rd Qu.   Max.
 77.1     87.8    90.3     92.6  101.1
```

EXERCISE: Verify that $\overline{u} = a + b\overline{x}$, $s_u^2 = b^2 s_x^2$, and $s_u = |b| s_x$ for these data (recall that $a = 1.8$ and $b = 32$). Also, with regard to the five values in the Five-Number Summary, what are the relationships that result (for each) under a linear transformation? How is the IQR affected by a linear transformation?

1.3 Density curves and normal distributions

Example 1.13. Low infant birth weight is a fairly common problem that expecting mothers experience. In Larimer County, CO, a group of researchers studied mothers aged from 25 to 34 years during 1999 to 2003. During this time, 1,242 male birth weights were recorded. The data appear in a relative frequency histogram in Figure 1.8.

SIDE NOTE: A **relative frequency histogram** is a special histogram. Percentages are plotted on the vertical axis (instead of counts). Plotting percentages does not alter the shape of the histogram.

DENSITY CURVE: On Figure 1.8, I have added a density curve to the histogram. One can think of a **density curve** as a smooth approximation to a histogram of data. Since density curves serve as approximations to real life data, it is sometimes convenient to think about them as **theoretical models** for the variable of interest (here, birth weights).

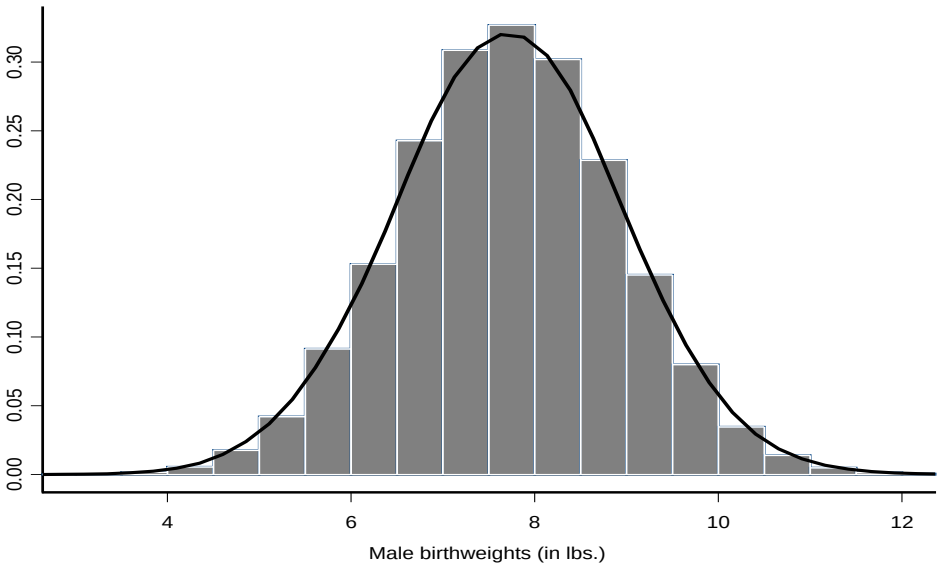


Figure 1.8: *Distribution of American male birth weights. A density curve has been superimposed over the relative frequency histogram.*

PROPERTIES: *A density curve describes the overall pattern of a distribution.* In general, a density curve associated with a quantitative variable x (e.g., birth weights, etc.) is a curve with the following properties:

1. the curve is non-negative
2. the area under the curve is one
3. the area under the curve between two values, say, a and b, represents the proportion of observations that fall in that range.

EXAMPLES: For example, the proportion of male newborns that weigh between 7 and 9 lbs is given by the **area** under the density curve *between* 7 and 9. Also, the proportion of newborns that are of low birth weight (i.e., < 5.5 lbs.) is the area under the density curve to the *left* of 5.5. In general, finding these areas might not be straightforward!

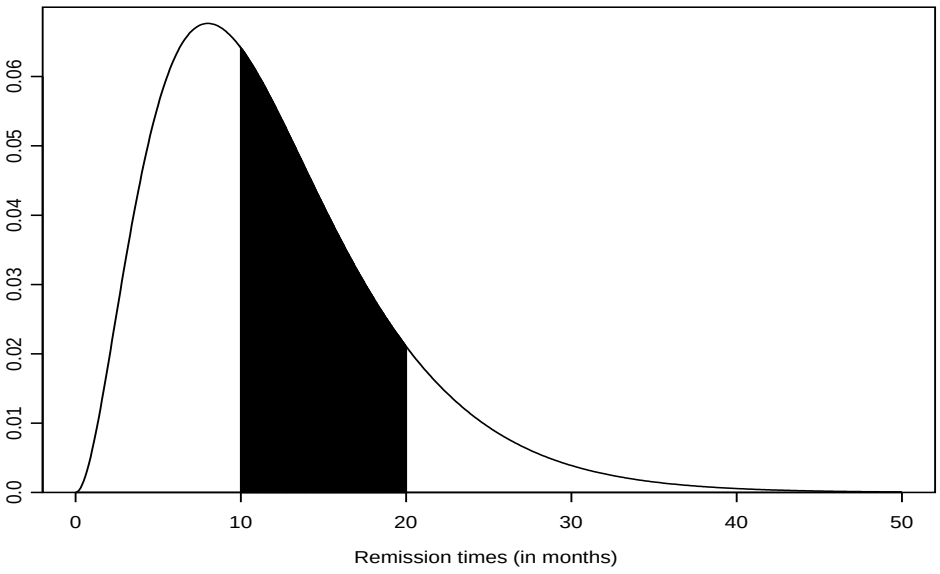


Figure 1.9: *Density curve for the remission times of leukemia patients. The shaded region represents the proportion of patient times between* 10 *and* 20 *months.*

MATHEMATICAL DEFINITION: A **density curve** is a function $f(x)$ which describes the distribution of values taken on by a quantitative variable. The function $f(x)$ has the following properties:

1. $f(x) > 0$ (nonnegative)
2. $\int_{-\infty}^{\infty} f(x)dx = 1$ (area under the curve is 1)
3. the proportion of observations that fall between a and b is given by $\int_a^b f(x)dx$ (the area under the curve between a and b).

Example 1.14. Researchers have learned that the best way to cure patients with acute lymphoid leukemia is to administer large doses of several chemotherapeutic drugs over a short period of time. Suppose that the density curve in Figure 1.9 represents the remission time for a certain group of leukemia patients. The equation of the curve is

$$f(x) = \frac{1}{128}x^2e^{-x/4},$$

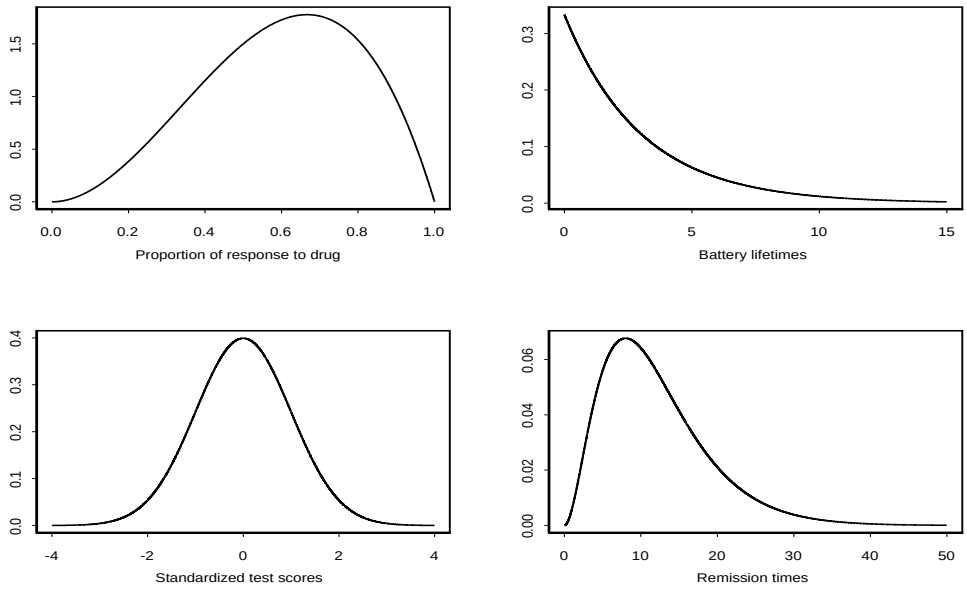


Figure 1.10: *Four density curves.*

for values of $x > 0$ (here, x is our variable representing remission time). If this is the true model for these remission times, what is the proportion of patients that will experience remission times between 10 and 20 months? That is, what is the area of the shaded region in Figure 1.9? Using calculus, we could compute

$$\int_{10}^{20} f(x)dx = \int_{10}^{20} \frac{1}{128}x^2 e^{-x/4}dx \approx 0.42.$$

Thus, about 42 percent of the patients will experience remission times between 10 and 20 months.

1.3.1 Measuring the center and spread for density curves

MAIN POINT: The notion of **center** and **spread** is the same for density curves as before. However, because we are talking about theoretical models (instead of raw data), we change our notation to reflect this. In particular,

- The **mean** for a density curve is denoted by μ.

- The **variance** for a density curve is denoted by σ^2.
- The **standard deviation** for a density curve is denoted by σ.

SUMMARY: Here is a review of the notation that we have adopted so far in the course.

	Observed data	Density curve
Mean	$\overline{x}$	μ
Variance	s^2	σ^2
Standard deviation	s	σ

IMPORTANT NOTE: The **sample values** $\overline{x}$ and s can be computed from observed data. The population values μ and σ are **theoretical values**. In practice, these values are often unknown (i.e., we can not compute these values from observed sample data).

COMPARING MEAN AND MEDIAN: The **mean** for a density curve μ can be thought of as a "balance point" for the distribution. On the other hand, the **median** for a density curve is the point that divides the area under the curve in half. The relationship between the mean and median is the same as they were before; namely,

- if a density curve is perfectly symmetric, the median and mean will be equal.
- if a density curve is skewed right, the mean will be greater than the median.
- if a density curve is skewed left, the median will be greater than the mean.

MODE: The **mode** for a density curve is the point where the density curve is at its maximum. For example, in Figure 1.9, the mode looks to be around 8 months.

1.3.2 Normal density curves

NORMAL DENSITY CURVES: The most famous (and important) family of density curves is the normal family or **normal distributions**. The function that describes the

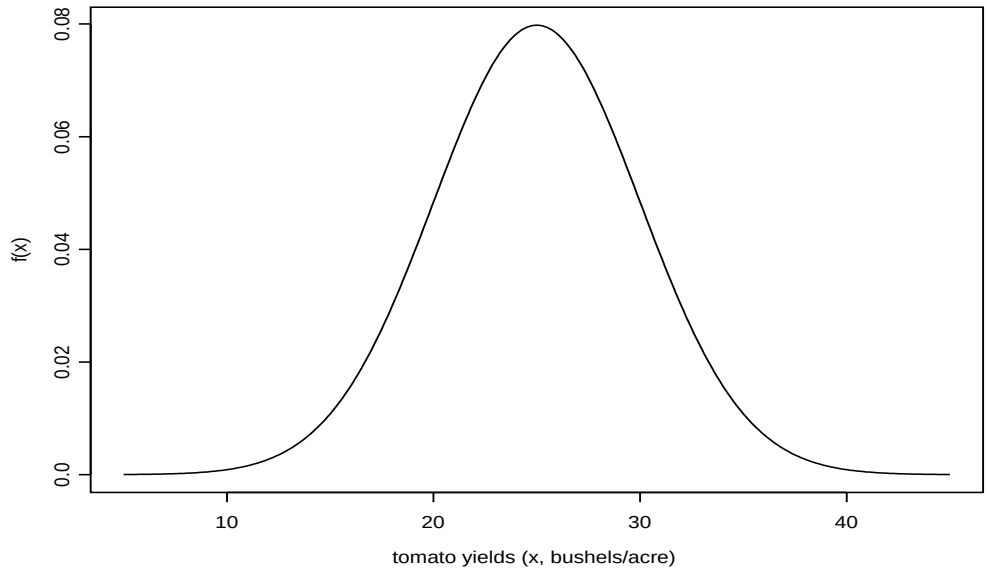


Figure 1.11: *A normal density curve with mean $\mu = 25$ and standard deviation $\sigma = 5$. A model for tomato yields.*

normal distribution with mean μ and standard deviation σ is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

FACTS: Here are some properties for the normal family of density curves.

- the curves are **symmetric** (thus, the empirical rule holds perfectly)
- area under the a normal density curve is one
- mean, median, and mode are all equal
- mound shaped and **unimodal**
- change of curvature points at $\mu \pm \sigma$.

EMPIRICAL RULE:

- 68 percent (68.26%) of the observations will be within **one** σ of μ,

- 95 percent (95.45%) of the observations will be within **two** σ of μ, and
- 99.7 percent (99.73%) of the observations will be within **three** σ of μ.

Example 1.15. For the normal density curve in Figure 1.11, which is used as a model for tomato yields (x, measured in bushels per acre), the empirical rule says that

- about 68 percent of the observations (yields) will be within 25 ± 5, or (20, 30) bushels/acre.
- about 95 percent of the observations (yields) will be within $25 \pm 2 \times 5$, or (15, 35) bushels/acre.
- about 99.7 percent of the observations (yields) will be within $25 \pm 3 \times 5$, or (10, 40) bushels/acre.

PRACTICE QUESTIONS USING EMPIRICAL RULE: Concerning Figure 1.11, a normal distribution with mean $\mu = 25$ and standard deviation $\sigma = 5$,

- what percentage of yields will be less than 20 bushels per acre?
- what percentage of yields will be greater than 35 bushels per acre?
- what percentage of yields will be between 10 and 35 bushels per acre?
- what percentage of yields will be between 15 and 30 bushels per acre?

SHORT-HAND NOTATION FOR NORMAL DISTRIBUTIONS: Because we will mention normal distributions often, a short-hand notation is useful. *We abbreviate the normal distribution with mean μ and standard deviation σ by*

$$\mathcal{N}(\mu, \sigma).$$

To denote that a variable X follows a normal distribution with mean μ and standard deviation σ, we write $X \sim \mathcal{N}(\mu, \sigma)$.

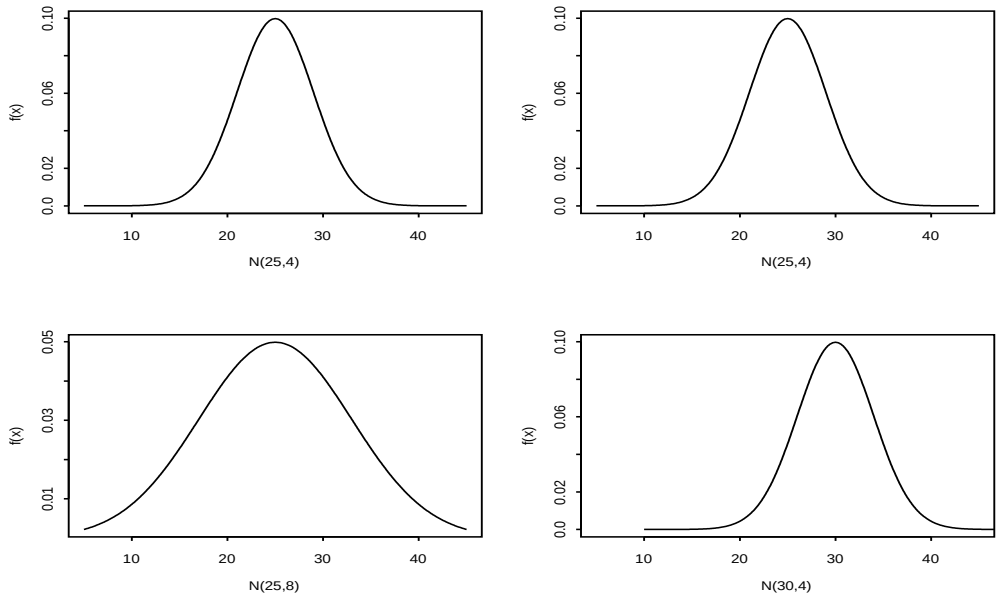


Figure 1.12: *Effects of changing μ and σ on the shape of the normal density curve.*

NORMAL DENSITY CURVE SHAPES: Figure 1.12 displays four normal distributions.

- The left two are the $\mathcal{N}(25, 4)$ and $\mathcal{N}(25, 8)$ density curves. Note how the $\mathcal{N}(25, 8)$ distribution has an increased spread (i.e., more variability).
- The right two are the $\mathcal{N}(25, 4)$ and $\mathcal{N}(30, 4)$. Note how the $\mathcal{N}(30, 4)$ distribution has been shifted over to the right (i.e., the mean has shifted).

1.3.3 Standardization

QUESTION: For the normal density curve in Figure 1.11, what is the percentage of yields between 26.4 and 37.9 bushels per acre?

GOAL: For any $\mathcal{N}(\mu, \sigma)$ distribution, we would like to compute areas under the curve. Of course, using the empirical rule, we know that we can find areas in special cases (i.e., when values happen to fall at $\mu \pm \sigma$, $\mu \pm 2\sigma$, and $\mu \pm 3\sigma$). How can we do this in general?

IMPORTANT RESULT: Suppose that the variable $X \sim \mathcal{N}(\mu, \sigma)$, and let x denote a specific value of X. We define the **standardized value** of x to be

$$z = \frac{x - \mu}{\sigma}.$$

Standardized values are sometimes called z**-scores**.

FACTS ABOUT STANDARDIZED VALUES:

- unitless quantities
- indicates how many standard deviations an observation falls above (below) the mean μ.

CAPITAL VERSUS LOWERCASE NOTATION: From here on out, our convention will be to use a **capital letter** X to denote a variable of interest. We will use a **lowercase letter** x to denote a specific value of X. This is standard notation.

STANDARD NORMAL DISTRIBUTION: Suppose that the variable $X \sim \mathcal{N}(\mu, \sigma)$. Then, the standardized variable

$$Z = \frac{X - \mu}{\sigma}.$$

has a normal distribution with mean 0 and standard deviation 1. We call Z a **standard normal variable** and write $Z \sim \mathcal{N}(0, 1)$. We call $\mathcal{N}(0, 1)$ the **standard normal distribution**.

Example 1.16. SAT math scores are normally distributed with mean $\mu = 500$ and standard deviation $\sigma = 100$. Let X denote the SAT math score so that $X \sim \mathcal{N}(500, 100)$

(a) Suppose that I got a score of $x = 625$. What is my standardized value? How is this value interpreted?
(b) What is the standardized value of $x = 342$? How is this value interpreted?
(c) Does the variable

$$\frac{X - 500}{100}$$

have a normal distribution? What is its mean and standard deviation?

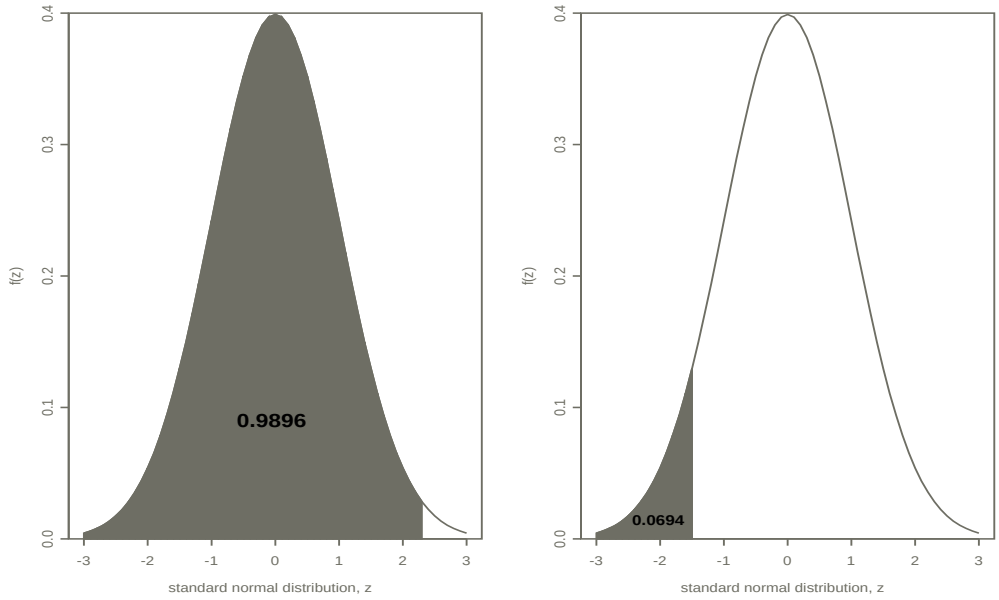


Figure 1.13: *Standard normal areas. Left: Area to the left of* $z = 2.31$. *Right: Area to the left of* $z = -1.48$.

FINDING AREAS UNDER THE STANDARD NORMAL DENSITY CURVE: **Table A** in your textbook provides areas under the standard normal curve. Note that for a particular value of z, the tabled entry gives the **area to the left** of z. For example,

- the area to the left of $z = 2.31$ is 0.9896
- the area to the left of $z = -1.48$ is 0.0694.

QUESTIONS FOR YOU:

- What is the area to the left of $z = 0.34$?
- What is the area to the right of $z = 2.31$?
- What is the area to the right of $z = -2.00$?
- What is the area between $z = -1.48$ and $z = 2.31$?

IMPORTANT: When you are making standard normal area calculations, I highly recommend that you sketch a picture of the distribution, place the value(s) of z that you are given, and shade in the appropriate area(s) you want. *This helps tremendously!*

ONLINE RESOURCE: The authors of the text have created an **applet** that will compute areas under the standard normal curve (actually for any normal curve). It is located at

`http://bcs.whfreeman.com/ips5e/`

1.3.4 Finding areas under any normal curve

FINDING AREAS: Computing the area under any normal curve can be done by using the following steps.

1. State the problem in terms of the variable of interest X, where $X \sim \mathcal{N}(\mu, \sigma)$.
2. Restate the problem in terms of Z by standardizing X.
3. Find the appropriate area using Table A.

Example 1.17. For the $\mathcal{N}(25, 5)$ density curve in Figure 1.11, what is the percentage of yields between 26.4 and 37.9 bushels per acre?

SOLUTION.

1. Here, the variable of interest is X, the yield (measured in bushels per acre). Stating the problem in terms of X, we would like to consider
$$26.4 < X < 37.9.$$
2. Now, we **standardize** by subtracting the mean and dividing by the standard deviation. Here, $\mu = 25$ and $\sigma = 5$.
$$\frac{26.4 - 25}{5} < \frac{X - 25}{5} < \frac{37.9 - 25}{5}$$
$$0.28 < Z < 2.58$$

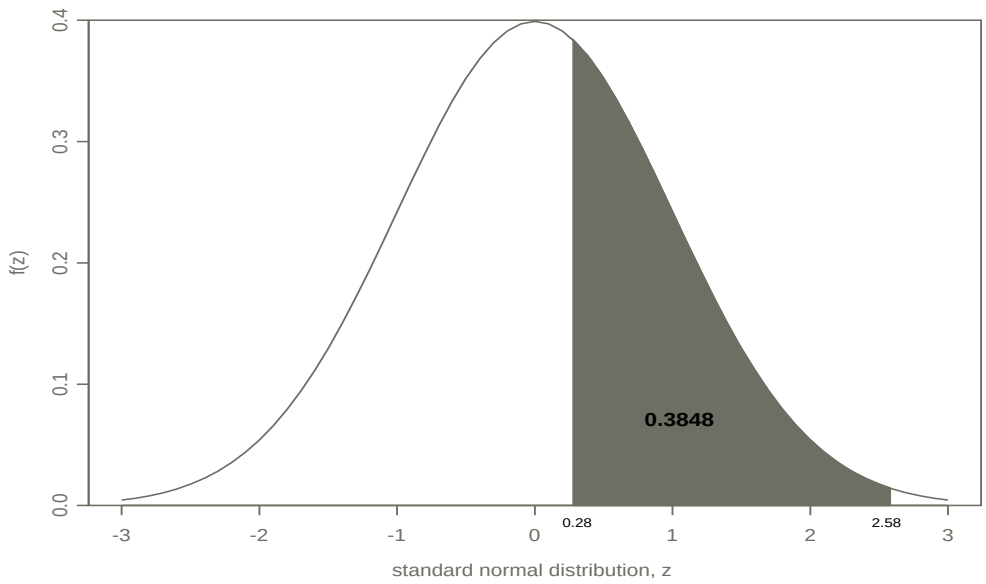


Figure 1.14: *Standard normal distribution. Area between* $z = 0.28$ *and* $z = 2.58$.

3. Finally, we find the area between $z = 0.28$ and $z = 2.58$ on the standard normal distribution using Table A.

   - the area to the left of $z = 2.58$ is 0.9951 (Table A)
   - the area to the left of $z = 0.28$ is 0.6103 (Table A).

   Thus, the area between $z = 0.28$ and $z = 2.58$ is $0.9951 - 0.6103 = 0.3848$.

ANSWER. So, about 38.48 percent of the tomato yields will be between 26.4 and 37.9 bushels per acre.

QUESTIONS FOR YOU: In Example 1.17,

- What proportion of yields will exceed 34.5 bushels per acre?
- What proportion of yields will be less than 18.8 bushels per acre?

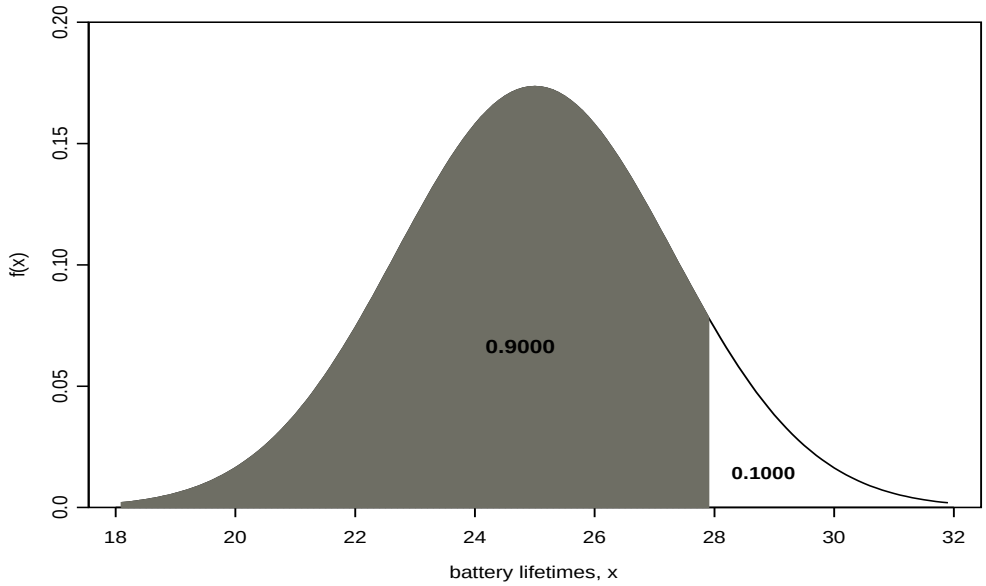


Figure 1.15: *Standard normal distribution. Unshaded area is the upper ten percent of the distribution.*

1.3.5 Inverse normal calculations: Finding percentiles

OBSERVATION: In the last subsection, our goal was to find the area under a normal density curve. This area represents a proportion of observations falling in a certain interval (e.g., between 26.4 and 37.9, etc.). Now, our goal is to go "the other way;" that is, we want to find the observed value(s) corresponding to a given proportion.

Example 1.18. The lifetime of a cardiac pacemaker battery is normally distributed with mean $\mu = 25.0$ days and standard deviation $\sigma = 2.3$ days. Ten percent of the batteries will last longer than how many days?

SOLUTION. Denote by X the lifetime of a pacemaker battery so that

$$X \sim \mathcal{N}(25.0, 2.3)$$

We want to find the 90th percentile of the $\mathcal{N}(25.0, 2.3)$ distribution. See Figure 1.15. We will solve this by first finding the 90th percentile of the standard normal distribution.

From Table A, this is given by $z = 1.28$. Note that the area to the left of $z = 1.28$ is 0.8997 (as close to 0.9000 as possible). Now, we **unstandardize**; that is, we set

$$\frac{x - 25}{2.3} = 1.28.$$

Solving for x, we get

$$x = 1.28(2.3) + 25 = 27.944.$$

Thus, 10 percent of the battery lifetimes will exceed 27.944 days.

FORMULA FOR PERCENTILES: Since, in general,

$$z = \frac{x - \mu}{\sigma},$$

we see that

$$x = \sigma z + \mu.$$

Thus, if z denotes the pth percentile of the $\mathcal{N}(0, 1)$ distribution, then $x = \sigma z + \mu$ is the pth percentile of the $\mathcal{N}(\mu, \sigma)$ distribution.

QUESTIONS FOR YOU:

- What is the 80th percentile of the standard normal distribution? The 20th?
- In Example 1.18, five percent of battery lifetimes will be below which value?
- In Example 1.18, what is the proportion of lifetimes between 21.4 and 28.2 days?

1.3.6 Diagnosing normality

Example 1.19. Since an adequate supply of oxygen is necessary to support life in a body of water, a determination of the amount of oxygen provides a means of assessing the quality of the water with respect to sustaining life. Dissolved oxygen (DO) levels provide information about the biological, biochemical, and inorganic chemical reactions occurring in aquatic environments. In a marine-biology study, researchers collected $n = 164$ water

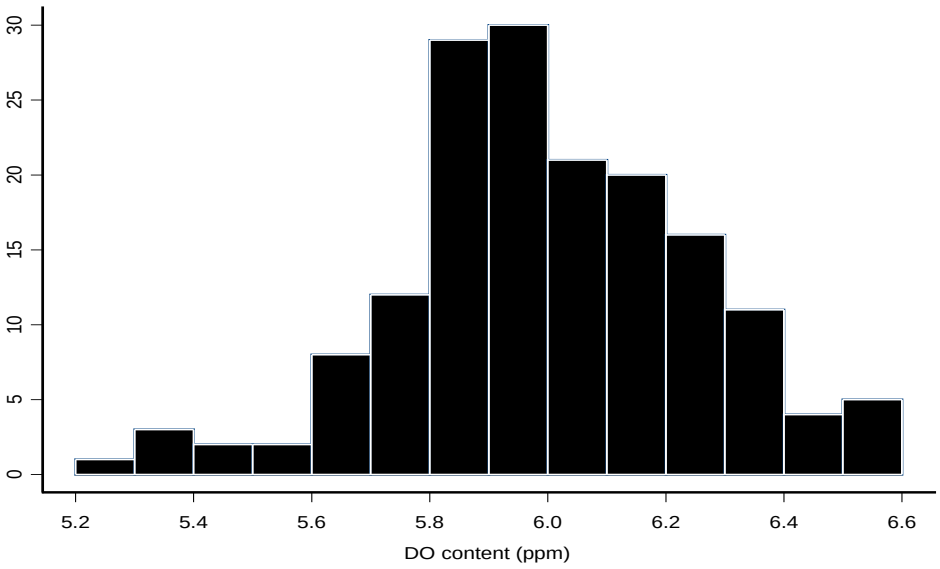


Figure 1.16: *Dissolved oxygen contents for* $n = 164$ *water specimens.*

specimens and recorded the DO content (measured in parts per million). A histogram of the data appears in Figure 1.16. *Are these data normally distributed? How can we tell?*

GOAL: Given a set of data $x_1, x_2, ..., x_n$, we would like to determine whether or not a normal distribution adequately "fits" the data.

CHECKING NORMALITY: Here are some things we can do to check whether or not data are well-represented by a normal distribution:

1. Plot the data!!!
2. Compute summary measures; check to see if the data follow the empirical rule.
3. Construct a **normal quantile plot.**

NORMAL SCORES: A set of n normal scores from the standard normal distribution are the z values which partition the density curve into $n+1$ equal areas. *That is, the normal scores are percentiles from the standard normal distribution!!*

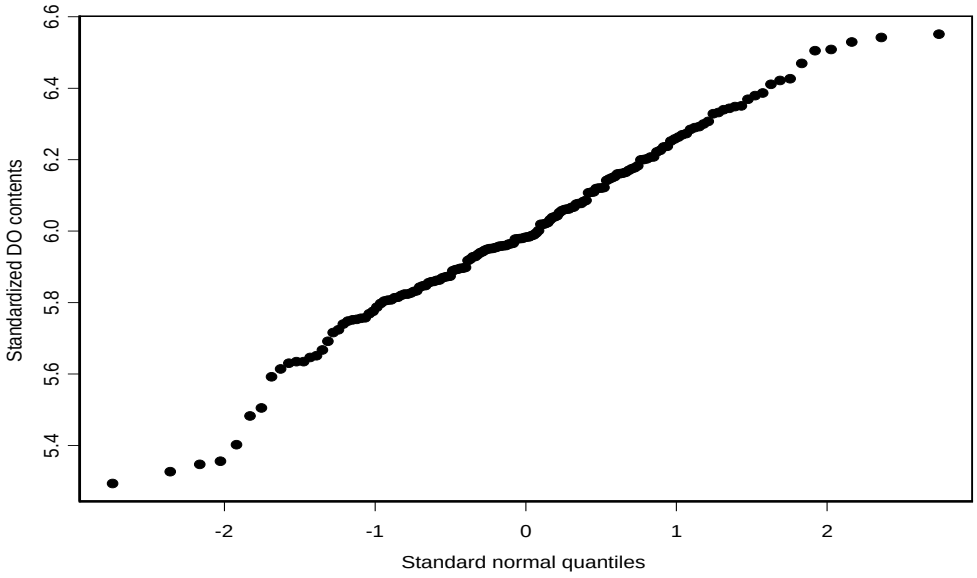


Figure 1.17: *Normal quantile plot for the DO content data in Example* 1.19.

EXERCISE. With $n = 4$, find the normal scores from the standard normal distribution.

NORMAL QUANTILE PLOTS: A **normal quantile plot** plots the observed data (ordered from low to high and then suitably standardized) versus the corresponding n normal scores from a standard normal distribution.

REALIZATION: If we plot our ordered standardized data versus the normal scores, then

- if the resulting plot is **relatively straight**, this supports the notion that the data are normally distributed (i.e., a normal density curve "fits" the data well).
- if the resulting plot has "heavy tails" and is **curved**, this supports the notion that the data are not normally distributed (i.e., a normal density curve does not "fit" the data well).

CONCLUSION: The normal quantile plot for the DO data in Example 1.19 is given in Figure 1.17. A normal distribution reasonably fits the dissolved oxygen data.

2 Looking at Data: Relationships

OVERVIEW: A problem that often arises in the biological and physical sciences, economics, industrial applications, and biomedical settings is that of investigating the mathematical relationship between *two or more* variables.

EXAMPLES:

- amount of alcohol in the body (BAL) versus body temperature (degrees C)
- weekly fuel consumption (degree days) versus house size (square feet)
- amount of fertilizer (pounds/acre) applied versus yield (kg/acre)
- sales ($1000s) versus marketing expenditures ($1000s)
- HIV status (yes/no) versus education level (e.g., primary, secondary, college, etc.)
- gender (M/F) versus promotion (yes/no). Are promotion rates different?
- Remission time (in days) versus treatment (e.g., surgery/chemotherapy/both)

REMARK: In Chapter 1, our focus was primarily on graphical and numerical summaries of data for a single variable (categorical or quantitative). In this chapter, we will largely look at situations where we have two **quantitative** variables.

TERMINOLOGY: Two variables measured on the same individuals are said to be **associated** if values of one variable tend to occur with values of the other variable.

Example 2.1. Many fishes have a lateral line system enabling them to experience *mechanoreception*, the ability to sense physical contact on the surface of the skin or movement of the surrounding environment, such as sound waves in air or water. In an experiment to study this, researchers subjected fish to electrical impulses. The frequency (number per second) of electrical impulses (EI) emitted from one particular fish was measured at several temperatures (measured in Celcius); the data are listed in Table 2.5.

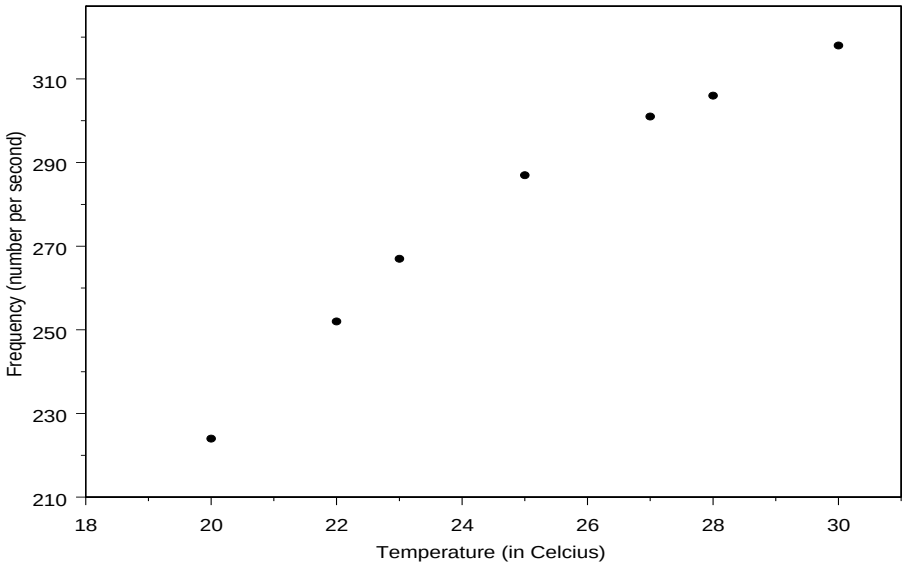


Figure 2.18: *EI frequency at different temperatures.*

The **scatterplot** for the data appears in Figure 2.18. It is clear that these variables are strongly related. As the water temperature increases, there is a tendency for the frequency of impulses to increase as well.

Table 2.5: *Electrical impulse data.*

Temperature	Frequency	Temperature	Frequency
20	224	27	301
22	252	28	306
23	267	30	318
25	287		

VARIABLES: In studies where there are multiple variables under investigation (e.g., temperature, EI frequency), it is common that one desires to study how one variable is affected by the other(s). In some problems (not all), it makes sense to focus on the behavior of one variable and, in particular, determine how another variable influences it.

In a scientific investigation, the main variable under consideration is called the **response variable**. An **explanatory variable** explains or causes changes in the response variable (there may be more than one explanatory variable!). In Example 2.1, EI frequency is the response variable, and temperature is the explanatory variable.

NOTATION: We denote the explanatory variable by x and the response variable by y.

NOTE: Explanatory variables are sometimes called **independent variables**. A response variable is sometimes called a **dependent variable**.

2.1 Scatterplots

SCATTERPLOTS: A **scatterplot** is a graphical display that plots observations on two quantitative variables. It is customary that the response variable is placed on the vertical axis. The explanatory variable is placed on the horizontal. Scatterplots give a visual impression of how the two variables behave together.

INTERPRETING SCATTERPLOTS: It is important to describe the overall pattern of a scatterplot by examining the following:

- **form**; are there curved relationships or different clusters of observations?
- **direction**; are the two variables positively related or negatively related?
- **strength**; how strong is the relationship? Obvious? Weak? Moderate?
- the presence of **outliers**.

LINEAR RELATIONSHIPS: If the form of the scatterplot looks to resemble a straight-line trend, we say that the relationship between the variables is **linear**.

TERMINOLOGY: Two variables are **positively related** if they tend to increase together. They are **negatively related** if an increase in one is associated with a decrease in the other. The data in Figure 2.18 display a strong positive relationship.

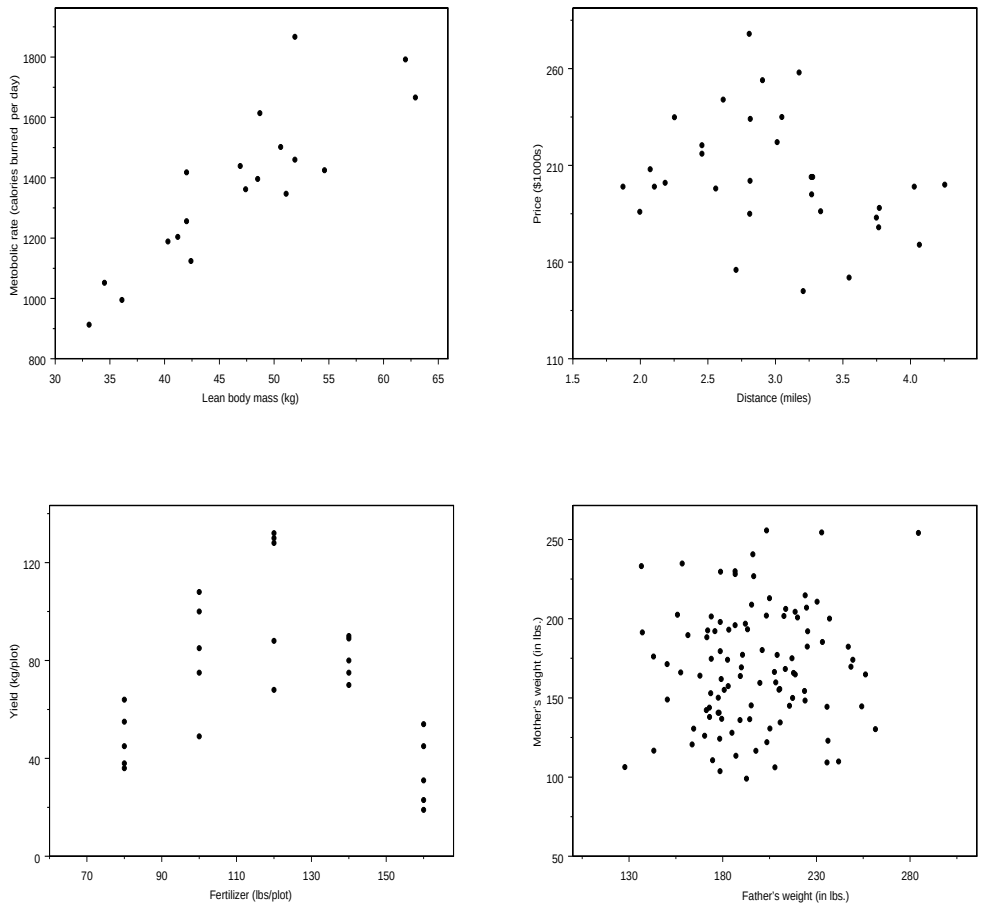


Figure 2.19: *Four scatterplots. Upper left: positive linear relationship. Upper right: mild linear negative relationship. Lower left: curved relationship. Lower right: random scatter.*

ADDING CATEGORICAL VARIABLES TO SCATTERPLOTS: In some situations, we might want to add a third variable to a scatterplot. As long as this **third** variable is categorical in nature, we can do this by using different plotting symbols for the levels of the categorical variable.

Example 2.2. An engineer is studying the effects of the pH for a cleansing tank and polymer type on the amount of suspended solids in a coal cleansing system. Data from the experiment are given in Table 2.6. The engineer believes that the explanatory variable pH (X) is important in describing the response variable Y, the amount of suspended

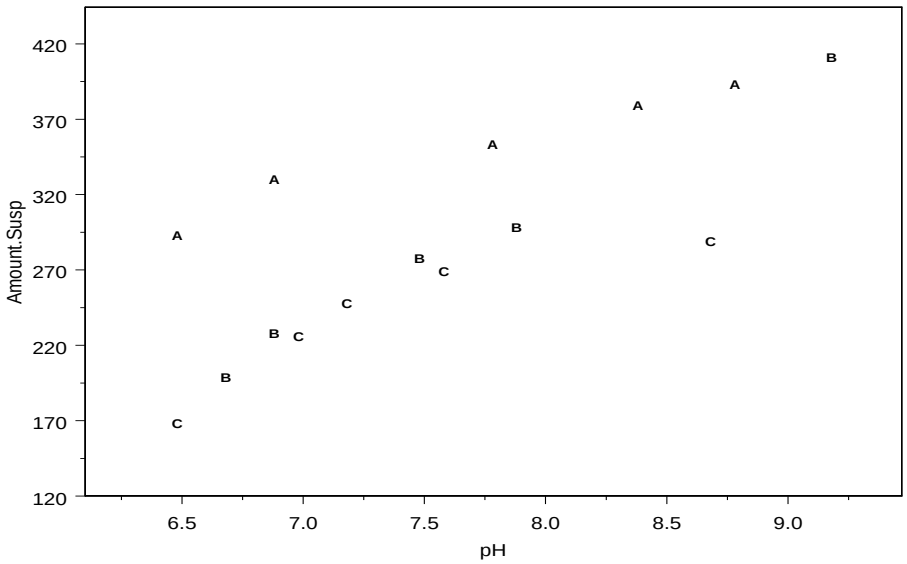


Figure 2.20: *Amount of suspended material, as a function of pH, for three polymers.*

solids (measured in ounces). However, she is also studying three different polymer types (generically denoted by A, B, and C). In Figure 2.20, different **plotting symbols** are used to differentiate among the three polymer types. What is the relationship between the amount of suspended solids and pH for each polymer?

Table 2.6: *Cleansing data for different polymers.*

Polymer A		Polymer B		Polymer C	
y	x	y	x	y	x
292	6.5	410	9.2	167	6.5
329	6.9	198	6.7	225	7.0
352	7.8	227	6.9	247	7.2
378	8.4	277	7.5	268	7.6
392	8.8	297	7.9	288	8.7

SIDE-BY-SIDE BOXPLOTS: When we have a quantitative response variable Y and a categorical explanatory variable X, we can display the relationship between these vari-

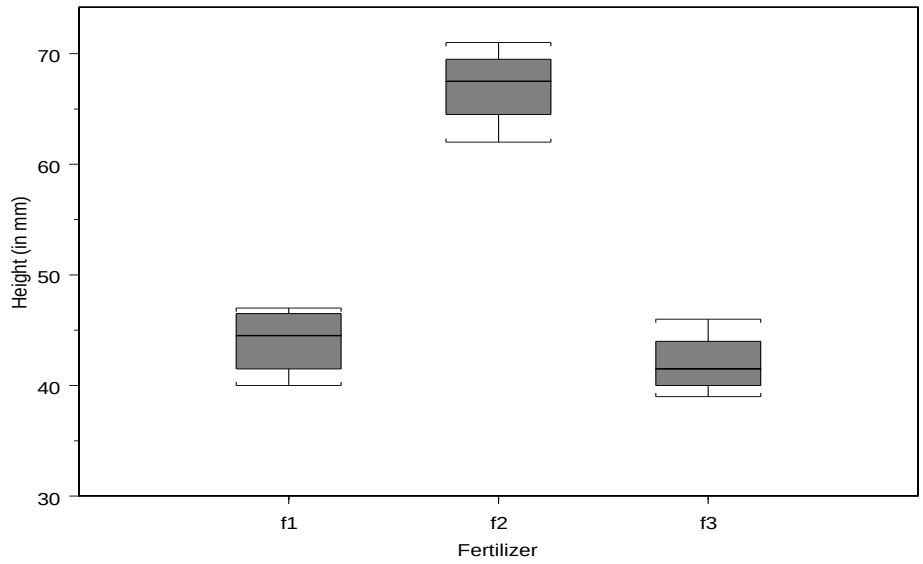


Figure 2.21: *Seedlings height data for three fertilizers.*

ables using side-by-side boxplots. These plots are very helpful with data from **designed experiments** where the response variable is often quantitative and the goal is often to compare two or more treatments (so that the categorical variable is "treatment").

Example 2.3. The operators of a nursery would like to investigate differences among three fertilizers (denoted by f1, f2, and f3) they might use on plants they are growing for commercial sale. The researchers have 24 seedlings and decide to use 8 seedlings for each fertilizer. At the end of six weeks, the heights of each seedlings, Y (measured in mm), are collected. The data from the experiment are displayed in Figure 2.21.

2.2 Correlation

SCENARIO: We would like to study the relationship between **two quantitative variables**, x and y. We observe the pair (X, Y) on each of n individuals in our sample, and we wish to use these data to say something about the relationship.

NOTE: Scatterplots give us graphical displays of the relationship between two quantitative variables. We now wish to summarize this relationship numerically.

TERMINOLOGY: The **correlation** is a numerical summary that describes the strength and direction of the linear relationship between two quantitative variables. With a sample of n individuals, denote by x_i and y_i the two measurements for the ith individual. The correlation is computed by the following formula

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

where $\overline{x}$ and $\overline{y}$ are the sample means and s_x and s_y are the sample standard deviations.

REMARK: Unless n is small, it is often best to use statistical software to compute the correlation. You should note that the terms

$$\frac{x_i - \overline{x}}{s_x} \quad \text{and} \quad \frac{y_i - \overline{y}}{s_y},$$

are the **sample standardized values** of x_i and y_i, respectively.

PROPERTIES OF THE CORRELATION:

- The correlation r is a **unitless number**; that is, there are no units attached to it (e.g., dollars, mm, etc.).
- It also makes no difference what you call x and what you call y; the correlation will be the same.
- The correlation r *always* satisfies

$$-1 \leq r \leq 1.$$

- If $r = 1$, then all data lie on a straight line with **positive** slope. If $r = -1$, then all the data lie on a straight line with **negative** slope. If $r = 0$, then there is **no linear relationship** present in the data.
- When $0 < r < 1$, there is a tendency for the values to vary together in a positive way (i.e., a positive linear relationship). When $-1 < r < 0$ there is a tendency for the values to vary together in a negative way (i.e., a negative linear relationship).

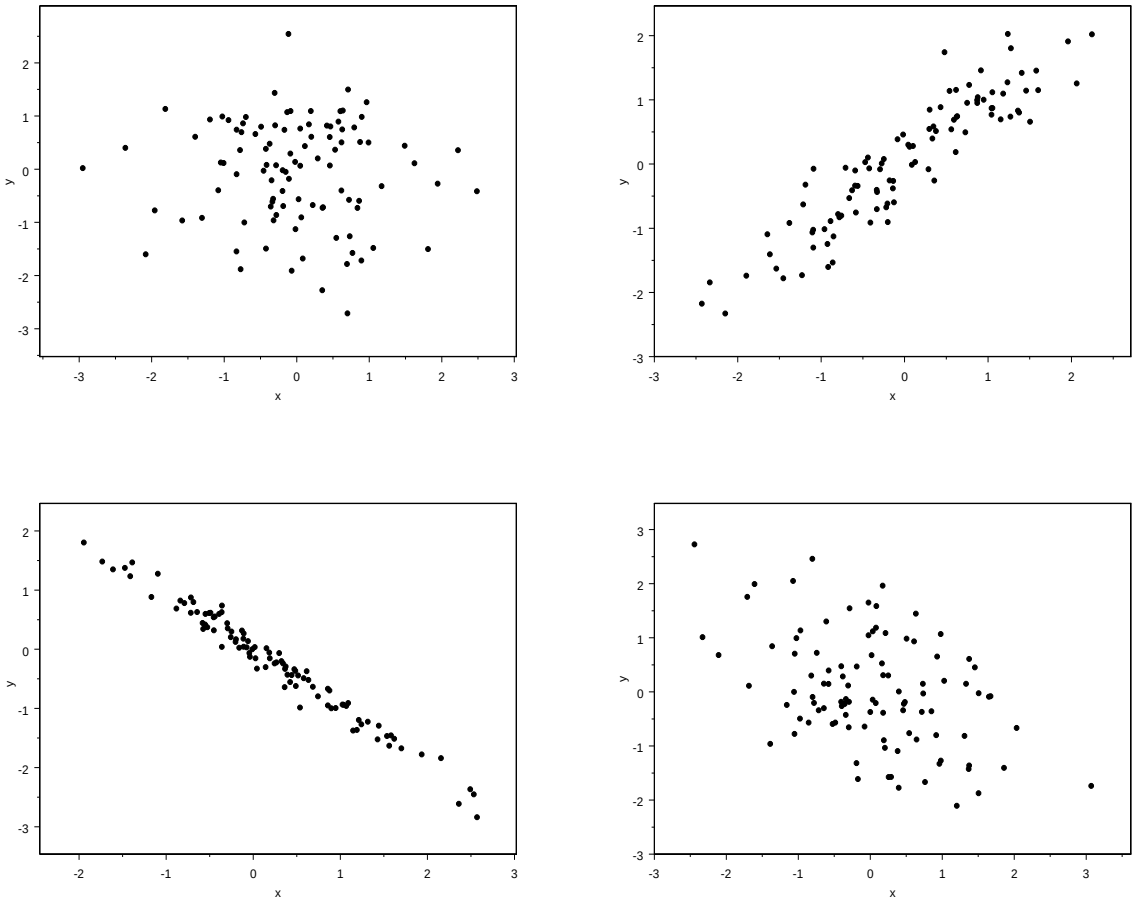


Figure 2.22: *Four scatterplots using data generated from Minitab. Upper left:* $r = 0$. *Upper right:* $r = 0.9$. *Lower left:* $r = -0.99$. *Lower right:* $r = -0.5$.

- *The correlation only measures linear relationships!!* It does not describe a curved relationship, no matter how strong that relationship is!
- Thus, we could have two variables X and Y that are perfectly related, but the correlation still be zero!! This could occur if the variables are related **quadratically**, for example. See Example 2.5.
- The value of r could be highly affected by **outliers**. This makes sense since sample means and sample standard deviations are affected by outliers (and these values are required to compute r).

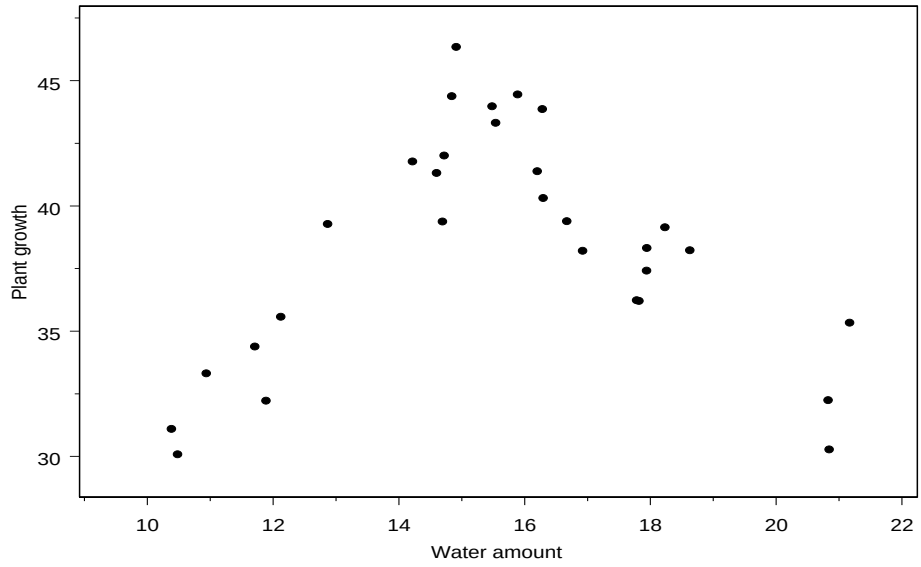


Figure 2.23: *Plant growth versus water amount.*

WARNING: The correlation is by no means a complete measure of a bivariate (i.e., two-variable) data set. *Always plot your data before computing the correlation!!*

Example 2.4. In Example 2.1, the correlation between EI frequency and temperature is $r = 0.981$. This value suggests that there is a strong positive linear relationship between the two variables.

Example 2.5. Researchers are trying to understand the relationship between the amount of water applied to plots (measured in cm) and total plant growth (measured in cm). A sample of $n = 30$ plots is taken from different parts of a field. The data from the sample is given in Figure 2.23. Using Minitab, the correlation between plant growth and water amount is $r = 0.088$. This is an example where the two variables under investigation (water amount and plant growth) have a very strong relationship, but the correlation is near zero. This occurs because the relationship is not linear; rather, it is **quadratic**. An investigator that did not plot these data and only looked at the value of r could be lead astray and conclude that these variables were not related!

2.3 Least-squares regression

REMARK: Correlation and scatterplots help us to document relationships (correlation only helps us assess *linear relationships*). The statistical technique known as **regression** allows us to formally model these relationships. Regression, unlike correlation, requires that we have an explanatory variable and a response variable.

NOTE: In this course, we will restrict attention to **regression models** for linear relationships with a single explanatory variable (this is called **simple linear regression**). Techniques for handling nonlinear relationships and/or more than one explanatory variable will be explored in the subsequent course.

Example 2.6. The following data are rates of oxygen consumption of birds (y) measured at different temperatures (x). Here, the temperatures were set by the investigator, and the O_2 rates, y, were observed for these particular temperatures.

x, (degrees Celcius)	-18	-15	-10	-5	0	5	10	19
y, (ml/g/hr)	5.2	4.7	4.5	3.6	3.4	3.1	2.7	1.8

The scatterplot of the data appears in Figure 2.24. There is clearly a negative linear relationship between the two variables.

REGRESSION MODELS: A **straight-line regression model** consists of two parts:

1. a straight-line **equation** that summarizes the general tendency of the relationship between the two variables, and
2. a **measure of variation** in the data around the line.

GOALS: We first need an approach to compute the equation of the line. Then we will look at a numerical summary that measures the variation about the line. Of course, we would like for the variation about the line to be **small** so that the regression line "fits" the data well.

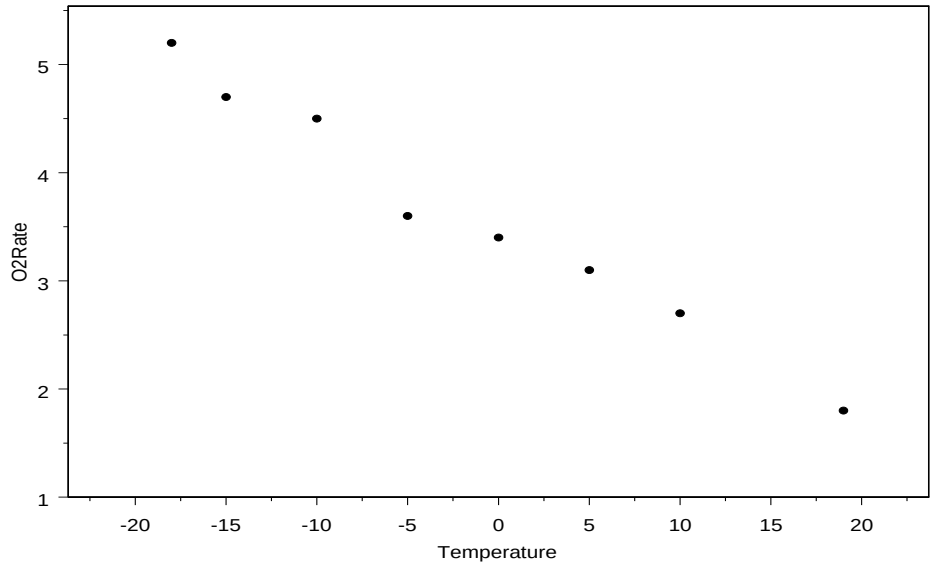


Figure 2.24: *Bird oxygen rate data for different temperatures.*

USEFULNESS: We can use the resulting regression equation to

- quantify the relationship between Y and X
- use the relationship to **predict** a new response y^* that we might observe at a given value, say, x^* (perhaps one not included in the experiment or study)
- use the relationship to **calibrate**; that is, given a new y value we might see, say, y^*, for which the corresponding value x is unknown, estimate the value x.

SCENARIO: We would like to find the equation of a straight line that best describes the relationship between **two quantitative variables**. We observe the pair (x, y) on each of n individuals in our sample, and we wish to use these data to compute the equation of the "best-fit line."

STRAIGHT LINE EQUATIONS: A REVIEW: Suppose that y is a response variable (plotted on the vertical axis) and that x is an explanatory variable (plotted on the

horizontal axis). A **straight line** relating y to x has an equation of the form

$$y = a + bx,$$

where the constant b represents the **slope** of the line and a is the **y-intercept**.

INTERPRETATION: The slope of a regression line gives the amount by which the response y changes when x increases by one unit. The y-intercept gives the value of the response y when $x = 0$.

2.3.1 The method of least squares

TERMINOLOGY: When we say, "fit a regression model," we basically mean that we are finding the values of a and b that are most consistent with the observed data. The **method of least squares** provides a way to do this.

RESIDUALS: For each y_i, and given values of a and b, note that the quantity

$$e_i = y_i - (a + bx_i)$$

measures the *vertical distance* from y_i to the line $a + bx_i$. This distance is called the ith **residual** for particular values of a and b. If a point falls above the line in the y direction, the residual is positive. If a point falls below the line in the y direction, the residual is negative. See Figure 2.25. We will discuss residuals more thorougly in Section 2.4.

RESIDUAL SUM OF SQUARES: A natural way to measure the overall deviation of the observed data from the line is with the **residual sum of squares**. This is given by

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

THE METHOD OF LEAST SQUARES: The most widely-accepted method for finding a and b is using **the method of least squares**. The method says to select the values of a and b that **minimize** the sum of squared residuals. A calculus argument can be used to find equations for a and b; these are given by

$$b = r \times \frac{s_y}{s_x} \quad \text{and} \quad a = \overline{y} - b\overline{x}.$$

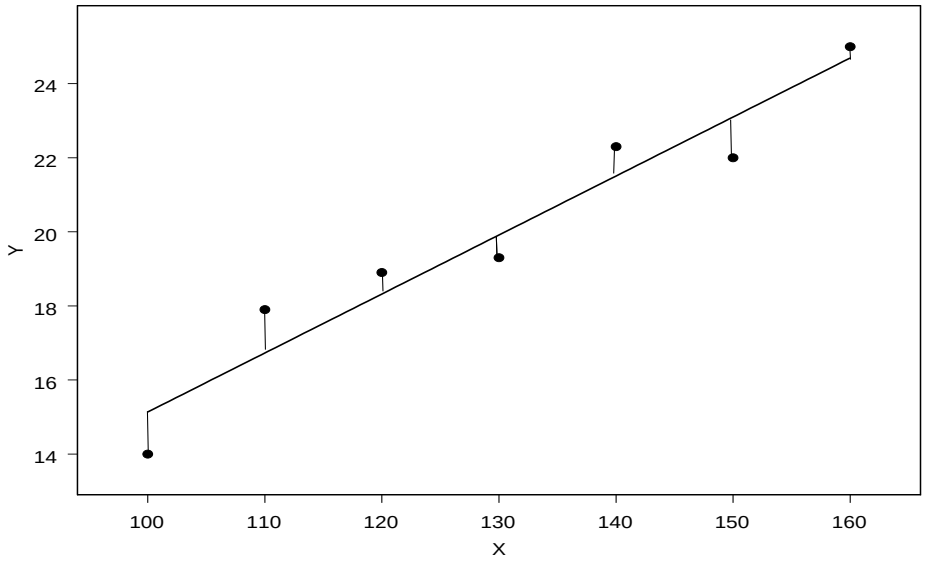


Figure 2.25: *A scatterplot with straight line and residuals associated with the straight line.*

EQUATION OF THE LEAST-SQUARES REGRESSION LINE: With the values of a and b just mentioned, the equation of the least-squares regression line is

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

We use the notation $\widehat{y}$ to remind us that this is a regression line **fit** to the data (x_i, y_i); $i = 1, 2, ..., n$.

TERMINOLOGY: The values $\widehat{y}_i = a + bx_i$ are called **fitted values**. These are values that fall directly on the line. We should see that the residuals are given by

$$\begin{aligned} e_i &= y_i - \widehat{y}_i \\ &= \text{Observed } y - \text{Fitted } y. \end{aligned}$$

REMARK: Computing the least-squares regression line by hand is not recommended because of the intense computations involved. It is best to use software. However, it is interesting to note that the values of a and b depend on five numerical summaries that we already know; namely, $\overline{x}$, $\overline{y}$, s_x, s_y, and r!!

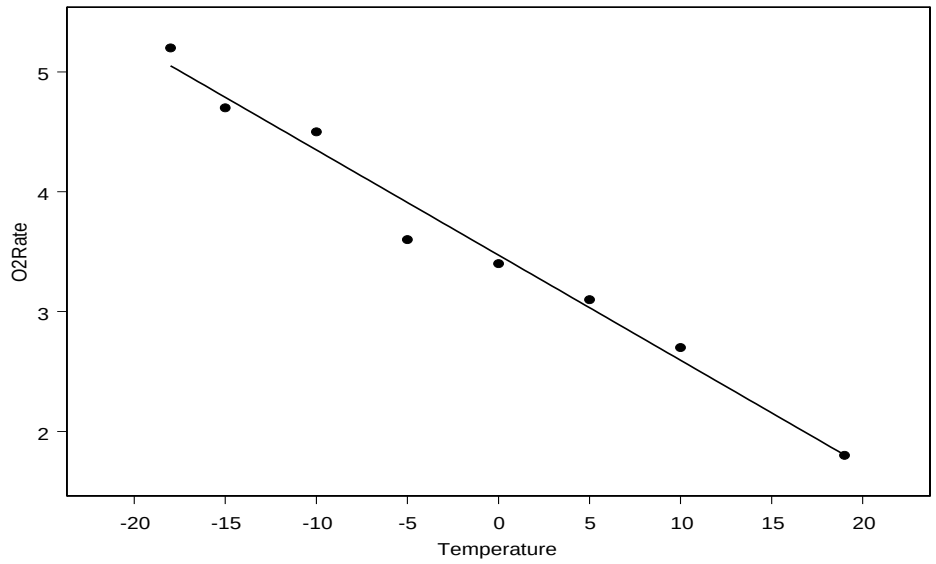


Figure 2.26: *Bird oxygen rate data for different temperatures. The least-squares regression line is superimposed.*

Example 2.7. We now compute the least-squares regression line for the bird-oxygen rate data in Example 2.6. Using Minitab, we obtain the following output.

```
The regression equation is O2Rate = 3.47 - 0.0878 Temperature

Predictor        Coef   SE Coef       T       P
Constant      3.47142   0.06012   57.74   0.000
Temperature  -0.08776   0.00499  -17.58   0.000

S = 0.168249   R-Sq = 98.1%   R-Sq(adj) = 97.8%
```

From the output, we see that $a = 3.47$ and $b = -0.0878$. Thus, using symbols, our best-fit regression line is

$$\widehat{y} = 3.47 - 0.0878x.$$

INTERPRETATION:

- The slope $b = -0.0878$ is interpreted as follows: "for a one-unit (degree) increase in temperature, we would expect for the oxygen rate to **decrease** by 0.0878 ml/g/hr."
- The y-intercept $a = 3.47$ is interpreted as follows: "for a temperature of $x = 0$, we would expect the oxygen rate to be 3.47 ml/g/hr."

2.3.2 Prediction and calibration

PREDICTION: One of the nice things about a regression line is that we can make **predictions**. For our oxygen data in Example 2.6, suppose we wanted to predict a future oxygen consumption rate (for a new bird used in the experiment, say) when the temperature is set at $x^* = 2.5$ degrees. Using our regression equation, the predicted value y^* is given by

$$y^* = 3.47 + (-0.0878)(2.5) = 3.252 \text{ ml/g/hr}.$$

Thus, for a bird subjected to 2.5 degrees Celsius, we would expect its oxygen rate to be approximately 3.252 ml/g/hr.

EXTRAPOLATION: It is sometimes desired to make predictions based on the fit of the straight line for values of x^* outside the range of x values used in the original study. This is called **extrapolation**, and can be very dangerous. In order for our inferences to be valid, we must believe that the straight line relationship holds for x values **outside** the range where we have observed data. In some situations, this **may** be reasonable; in others, we may have no theoretical basis for making such a claim without data to support it. Thus, it is very important that the investigator have an honest sense of the relevance of the straight line model for values outside those used in the study if inferences such as estimating the mean for such x^* values are to be reliable.

CALIBRATION: In a standard prediction problem, we are given a value of x (e.g., temperature) and then use the regression equation to solve for y (e.g., oxygen rate). A

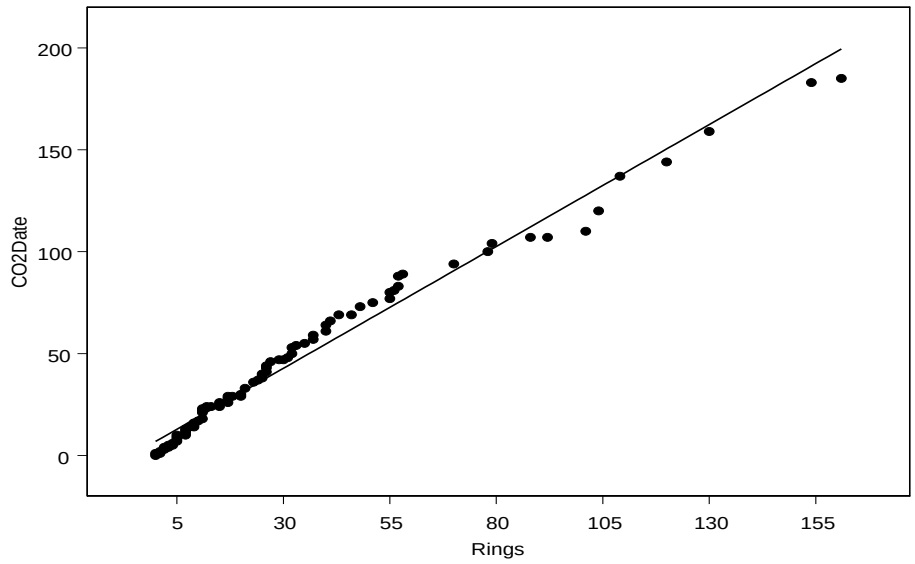


Figure 2.27: *Tree ages measured by two different methods.*

calibration problem is the exact opposite! Namely, suppose now that we have a value of y, say, y^*, and the goal is to estimate the unknown corresponding value of x, say, x^*.

Example 2.8. Consider a situation where interest focuses on two different methods of calculating the age of a tree. One way is by counting tree rings. This is considered to be very accurate, but requires sacrificing the tree. Another way is by a carbon-dating process. Suppose that data are obtained for $n = 100$ trees on age by the counting method (x) and age by carbon dating (y), both measured in years. A scatterplot of these data, with the straight-line fit, is given in Figure 2.27. Suppose that the carbon-dating method is applied to a new tree not in the study yielding an age $y^* = 34$ years. What can we say about the true age of the tree, x^*, (that is, its age by the very accurate counting method) without sacrificing the tree?

ANALYSIS: Here, we are given the value of the response $y^* = 34$ (obtained from using the carbon-dating method). The idea is to use the least-squares regression line to estimate x^* (the corresponding value using the ring-counting method). The almost obvious solution

arises from solving the regression equation $y = a + bx$ for x to obtain

$$x = \frac{y - a}{b}.$$

For our tree age data in Example 2.8, I used the R software package to compute the regression line as

$$\widehat{y} = 6.934 + 1.22x.$$

Thus, for a tree yielding an age $y^* = 34$ years (using carbon dating) the true age of the tree, x^*, is estimated as

$$x^* = \frac{34 - 6.934}{1.22} = 22.19 \text{ years}.$$

2.3.3 The square of the correlation

SUMMARY DIAGNOSTIC: In a regression analysis, one way to measure how well a straight line fits the data is to compute the **square of the correlation** r^2. This is interpreted as *the proportion of total variation in the data explained by the straight-line relationship with the explanatory variable.*

NOTE: Since $-1 \leq r \leq 1$, it must be the case that

$$0 \leq r^2 \leq 1.$$

Thus, an r^2 value "close" to 1 is often taken as evidence that the regression model does a good job at describing the variability in the data.

IMPORTANT: It is critical to understand what r^2 does and does not measure. Its value is computed under the assumption that the straight-line regression model **is correct**. Thus, if the relationship between X and Y really is a straight line, r^2 assesses how much of the variation in the data may actually be attributed to that relationship rather than just to inherent variation.

- If r^2 is small, it may be that there is a lot of random inherent variation in the data, so that, although the straight line is a reasonable model, it can only explain so much of the observed overall variation.

- Alternatively, r^2 may be close to 1, but the straight-line model may not be the most appropriate model! In fact, r^2 may be quite "high," but, in a sense, is irrelevant, because it assumes the straight line model is correct. In reality, a better model may exist (e.g., a quadratic model, etc.).

Example 2.9. With our bird-oxygen rate data from Example 2.6, I used Minitab to compute the correlation to be $r = -0.9904$. The square of the correlation is

$$r^2 = (-0.9904)^2 = 0.9809.$$

Thus, about 98.1 percent of the variation in the oxygen rate data, Y, is explained by the least squares regression of Y on temperature (X). This is a very high percentage! The other 1.9 percent is explained by other variables (not accounted for in our straight-line regression model).

TRANSFORMATIONS: In some problems, it may be advantageous to perform regression calculations on a **transformed scale**. Sometimes, it may be that data measured on a different scale (e.g., square-root scale, log scale, etc.) will obey a straight-line relationship whereas the same data measured on the original scale (before transforming) do not! For more information on transformations, see Example 2.14 (MM) on pages 143-4.

2.4 Cautions about correlation and regression

A CLOSER LOOK AT RESIDUALS: When we fit a regression equation to data, we are using a mathematical formula to explain the relationship between variables. However, we know that very few relationships are perfect! Thus, the residuals represent the "left-over" variation in the response after fitting the regression line. As noted earlier,

$$\begin{aligned} e_i &= y_i - \widehat{y}_i \\ &= \text{Observed } y - \text{Predicted } y. \end{aligned}$$

The good news is that software packages (such as SAS or Minitab) will store the residuals for you. *An important part of any regression analysis is looking at the residuals!*

Example 2.10. With our bird-oxygen rate data from Example 2.6, one of the observed data pairs was $(x_i, y_i) = (5, 3.1)$. In this example, we will compute the least-squares residual associated with this observation. From Example 2.7, we saw that the least-squares regression line was

$$\widehat{y} = 3.47 - 0.0878x.$$

Thus, the **fitted value** associated with $x = 5$ is

$$\widehat{y}_i = 3.47 - 0.0878(5) = 3.031$$

and the **residual** is

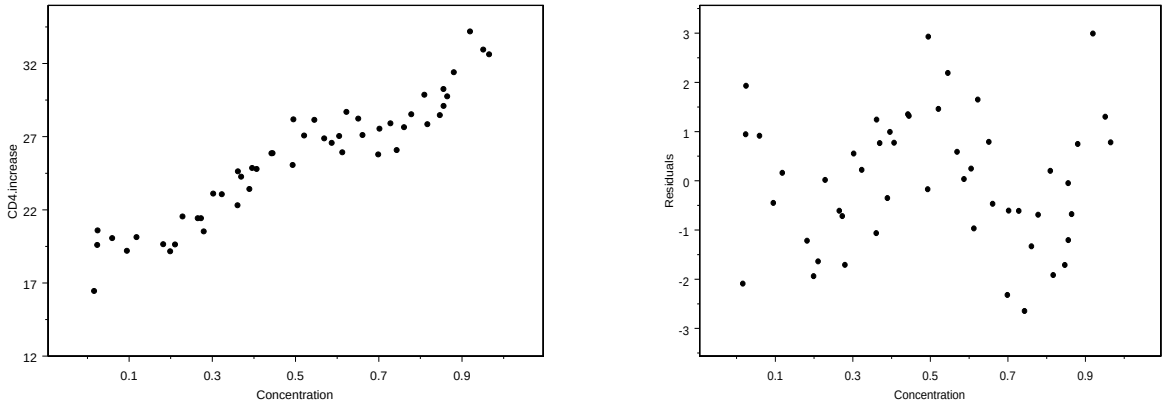
$$\begin{aligned} e_i &= y_i - \widehat{y}_i \\ &= 3.1 - 3.031 = 0.069. \end{aligned}$$

The other seven residuals from Example 2.6 are computed similarly.

2.4.1 Residual plots

TERMINOLOGY: A **residual plot** is a scatterplot of the regression residuals against the explanatory variable. Residual plots help us assess the fit of a regression line. In particular, *residual plots that display nonrandom patterns suggest that there are some problems with our straight-line model assumption.*

Example 2.11. In order to assess the effects of drug concentration (X) on the resulting increase in CD4 counts (Y), physicians used a sample of $n = 50$ advanced HIV patients with different drug concentrations and observed the resulting CD4 count increase. Data from the study appear in Figure 2.28(a). There looks to be a significant linear trend between drug concentration and CD4 increase. The residual plot from the straight-line fit is in Figure 2.28(b). This plot may look random at first glance, but, upon closer inspection, one will note that there is a "w-shape" in it. This suggests that the true relationship between CD4 count increase and drug concentration is curved.



(a) Scatterplot. (b) Residual plot.

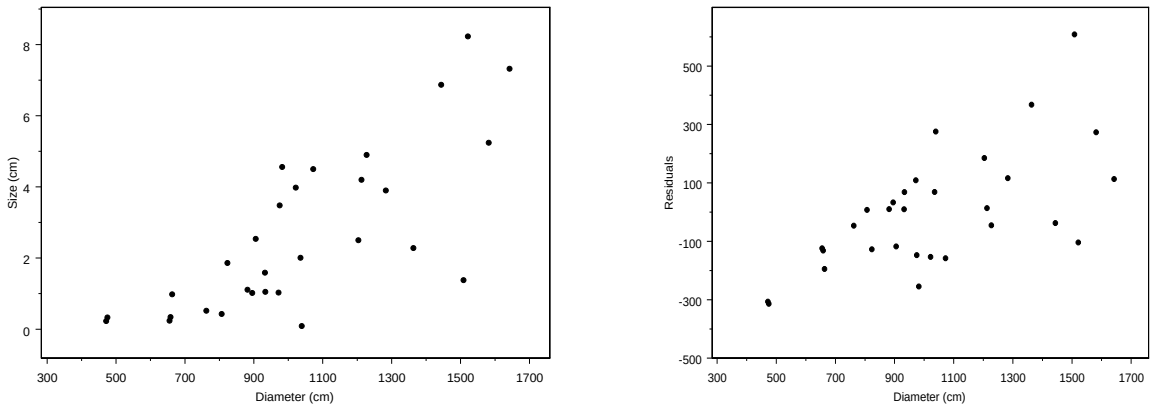
Figure 2.28: HIV *study.* CD4 *count increase versus drug concentration.*

INTERESTING FACT: For any data set, the mean of the residuals from the least-squares regression is always zero! This fact is useful when interpreting residual plots.

Example 2.12. An entomological experiment was conducted to study the survivability of stalk borer larvae. It was of interest to develop a model relating the mean size of larvae (cm) as a function of the stalk head diameter (cm). Data from the experiment appear in Figure 2.29(a). There looks to be a moderate linear trend between larvae size and head diameter size. The residual plot from the straight-line regression fit is in Figure 2.28(b). This plot displays a "fanning out" shape; this suggests that the variability increases for larger diameters. This pattern again suggests that there are problems with the straight-line model we have fit.

2.4.2 Outliers and influential observations

OUTLIERS: Another problem is that of **outliers**; i.e., data points that do not fit well with the pattern of the rest of the data. In straight-line regression, an outlier might be an observation that falls far off the apparent approximate straight line trajectory followed by the remaining observations. Practitioners often "toss out" such anomalous points, which may or may not be a good idea. If it is clear that an "outlier" is the result of a



(a) Scatterplot. (b) Residual plot.

Figure 2.29: *Entomology experiment. Larvae size versus head diameter size.*

mishap or a gross recording error, then this may be acceptable. On the other hand, if no such basis may be identified, the outlier may, in fact, be a genuine response; in this case, it contains information about the process under study, and may be reflecting a legitimate phenomenon. In this case, "throwing out" an outlier may lead to misleading conclusions, because a legitimate feature is being ignored.

TERMINOLOGY: In regression analysis, an observation is said to be **influential** if its removal from consideration causes a large change in the analysis (e.g., large change in the regression line, large change in r^2, etc.). An influential observation need not be an outlier! Similarly, an outlier need not be influential!

2.4.3 Correlation versus causation

REMARK: Investigators are often tempted to infer a **causal relationship** between X and Y when they fit a regression model or perform a correlation analysis. However, a significant association between X and Y does not necessarily imply a causal relationship!

Example 2.13. A Chicago newspaper reported that "there is a strong correlation between the numbers of fire fighters (X) at a fire and the amount of damage (Y, measured

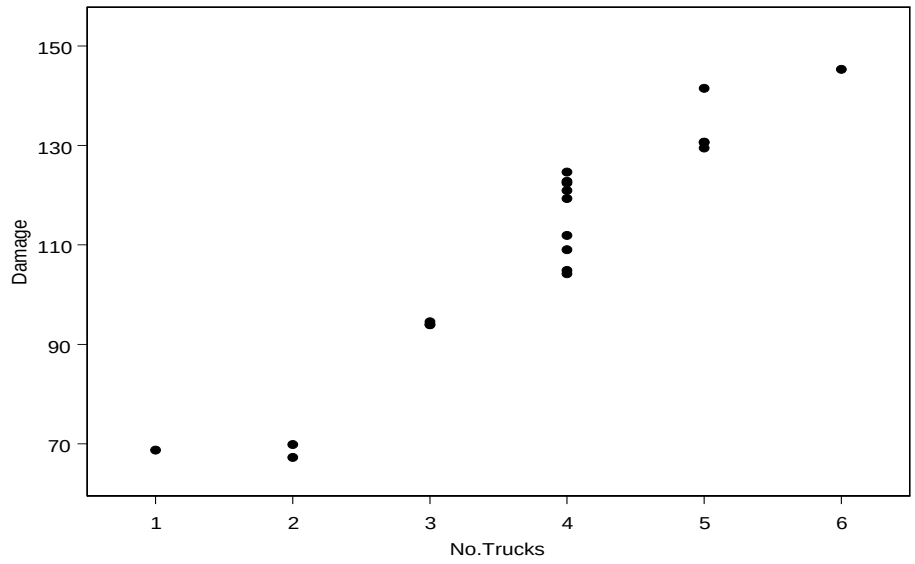


Figure 2.30: *Chicago fire damages* (\$1000s) *and the number of fire trucks.*

in \$1000's) that the fire does." Data from 20 recent fires in the Chicago area appear in Figure 2.30. From the plot, there appears to be a strong linear relationship between X and Y. Few, however, would infer that the increase in the number of fire trucks **causes** the observed increase in damages! Often, when two variables X and Y have a strong association, it is because both X and Y are, in fact, each associated with a third variable, say W. In the example, both X and Y are probably strongly linked to W, the severity of the fire, so it is understandable that X and Y would increase together.

MORAL: This phenomenon is the basis of the remark "*Correlation does not necessarily imply causation.*" An investigator should be aware of the temptation to infer causation in setting up a study, and be on the lookout for **lurking variables** like W above that are actually the driving force behind observed results. In general, the best way to control the effects of "lurking" variables is to use a carefully designed experiment. In observational studies, it is very difficult to make causal statements. Oftentimes, the best we can do is make statements documenting the observed association, and nothing more. For more information on causation, see Section 2.5 (MM).

3 Producing Data

3.1 Introduction

EXPLORATORY DATA ANALYSIS: Up until now, we have focused mainly on **exploratory data analysis**; that is, we have analyzed data to explore distributions and possible relationships between variables. In a nutshell, exploratory data analysis is performed to answer the question, *"What do we see in our sample of data?"* The use of graphical displays and summary statistics are an important part of this type of analysis.

STATISTICAL INFERENCE: A further analysis occurs when we wish to perform **statistical inference**. This has to do with generalizing the results of the sample to that of the population from which the sample data arose. Statistical inference is more numerical in nature and consists largely of confidence intervals of hypothesis tests. These important topics will be covered later in the course.

TERMINOLOGY: This chapter deals with producing data. **Primary data** are data that one collects proactively through observational studies and experiments. **Secondary data** are data gathered from other sources (e.g., journals, internet, census data, etc.).

TERMINOLOGY: An **observational study** observes individuals and measures variables of interest but does not attempt to influence the responses. An **experiment** deliberately imposes a treatment on individuals in order to observe their responses.

Example 3.1. The Human Relations Department at a major corporation wants to collect employee opinions about a prospective pension plan adjustment. Survey forms are sent out via email asking employees to answer 10 multiple-choice questions.

Example 3.2. In a clinical trial, physicians on a Drug and Safety Monitoring Board want to determine which of two drugs is more effective for treating HIV in its early stages. Patients in the trial are randomly assigned to one of two treatment groups. After 6 weeks on treatment, the net CD4 count change is observed for each patient.

DISCUSSION: Example 3.1 describes an observational study; there is no attempt to influence the responses of employees. Instead, information in the form of data from the questions is merely observed for each employee returning the survey form. Example 3.2 describes an experiment because we are trying to influence the response from patients (CD4 count) by giving different drugs.

TOWARD STATISTICAL INFERENCE: In Example 3.1, it might be the hope that information gathered from those employees returning the survey form are, in fact, representative of the entire company. When would this not be true? What might be ineffective with this type of **survey design**? In Example 3.2, we probably would like to compare the two drugs being given to patients in terms of the CD4 count change. However, would these results necessarily mean that similar behavior would occur in the entire population of advanced-HIV patients? This is precisely the question that we would like to answer, and using appropriate statistical inference procedures will help us answer it.

3.2 Experiments

3.2.1 Terminology and examples

TERMINOLOGY: In the language of experiments, individuals are called **experimental units**. An experimental condition applied to experimental units is called a **treatment**.

Example 3.3. Does aspirin reduce heart attacks? The most evidence for this claim comes from the Physicians Health Study, a large **double-blinded** experiment involving 22,000 male physicians. One group of about 11,000 took an aspirin every second day, while the rest of them took a **placebo** (i.e., a sugar pill designed to look like an aspirin).

- **Treatment**: Drug (e.g., placebo/aspirin)
- **Experimental units**: Physicians (they receive the treatment)
- **Response**: Heart attack/not.

TERMINOLOGY: In Example 3.3, we might call the group that received the placebo the **control group**. This group enables us to control the effects of outside variables on the outcome, and it gives a frame of reference for comparisons.

TERMINOLOGY: In the language of experiments, explanatory variables are often called **factors**. These are often categorical in nature. Factors are made up of different **levels**.

Example 3.4. A soil scientist wants to study the relationship between grass quality (for golf greens) and two factors:

- **Factor 1**: Fertilizer (ammonium sulphate, urea, isobutylidene dairy, sulfur urea)
- **Factor 2**: Number of months of dead-grass buildup (three, five, seven)
- **Response**: Grass quality, measured by the amount of chlorophyll content in the grass clippings (mg/gm).

Here, there are 12 **treatment combinations** ($4 \times 3 = 12$) obtained from forming all possible combinations of the levels of each factor; i.e.,

AS3	U3	IBID3	SU3
AS5	U5	IBID5	SU5
AS7	U7	IBID7	SU7

A golf green is divided up into 12 plots, roughly of equal size. *How should the 12 treatment combinations be applied to the plots?* This example describes an experiment with a **factorial treatment structure**, since each treatment combination is represented. If we had two greens, we could **replicate** the experiment (in this case, we would have $4 \times 3 \times 2 = 24$ observations).

Example 3.5. The response time (measured in milliseconds) was observed for three different types of circuits used in automatic value shutoff mechanisms. Twelve machines were used in the experiment and were randomly assigned to the three circuit types. Four replicates were observed for each circuit type. The data observed in the experiment are given in Table 3.7.

Table 3.7: *Circuit response data.*

Circuit type	Times	Means
1	9, 20, 12, 15	$\overline{x}_1 = 13.50$
2	23, 26, 20, 21	$\overline{x}_2 = 22.50$
3	6, 3, 10, 6	$\overline{x}_3 = 6.25$

- **Treatment**: Circuit type with three levels (1, 2, 3)
- **Response**: Time
- **Experimental unit**: Machine.

REMARK: When experimental units are randomly assigned to the treatments without restriction, we call such a design a **completely randomized design** (**CRD**). The experiment described in Example 3.5 is a CRD.

DISCUSSION: We see that the treatment means $\overline{x}_1, \overline{x}_2, \overline{x}_3$ are different. This is not surprising because each machine is inherently different. The big question is this: *are the observed differences real or could they have resulted just by chance?* Results that are real and are more likely to have not been caused by mere chance are said to be **statistically significant**. A major part of this course (coming up!) will be learning how to determine if results are statistically significant in an experimental setting.

EXPERIMENTS: An **experiment** is an investigation set up to answer research questions of interest.

- In our context, an experiment is most likely to involve a comparison of **treatments** (e.g., circuits, fertilizers, rations, drugs, methods, varieties, etc.).
- The outcome of an experiment is information in the form of observations on a **response variable** (e.g., yield, number of insects, weight gain, length of life, etc.).
- Because uncertainty in the responses due to sampling and biological variation, we

can *never* provide definitive answers to the research question(s) of interest based on such observations. *However, we can make inferences that incorporate and quantify inherent uncertainty.*

3.2.2 Designing experiments

IMPORTANT NOTES: Before an experiment may be designed, the question(s) of interest must be well-formulated. Nothing should start until this happens! The investigator and statistician should work together to identify important features and the appropriate design. This is extremely important as the design of the experiment lends naturally to the analysis of the data collected from the experiment. If the design changes, the analysis will as well. An experiment will most likely give **biased** results if it is not designed properly or is analyzed incorrectly. Before we talk about the fine points, consider this general outline of how experiments are performed:

General Outline of Experimentation

Sample experimental units from population

$\downarrow$

Randomize experimental units to treatments

$\downarrow$

Record data on each experimental unit

$\downarrow$

Analyze variation in data

$\downarrow$

Make statement about the differences among treatments.

PREVAILING THEME: The prevailing theme in **experimental design** is to allow explicitly for variation in and among samples, but design the experiment to **control** this variation as much as possible.

- This is certainly possible in experiments such as field trials in agriculture, clinical trials in medicine, reliability studies in engineering, etc.
- This is not usually possible in observational studies since we have no control over the individuals.

THE THREE BASIC PRINCIPLES: The basic principles of experimental design are **randomization**, **replication**, and **control**. We now investigate each principle in detail.

PRINCIPLE 1: RANDOMIZATION: This device is used to ensure that samples are as "alike as possible" except for the treatments. Instead of assigning treatments systematically, we assign them so that, once all the acknowledged sources of variation are accounted for, it can be assumed that no obscuring or **confounding** effects remain.

Example 3.6. An entomologist wants to determine if two preparations of a virus would produce different effects on tobacco plants. Consider the following experimental design:

- He took 10 leaves from each of 4 plots of land (so that there are 40 leaves in the experiment).
- For each plot, he randomly assigned 5 leaves to Preparation 1 and 5 leaves to Preparation 2.

Randomization in this experiment is **restricted**; that is, leaves were randomly assigned to treatments *within* each plot. Notice how the researcher has acknowledged the possible source of variation in the different plots of land. In this experiment, the researcher wants to compare the two preparations. *With this in mind, why would it be advantageous to include leaves from different plots?* This experiment is an example of a **randomized complete block design** (RCBD) where

- **Block**: Plot of land
- **Treatment**: Preparations (2)
- **Experimental unit**: Tobacco leaf.

AN INFERIOR DESIGN: In Example 3.6, suppose that our researcher used a CRD and simply randomized 20 leaves to each preparation. In this situation, he would have lost the ability to account for the possible variation among plots! The effects of the different preparations would be confounded with the differences in plots.

TERMINOLOGY: In the language of experiments, a **block** is a collection of experimental units that are thought to be "more alike" in some way.

Example 3.7. Continuing with Example 3.2, suppose it is suspected that males react to the HIV drugs differently than females. In light of this potential source of variation, consider the following three designs:

- **Design 1**: assign all the males to Drug 1, and assign all the females to Drug 2.
- **Design 2**: ignoring gender, randomize each individual to one of the two drugs.
- **Design 3**: randomly assign drugs within gender; that is, randomly assign the two drugs within the male group, and do the same for the female group.

Design 1 would be awful since if we observed "a difference," we would have no way of knowing whether or not it was from the treatments (i.e., drugs) or due to the differences in genders. Here, one might say that "drugs are **completely confounded** with gender." Design 2 is better than Design 1, but we still might not observe the differences due to drugs since differences in genders might not "average out" between the samples. However, in Design 3, differences in treatments may be assessed despite the differences in response due to the different genders! Design 3 is an example of an RCBD with

- **Block**: Gender
- **Treatment**: Drugs (2)
- **Experimental unit**: Human subject.

QUESTION: Randomization of experimental units to treatments is an important aspect of experimental design. *But, how does one physically randomize?*

TABLE OF RANDOM DIGITS: To avoid bias, treatments are assigned using a suitable randomization mechanism. Back in the days before we had computers, researchers could use a device known as a **Table of Random Digits**. See Table B (MM). For example, to assign each of eight fan blades to two experimental treatments (e.g., lubricant A and lubricant B) in a CRD, we list the blades as follows by their serial number and code:

```
xc4553d  0  xc4550d  4
xc4552e  1  xc4551e  5
xc4567d  2  xc4530e  6
xc4521d  3  xc4539e  7
```

To choose which four are allocated to lubricant A, we can generate a list of random numbers; then, match the randomly selected numbers to the list above. The first 4 matches go to lubricant A. Suppose the list was `26292 31009`. Reading from left to right in the sequence we see that numbers `2`, `6`, `3`, and `1` would be assigned to lubricant A. The others would be allocated to lubricant B.

REMARK: Tables of Random Digits are useful but are largely archaic. In practice, it is easier to use **random-number generators** from statistical software packages.

Example 3.8: *Matched-pairs design.* A certain stimulus is thought to produce an increase in mean systolic blood pressure (SBP) in middle-aged men.

DESIGN ONE: Take a random sample of men, then randomly assign each man to receive the stimulus or not using a CRD. Here, the two groups can be thought of as **independent** since one group receives one stimulus and the other group receives the other.

DESIGN TWO: Consider an alternative design, the so-called **matched-pairs design**. Rather than assigning men to receive one treatment or the other (stimulus/no stimulus), obtain a response from each man under both treatments! That is, obtain a random sample of middle-aged men and take two readings on each man, with and without the stimulus. In this design, because readings of each type are taken on the same man, the difference between before and after readings on a given man should be less variable than

Table 3.8: *Sources of variation present in different designs.*

Design	Type of Difference	Sources of Variation
Independent samples (CRD)	among men	among men, within men
Matched pairs setup	within men	within men

the difference between a before-response on one man and an after-response on a different man. The man-to-man variation inherent in the latter difference is not present in the difference between readings taken on the same subject!

ADVANTAGE OF MATCHED PAIRS: In general, by obtaining a pair of measurements on a single individual (e.g., man, rat, pig, plot, tobacco leaf, etc.), where one of the measurements corresponds to treatment 1 and the other measurement corresponds to treatment 2, you eliminate the variation *among* the individuals. Thus, you may compare the treatments (e.g., stimulus/no stimulus, ration A/ration B, etc.) under more **homogeneous** conditions where only variation within individuals is present (that is, the variation arising from the difference in treatments).

REMARK: In some situations, of course, pairing might be impossible or impractical (e.g., destructive testing in manufacturing, etc.). However, in a matched-pairs experiment, we still may think of two populations; e.g., those of all men with and without the stimulus. What changes in this setting is really how we have "sampled" from these populations. The two samples are no longer independent, because they involve the same individual.

A NOTE ON RANDOMIZATION: In matched-pairs experiments, it is common practice, when possible, to **randomize** the order in which treatments are assigned. This may eliminate "common patterns" (that may confound our ability to determine a treatment effect) from always following, say, treatment A with treatment B. In practice, the experimenter could flip a fair coin to determine which treatment is applied first. If there are **carry-over effects** that may be present, these would have to be dealt with accordingly. We'll assume that there are no carry-over effects in our discussion here.

PRINCIPLE 2: CONTROL: The basic idea of **control** is to eliminate confounding effects of other variables. Two variables are said to be **confounded** when their effects cannot be distinguished from each other. Consider the following two examples.

Example 3.9. In an agricultural study, researchers want to know which of three fertilizer compounds produces the highest yield. Suppose that we keep Fertilizer 1 plants in one greenhouse, Fertilizer 2 plants in another greenhouse, and Fertilizer 3 plants in yet another greenhouse. This choice may be made for convenience or simplicity, but this has the potential to introduce a big problem; namely, we may never know whether any observed differences are due to the actual differences in the treatments. Here, the observed differences could be due to the different greenhouses! The effects of fertilizer and greenhouse location are confounded.

Example 3.10. Consider a cancer clinical trial (a medical experiment where a drug's efficacy is of primary interest) to compare a new, experimental treatment to a standard treatment. Suppose that a doctor assigns patients with advanced cases of disease to a new experimental drug and assigns patients with mild cases to the standard drug, thinking that the new drug is promising and should thus be given to the sicker patients. Here, the effects of the drugs are confounded with the seriousness of the disease.

SOME GENERAL COMMENTS:

- One way to control possible confounding effects is to use **blocking**. Also, the use of blocking allows the experimenter to make treatment comparisons under more homogeneous conditions. How could blocking be used in the last two examples?
- Assignment of treatments to the samples should be done so that potential sources of variation do not obscure the treatment differences.
- If an experiment is performed, but the researchers fail to design the experiment to "block out" possible confounding effects, all results could be meaningless. *Thus, it pays to spend a little time at the beginning of the investigation and think hard about designing the experiment appropriately.*

PRINCIPLE 3: REPLICATION: Since we know that individuals will **vary** within samples (and among different samples from a population), we should collect data on more than one individual (e.g., pig, plant, person, plot, etc.). Doing so will provide us with more information about the population of interest and will reduce chance variation in the results.

Example 3.11. A common way to test whether a particular hearing aid is appropriate for a patient is to play a tape on which words are pronounced clearly but at low volume, and ask the patient to repeat the words as heard. However, a major problem for those wearing hearing aids is that the aids amplify background noise as well as the desired sounds. In an experiment, 24 subjects with normal hearing listened to standard audiology tapes of English words at low volume, with a noisy background (there were 25 words per tape). Recruited subjects were to repeat the words and were scored correct or incorrect in their perception of the words. The order of list presentation was randomized. Replication was achieved by using 24 subjects.

Example 3.12. A chemical engineer is designing the production process for a new product. The chemical reaction that produces the product may have higher or lower yield depending on the temperature and stirring rate in the vessel in which the reaction takes place. The engineer decides to investigate the effects of temperature (50C and 60C) and stirring rate (60rpm, 90rpm, and 120rpm) on the yield of the process. Here, we have 6 treatment combinations. One full replication would require 6 observations, one for each of the six treatment combinations. This is an example of an experiment with a **factorial treatment structure**.

REMARK: There are statistical benefits conferred by replicating the experiment under identical conditions (we'll become more familiar with these benefits later on); namely,

- increased **precision** and
- higher **power** to detect potential significant differences.

3.3 Sampling designs and surveys

TERMINOLOGY: A **survey** is an observational study where individuals are asked to respond to questions. In theory, no attempt is made to influence the responses of those participating in the survey. The **design** of the survey refers to the method used to choose the sample from the population. The survey is given to the sample of individuals chosen.

3.3.1 Sampling models

TERMINOLOGY: **Nonprobability samples** are samples of individuals that are chosen without using randomization methods. Such samples are rarely representative of the population from which the individuals were drawn (i.e., such samples often give **biased** results).

- A **voluntary response sample** (VRS) consists of people who choose themselves by responding to a general appeal (e.g., online polls, etc.). A famous VRS was collected by Literary Digest in 1936. Literary Digest predicted Republican presidential candidate Alfred Landon would defeat the incumbent president FDR by a 3:2 margin. In that election, FDR won 62 percent of the popular vote! The sampling procedure included two major mistakes. First, most of those contacted were from Literary Digest's subscription list! Second, only 2.3 percent of the ballots were returned (a voluntary response sample with major **non-response**).
- A **convenience sample** chooses individuals that are easiest to contact. For example, students standing by the door at the Student Union taking surveys are excluding those students who do not frequent the Union.

TERMINOLOGY: A **probability sample** is a sample where the individuals are chosen using randomization. While there are many sampling designs that can be classified as probability samples, we will discuss four: simple random samples (SRS), stratified random samples, systematic samples, and cluster samples.

SIMPLE RANDOM SAMPLE: A **simple random sample (SRS)** is a sampling model where individuals are chosen without replacement, and each sample of size n has an equal chance of being selected. See Example 3.18 (MM) on page 220.

- In the SRS model, we are choosing individuals so that our sample will hopefully be representative. Each individual in the population has an equal chance of being selected.
- We use the terms "random sample" and "simple random sample" interchangeably.

DIFFICULTIES: In practice, researchers most likely will not (i) identify exactly what the population is, (ii) identify all of the members of the population (an impossible task, most likely), and (iii) choose an SRS from this list. It is more likely that reasonable attempts will have been made to choose the individuals at random from those available. Hence, in theory, the sample obtained for analysis might not be "a true random sample;" however, in reality, the SRS model assumption might not be that far off.

SYSTEMATIC SAMPLE: A **systematic sample** is chosen by listing individuals in a **frame** (i.e., a complete list of individuals in a population) and selecting every jth individual on the list for the sample. In a systematic sample, like an SRS, each individual has the same chance of being selected. However, unlike the SRS, each sample of size n does not have the same chance of being selected (this is what distinguishes the two sampling models).

Example 3.13. In a plant-disease experiment, researchers select every 10th plant in a particular row for a sample of plants. The plants will then be tested for the presence of a certain virus.

STRATIFIED RANDOM SAMPLE: To select a **stratified random sample**, first divide the population into groups of similar individuals, called **strata**. Then choose a separate SRS in each stratum and combine these to form the full sample. Examples of strata include gender, cultivar, age, breed, salary, etc.

Table 3.9: *Education level for plasma donors in rural eastern China between* 1990-1994.

Education level	Count	Percentage
Illiterate	645	46.4
Primary	550	39.6
Secondary	195	14.0
Total	1390	100.0

REASONS FOR STRATIFICATION:

- Stratification broadens the scope of the study.
- Strata may be formed because they themselves may be of interest separately.

Example 3.14. In Example 1.3 (notes), we examined the following data from blood plasma donors in rural China. The three groups of individuals (illiterate, primary, and secondary) form strata for this population. Furthermore, it might be reasonable to assume that the samples within strata are each an SRS.

CLUSTER SAMPLE: **Cluster samples** proceed by first dividing the population into clusters; i.e., groups of individuals which may or may not be alike. Then, one randomly selects a sample of clusters, and uses every individual in those chosen clusters.

Example 3.15. For a health survey, we want a sample of $n = 400$ from a population of 10,000 dwellings in a city. To use an SRS would difficult because no frame is available and would be too costly and time-consuming to produce. Instead, we can use the city blocks as clusters. We then sample approximately 1/25 of the clusters.

NOTE: We have looked at four different probability sampling models. In practice, it is not uncommon to use one or more of these techniques in the same study (this is common in large government-based surveys). For example, one might use a cluster sample to first select city blocks. Then, one could take a stratified sample from those city blocks. In this situation, such a design might be called a **multistage sampling design**.

3.3.2 Common problems in sample surveys

UNDERCOVERAGE: This occurs when certain segments of the population are excluded from the study. For example, individuals living in certain sections of a city might be difficult to contact.

NONRESPONSE: This occurs when an individual chosen for the sample can not be contacted or refuses to participate. Non-response is common in mail and phone surveys.

RESPONSE BIAS: This could include **interviewer bias** and **respondent bias**. An interviewer might bias the results by asking the questions negatively. For example,

"In light of the mounting casualties that we have seen in Iraq recently, do you approve of the way President Bush has handled the war in Iraq?"

On the other hand, respondent bias might occur if the interviewee is not comfortable with the topic of the interview, such as sexual orientation or criminal behaviors. These types of biases can not be accounted for statistically. Also important is **how the questions are worded**. The following question appeared in an American-based survey in 1992:

"Does it seem possible or does it seem impossible to you that the Nazi extermination of the Jews never happened?"

When 22 percent of the sample said that it was "possible," the news media wondered how so many Americans could be uncertain. A much simpler version of this question was later asked and only 1 percent of the respondents said it was "possible."

3.4 Introduction to statistical inference

RECALL: **Statistical inference** deals with generalizing the results of a sample to that of the population from which the sample was drawn. This type of analysis is different from exploratory data analysis, which looks only at characteristics of the sample through graphical displays (e.g., histograms, scatterplots, etc.) and summary statistics.

Example 3.16. A Columbia-based health club wants to estimate the proportion of Columbia residents who enjoy running as a means of cardiovascular exercise. This proportion is unknown (that is why they want to estimate it!). Let p denote the proportion of Columbia residents who enjoy running as a means of exercise. If we knew p, then there would be no reason to estimate it. In lieu of this knowledge, we can take a sample from the population of Columbia residents and estimate p with the data from our sample. Here, we have two values:

$$p = \text{the true proportion of Columbia residents who enjoy running (unknown)}$$
$$\widehat{p} = \text{the proportion of residents who enjoy running observed in our sample.}$$

Suppose that in a **random sample** of $n = 100$ Columbia residents, 19 said that they enjoy running as a means of exercise. Then, $\widehat{p} = 19/100 = 0.19$.

TERMINOLOGY: A **parameter** is a number that describes a population. A parameter can be thought of as a fixed number, but, in practice, we do not know its value.

TERMINOLOGY: A **statistic** is a number that describes a sample. The value of a statistic can be computed from our sample data, but it can change from sample to sample! We often use a statistic to estimate an unknown parameter.

REMARK: In light of these definitions, in Example 3.16, we call

$$p = \text{population proportion}$$
$$\widehat{p} = \text{sample proportion,}$$

and we can use $\widehat{p} = 0.19$ as an **estimate** of the true p.

HYPOTHETICALLY: Suppose now that I take another SRS of Columbia residents of size $n = 100$ and 23 of them said that they enjoy running as a means of exercise. From this sample, our estimate of p is $\widehat{p} = 23/100 = 0.23$. That the two samples gave different estimates should not be surprising (the samples most likely included different people). *Statistics' values vary from sample to sample because of this very fact.* On the other hand, the value of p, the population proportion, does not change.

OF INTEREST: In light of the fact that statistics' values will change from sample to sample (this is called **sampling variability**), it is natural to want to know what would happen if we repeated the sampling procedure many times. In practice, we can not do this because, chances are, we only get to take one sample from our population. However, we can imitate this notion by using **simulation**.

SIMULATION: Here is what happened when I simulated 10 different values of $\widehat{p}$ under the assumption that $p = 0.2$. I used R to perform the simulations.

```
> p.hat
0.18 0.25 0.20 0.20 0.17 0.16 0.19 0.19 0.14 0.19
```

Now, I simulated a few more values of $\widehat{p}$ under the assumption that $p = 0.2$:

```
> p.hat
0.17 0.16 0.20 0.20 0.16 0.25 0.21 0.17 0.14 0.26 0.12 0.20 0.18
0.15 0.28 0.20 0.21 0.24 0.25 0.22 0.18 0.20 0.19 0.16 0.21 0.21
0.25 0.18 0.20 0.14 0.14 0.21 0.20 0.17 0.18 0.15 0.21 0.12 0.18
0.23 0.18 0.22 0.26 0.18 0.13 0.19 0.17 0.28 0.18 0.21 0.22 0.18
0.19 0.22 0.23 0.17 0.26 0.21 0.19 0.19 0.20 0.10 0.17 0.18 0.18
0.18 0.21 0.20 0.23 0.23 0.26 0.18 0.18 0.16 0.24 0.22 0.16 0.21
0.27 0.18 0.19 0.26 0.25 0.24 0.10 0.18 0.18 0.25 0.18 0.21 0.20
0.21 0.20 0.18 0.22 0.19 0.26 0.17 0.16 0.20
```

It should be clear that I can generate as many values of $\widehat{p}$ as I want. Thus, to answer the question, *"What would happen if I took many samples?"* we can do the following:

1. Take a large number of samples assuming that $p = 0.2$.
2. Calculate the sample proportion $\widehat{p}$ for each sample.
3. Make a histogram of the values of $\widehat{p}$.
4. Examine the distribution displayed in the histogram.

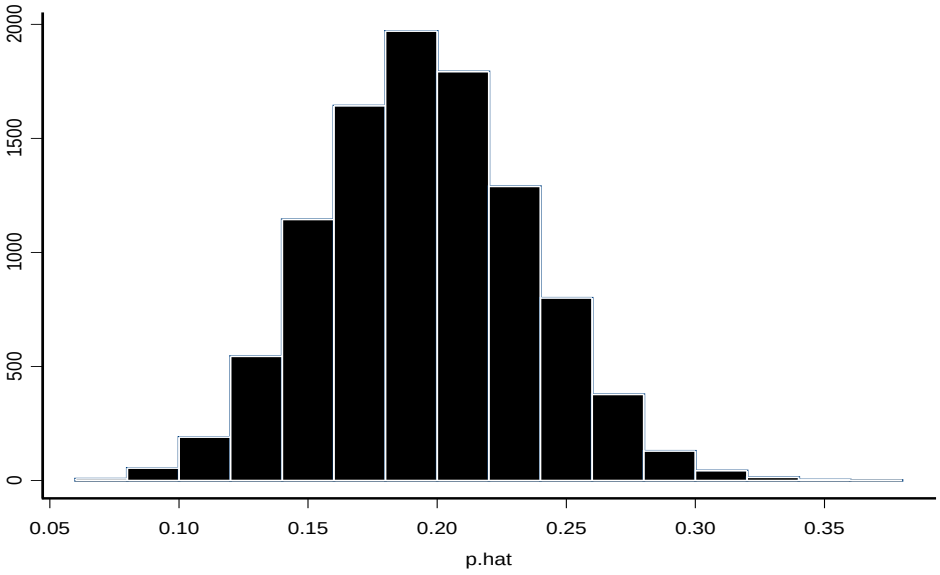


Figure 3.31: *Sampling distribution of the sample proportion when* $p = 0.2$.

RESULTS: When I did this using 10,000 random samples, each of size $n = 100$, I got the histogram in Figure 3.31. It is interesting to note that the shape looks **normally distributed** with a mean of around 0.2. In fact, when I took the sample mean of the 10,000 randomly-generated $\widehat{p}$ values, I got

```
> mean(p.hat)
0.200419
```

This is close to the true proportion $p = 0.2$. What we have just done is generated (using computer simulation) the sampling distribution of $\widehat{p}$ when $p = 0.2$.

SAMPLING DISTRIBUTIONS: The **sampling distribution** of a statistic is the distribution of values taken by the statistic in repeated sampling using samples of the same size.

REMARK: Sampling distributions are important distributions in statistical inference. As with any distribution, we will be interested in the following:

- **center** of the distribution

- **spread** (variation) in the distribution
- **shape**: is the distribution symmetric or skewed?
- the presence of **outliers**.

TWO QUESTIONS FOR THOUGHT:

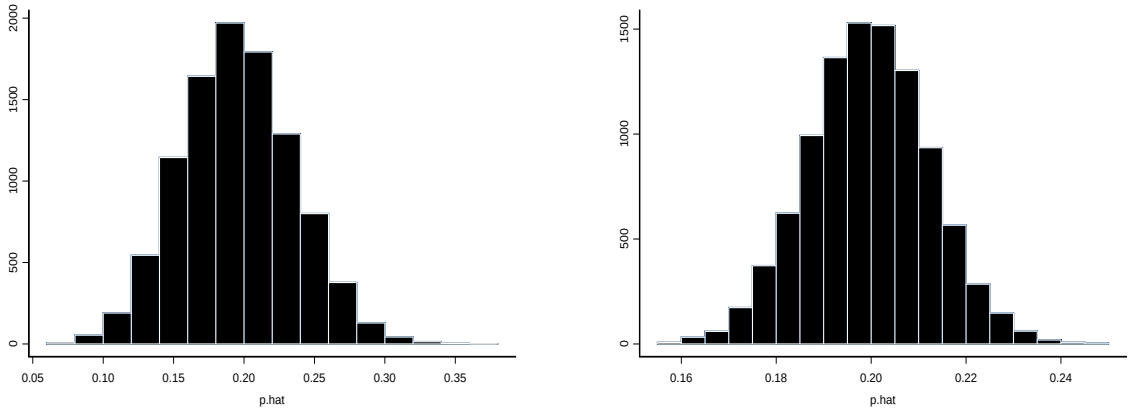
- First, based on the appearance of the sampling distribution for $\widehat{p}$ when $p = 0.2$ (see Figure 3.31), what range of values of $\widehat{p}$ are common?
- Second, suppose that in your investigation, you did not get a value of $\widehat{p}$ in this commonly-observed range. What does this suggest about the value of p?

TERMINOLOGY: A statistic is said to be an **unbiased estimator** for a parameter if the mean of the statistic's sampling distribution is equal to the parameter. If the mean of the statistic's sampling distribution is not equal to this parameter, the statistic is called a **biased estimator**. It should be clear that bias (or lack thereof) is a property that concerns the center of a sampling distribution.

TERMINOLOGY: The **variability of a statistic** is described by the spread in the statistic's sampling distribution. This variability often depends largely on the sample size n. When n is larger, and the sample is a probability sample (e.g., SRS), the variability often is smaller.

IDEAL SITUATION: Better statistics are those which have small (or no) bias and small variability. If a statistic is unbiased (or has very low bias) then we know that we are approximately "right on average." If we have a statistic with small variability, then we know there is not a large spread in that statistic's sampling distribution. The combination of "right on average" and "small variation" is the ideal case! See p. 237 (MM) for an illustration.

MANAGING BIAS AND VARIABILITY: To reduce bias, use random sampling. In many situations, SRS designs produce unbiased estimates of population parameters. To



(a) $n = 100$ (b) $n = 1000$

Figure 3.32: *Sampling distribution of the sample proportion when* $p = 0.2$.

reduce variation, use a larger sample size n. Often, the variation in the sampling distribution decreases as n increases. We'll see examples of this later on.

ILLUSTRATION: To illustrate how increasing the sample size decreases the variation in a statistic's sampling distribution, consider Figure 3.32. On the left, we see the sampling distribution of $\widehat{p}$ when $p = 0.2$ and $n = 100$. On the right, we see the the sampling distribution of $\widehat{p}$ when $p = 0.2$ and $n = 1000$. I simulated both distributions using R. Note how the variation in the right distribution is greatly reduced. This happens because the sample size is larger. Note also that increasing the sample size didn't affect the mean of the sampling distribution of $\widehat{p}$; i.e., $\widehat{p}$ is unbiased!

REMARK: Randomization is an important aspect of sampling design in experiments and observational studies. Whenever you sample individuals, there are two types of error: sampling error and non-sampling error. **Sampling error** is induced from natural variability among individuals in our population. The plain fact is that we usually can never get to see every individual in the population. Sampling error arises from this fact. **Non-sampling error** occurs from other sources such as measurement error, poor sampling designs, nonresponse, undercoverage, missing data, etc. Randomization helps to reduce the non-sampling error; however, it usually can not fully eliminate it.

4 Probability: The Study of Randomness

4.1 Randomness

TERMINOLOGY: We call a phenomenon **random** if its individual outcomes are uncertain, but there is nonetheless a regular distribution of outcomes in a large number of repetitions. The **probability** of any outcome is the proportion of times the outcome would occur in a very long series of independent repetitions.

NOTE: This interpretation of probability just described (as a long-run proportion) is called the **relative frequency approach** to measuring probability.

EXAMPLES: Here are examples of outcomes we may wish to assign probabilities to:

- tomorrow's temperature exceeding 80 degrees
- manufacturing a defective part
- the NASDAQ losing 5 percent of its value
- rolling a "2" on an unbiased die.

Example 4.1. *An example illustrating the relative frequency approach to probability.* Suppose we roll a die $n = 1000$ times and record the number of times we observe a "2." Let A denote this outcome. The quantity

$$\frac{\text{number of times } A \text{ occurs}}{\text{number of trials performed}} = \frac{f}{n}$$

is called the **relative frequency** of the outcome. If we performed this experiment repeatedly, the relative frequency approach says that $P(A) \approx f/n$, for large n.

NOTATION: The symbol $P(A)$ is shorthand for "the probability that A occurs."

SIMULATION: I simulated the experiment in Example 4.1 four times; that is, I rolled a single die 1,000 times on four separate occasions (using a computer). Each occasion

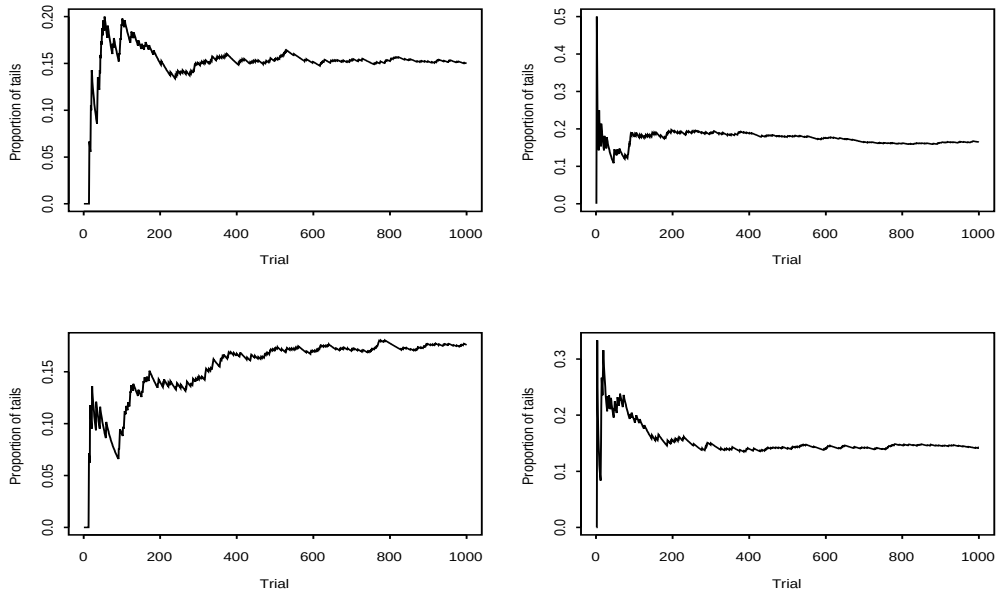


Figure 4.33: *The proportion of tosses which result in a* "2"*; each plot represents* $1,000$ *rolls of a fair die.*

is depicted in its own graph in Figure 4.33. The vertical axis measures the relative frequency (proportion of occurrences) of the outcome $A = \{\text{roll a } 2\}$. You will note that each relative frequency gets close to $1/6$ as n, the number of rolls, increases. The relative frequency approach says that this relative frequency will "converge" to $1/6$, if we were allowed to continually roll the die.

4.2 Probability models

TERMINOLOGY: In probability applications, it is common to perform an experiment and then observe an **outcome**. The set of all possible outcomes for an experiment is called the **sample space**, hereafter denoted by S.

Example 4.2. A rat is selected and we observe the sex of the rat:

$$S = \{\text{male, female}\}.$$

Example 4.3. The Michigan state lottery calls for a three-digit integer to be selected:

$$S = \{000, 001, 002, ..., 998, 999\}.$$

Example 4.4. An industrial experiment consists of observing the lifetime (measured in hours) of a certain battery:

$$S = \{w : w \geq 0\}.$$

That is, any positive value (in theory) could be observed.

TERMINOLOGY: An **event** is an outcome or set of outcomes in a random experiment. Put another way, an event is a **subset** of the sample space. We typically denote events by capital letters towards the beginning of the alphabet (e.g., A, B, C, etc.).

EXAMPLES: In Example 4.2, let $A = \{\text{female}\}$. In Example 4.3, let $B = \{003, 547, 988\}$. In Example 4.4, let $C = \{w : w < 30\}$. What are $P(A)$, $P(B)$, and $P(C)$? To answer these questions, we have to learn more about how probabilities are assigned to outcomes in a sample space.

PROBABILITY RULES: Probabilities are assigned according to the following rules:

1. The probability $P(A)$ of any event A satisfies $0 \leq P(A) \leq 1$.

2. If S is the sample space in a probability model, then $P(S) = 1$.

3. Two events A and B are **disjoint** if they have no outcomes in common (i.e., they can not both occur simultaneously). If A and B are disjoint, then

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

4. The **complement** of any event A is the event that A does not occur. The complement is denoted by A^c. The **complement rule** says that

$$P(A^c) = 1 - P(A).$$

NOTE: **Venn Diagrams** are helpful in depicting events in a sample space! See those on p. 263 (MM).

Example 4.5. Consider the following probability model for the ages of students taking distance courses at USC:

Age group	18-23 years	24-29 years	30-39 years	40 years and over
Probability	0.47	0.27	0.14	0.12

First, note that each outcome has a probability that is between 0 and 1. Also, the probabilities sum to one because the four outcomes make up S.

1. What is the probability that a distance-course student is 30 years or older?
2. What is the probability that a distance-course student is less than 40 years old?
3. What is the probability that a distance-course student is 25 years old?

4.2.1 Assigning probabilities

Example 4.6. In an experiment, we observe the number of insects that test positive for a virus. If 6 insects are under study, the sample space is $S = \{0, 1, 2, ..., 6\}$. Define

$$A_1 = \{\text{exactly one insect is infected}\} = \{1\}$$
$$A_2 = \{\text{no insects are infected}\} = \{0\}$$
$$A_3 = \{\text{more than one insect is infected}\} = \{2, 3, 4, 5, 6\}$$
$$A_4 = \{\text{at most one insect is infected}\} = \{0, 1\}.$$

Note that events A_1 and A_2 each contain only one outcome. The events A_3 and A_4 each contain more than one outcome. *How do we assign probabilities to these events?* To do this, we need to know how probabilities are assigned to each outcome in S.
EXERCISE: Draw a Venn Diagram for the events A_1, A_2, A_3, and A_4.

TERMINOLOGY: A sample space is called **finite** if there are a fixed and limited number of outcomes. Examples 4.2, 4.3, 4.5, and 4.6 have finite sample spaces. Example 4.4 does not have a finite sample space.

ASSIGNING PROBABILITIES: In a finite sample space, computing the probability of an event can be done by

(1) identifying all outcomes associated with the event

(2) adding up the probabilities associated with each outcome.

Example 4.7. In Example 4.6, consider the following probability model for the number of insects which test positive for a virus:

Outcome	0	1	2	3	4	5	6
Probability	0.55	0.25	0.10	0.04	0.03	0.02	0.01

With this probability model, we can now compute $P(A_1)$, $P(A_2)$, $P(A_3)$, and $P(A_4)$. In particular, $P(A_1) = 0.25$, $P(A_2) = 0.55$,

$$P(A_3) = 0.10 + 0.04 + 0.03 + 0.02 + 0.01 = 0.20,$$

and

$$P(A_4) = 0.55 + 0.25 = 0.80.$$

QUESTION: What is $P(A_1 \text{ or } A_2)$? $P(A_3 \text{ or } A_4)$? $P(A_1 \text{ or } A_4)$?

EQUALLY LIKELY OUTCOMES: In discrete sample spaces with equally likely outcomes, assigning probabilities is very easy. In particular, if a random phenomenon has a sample space with k possible outcomes, all **equally likely**, then

- each individual outcome has probability $1/k$, and
- the probability of any event A is

$$P(A) = \frac{\text{number of outcomes in } A}{\text{number of outcomes in } S} = \frac{\text{number of outcomes in } A}{k}.$$

Example 4.8. *Equally-likely outcomes.* In Example 4.3, the sample space for the Michigan lottery experiment is

$$S = \{000, 001, 002, ..., 998, 999\}.$$

Suppose that I bought three lottery tickets, numbered 003, 547, and 988. What is the probability that my number is chosen? That is, what is $P(B)$, where $B = \{003, 547, 988\}$? If each of the $k = 1000$ outcomes in S is equally likely (this should be true if the lottery numbers are randomly selected), then each outcome has probability $1/k = 1/1000$. Also, note that there are 3 outcomes in B; thus,

$$P(B) = \frac{\text{number of outcomes in } B}{1000} = \frac{3}{1000}.$$

Example 4.9. *Equally-likely outcomes.* Four equally qualified applicants (a, b, c, d) are up for two positions. Applicant a is a minority. Positions are chosen at random. What is the probability that the minority is hired? Here, the sample space is

$$S = \{ab, ac, ad, bc, bd, cd\}.$$

We are assuming that the order of the positions is not important. If the positions are assigned at random, each of the six sample points is equally likely and has 1/6 probability. Let E denote the event that a minority is hired. Then, $E = \{ab, ac, ad\}$ and

$$P(E) = \frac{\text{number of outcomes in } E}{6} = \frac{3}{6}.$$

4.2.2 Independence and the multiplication rule

INFORMALLY: Some events in a sample space may be "related" in some way. For example, if A denotes the event that it rains tomorrow, and B denotes the event that it will be overcast tomorrow, then we know that the occurrence of A is related to the occurrence of B. If B does not occur, this changes our measure of $P(A)$ from what it would be if B did occur.

TERMINOLOGY: When the occurrence or non-occurrence of A has no effect on whether or not B occurs, and vice-versa, we say that the events A and B are **independent**.

MATHEMATICALLY: If the events A and B are independent, then

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

This is known as the **multiplication rule** for independent events. *Note that this rule only applies to events that are independent. If events A and B are not independent, this rule does not hold!*

Example 4.10. A red die and a white die are rolled. Let $A = \{4 \text{ on red die}\}$ and $B = \{\text{sum is odd}\}$. Are the events independent?

SOLUTION. The sample space here is

$$\begin{aligned} S = \{&(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\ &(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\ &(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}. \end{aligned}$$

In each outcome, the first value corresponds to the red die; the second value corresponds to the white die. Of the 36 outcomes in S (each of which are assumed equally likely),

- 6 are favorable to A,
- 18 are favorable to B, and
- 3 are favorable to both A and B.

Thus, since

$$\frac{3}{36} = P(A \text{ and } B) = P(A)P(B) = \frac{6}{36} \times \frac{18}{36},$$

the events A and B are independent.

Example 4.11. In an engineering system, two components are placed in a **series** so that the system is functional as long as **both** components are. Let A_1 and A_2, denote the events that components 1 and 2 are functional, respectively. From past experience we know that $P(A_1) = P(A_2) = 0.95$. Assuming *independence* between the two components, the probability the system is functional (i.e., the **reliability** of the system) is

$$P(A_1 \text{ and } A_2) = P(A_1)P(A_2) = (0.95)^2 = 0.9025.$$

If the two components are not independent, then we do not have enough information to determine the reliability of the system.

EXTENSION: The notion of independence still applies if we are talking about more than two events. If the events $A_1, A_2, ..., A_n$ are independent, then

$$P(\text{all } A_i \text{ occur}) = P(A_1)P(A_2)\cdots P(A_n).$$

FACT: If A and B are independent events, so are (a) A^c and B (b) A and B^c, and (c) A^c and B^c. *That is, complements of independent events are also independent.*

Example 4.12. Suppose that in a certain population, individuals have a certain disease with probability $p = 0.05$. Suppose that 10 individuals are observed. Assuming independence among individuals, what is the probability that

(a) no one has the disease?

(b) at least one has the disease?

4.3 Random variables

TERMINOLOGY: A **random variable** is a variable whose numerical value is determined by chance. We usually denote random variables by capital letters near the end of the alphabet (e.g., X, Y, Z, etc.).

NOTATION: We denote a **random variable** X with a capital letter; we denote an **observed value** by x, a lowercase letter. This is standard notation.

4.3.1 Discrete random variables

TERMINOLOGY: A **discrete random variable**, say, X has a finite (limited) number of possible values. The **probability distribution** of X lists these values and the probabilities associated with each value:

x	x_1	x_2	x_3	$\cdots$	x_k
Probability	p_1	p_2	p_3	$\cdots$	p_k

The probabilities p_i **must** satisfy:

1. Every probability p_i is a number between 0 and 1.
2. The probabilities $p_1, p_2, ..., p_k$ must **sum** to 1.

RULE: We find the probability of any event by adding the probabilities p_i of the particular values x_i that make up the event.

Example 4.14. Suppose that an experiment consists of flipping two fair coins. The sample space consists of four outcomes:

$$S = \{(H,H), (H,T), (T,H), (T,T)\}.$$

Now, let Y count the **number of heads** in the two flips. Before we perform the experiment, we do not know the value of Y (it is a random variable). Assuming that each outcome in S is **equally likely**, we can compute the probability distribution for Y:

y	0	1	2
Probability	0.25	0.50	0.25

What is the probability that I flip **at least one** head? This corresponds to the values $y = 1$ and $y = 2$; hence, the probability is

$$P(Y \geq 1) = P(Y = 1) + P(Y = 2) = 0.50 + 0.25 = 0.75.$$

Example 4.15. During my morning commute, there are 8 stoplights between my house and I-77. Consider the following probability distribution for X, the number of stop lights at which I stop (note that I stop only at red lights).

x	0	1	2	3	4	5	6	7	8
Probability	0.20	0.20	0.30	0.20	0.04	0.02	0.02	0.01	0.01

Using this model, what is the probability that I must stop (a) at most once? (b) at least five times? (c) not at all?

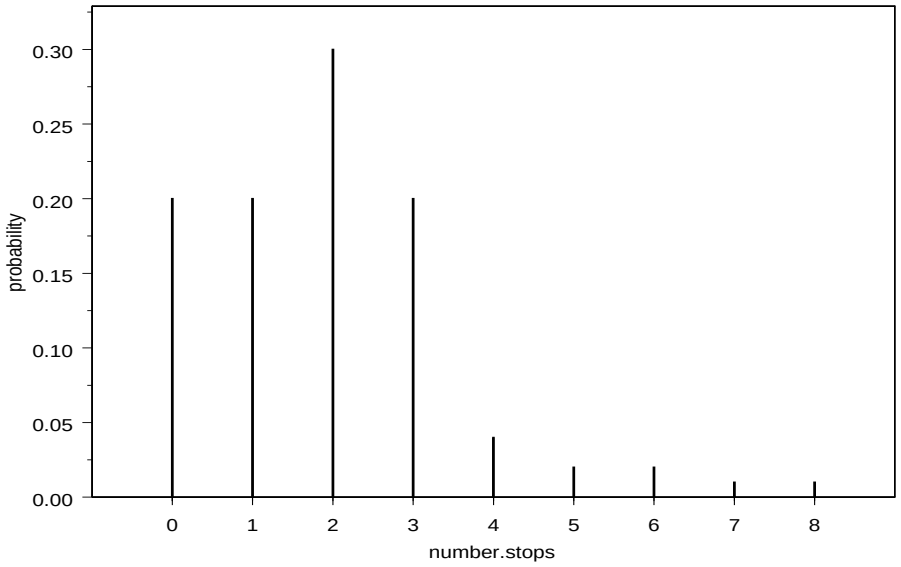


Figure 4.34: *Probability histogram for the number of traffic light stops.*

SOLUTION (a):

$$\begin{aligned} P(X \leq 1) &= P(X=0) + P(X=1) \\ &= 0.20 + 0.20 = 0.40. \end{aligned}$$

SOLUTION (b):

$$\begin{aligned} P(X \geq 5) &= P(X=5) + P(X=6) + P(X=7) + P(X=8) \\ &= 0.02 + 0.02 + 0.01 + 0.01 \\ &= 0.06. \end{aligned}$$

SOLUTION (c): $P(X=0) = 0.20$.

TERMINOLOGY: **Probability histograms** are graphical displays that show the probability distribution of a discrete random variable. The values for the variable are placed on the horizontal axis; the probabilities are placed on the vertical axis. A probability histogram for Example 4.15 is given in Figure 4.34.

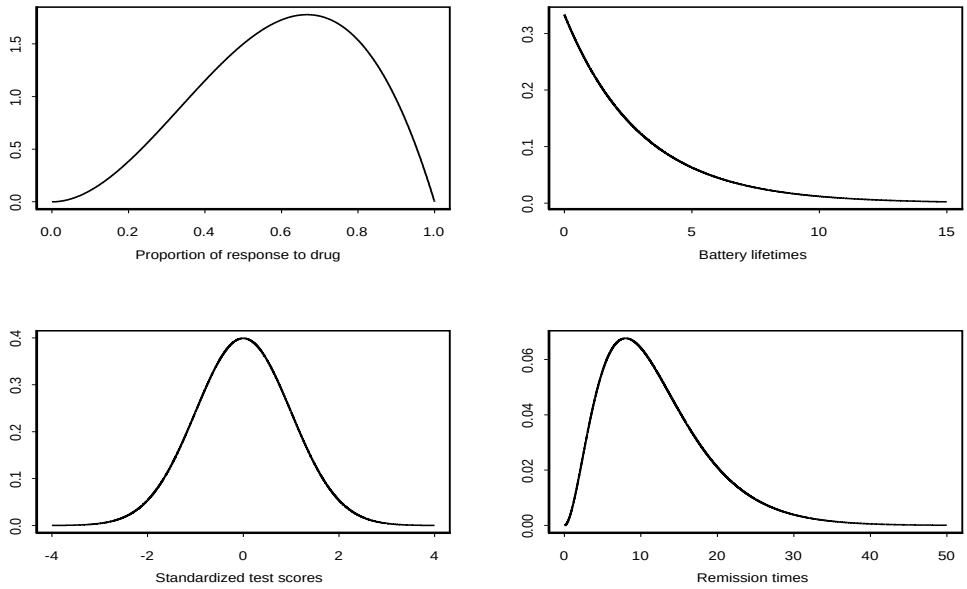


Figure 4.35: *Four density curves.*

4.3.2 Continuous random variables

TERMINOLOGY: **Continuous random variables** are random variables that take on values in intervals of numbers (instead of a fixed number of values like discrete random variables).

Example 4.16. Let Y denote the weight, in ounces, of the next newborn boy in Columbia, SC. Here, Y is continuous random variable because it can (in theory) assume any value larger than zero.

TERMINOLOGY: The probability distribution of a continuous random variable is represented by a **density curve**. Examples of density curves appear in Figure 4.35.

COMPUTING PROBABILITIES: *How do we compute probabilities associated with continuous random variables?* We do this by **finding areas** under density curves. See Figure 4.36. Associating the area under a density curve with probability is not new. In Chapter 1, we associated areas with the *proportion of observations* falling in that range.

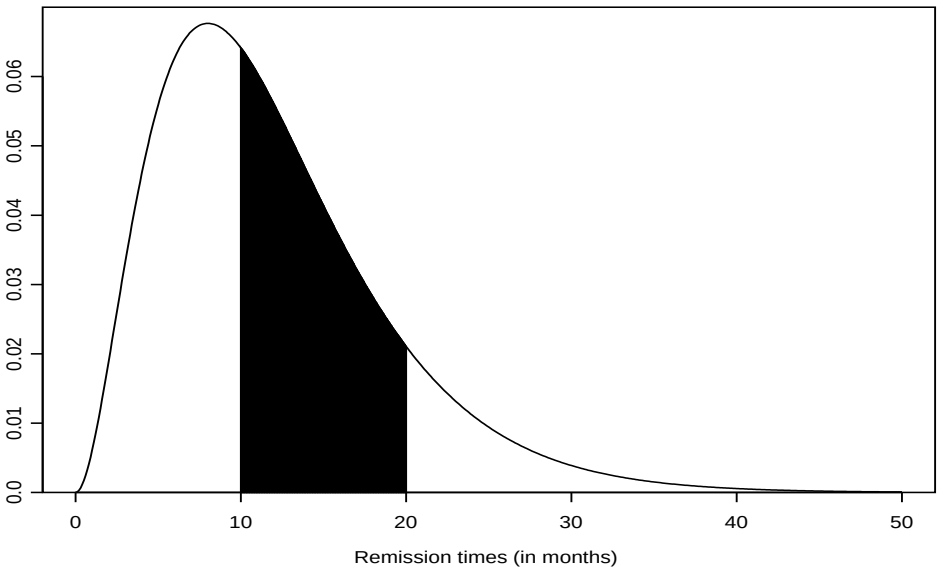


Figure 4.36: *Density curve for X, the remission times of leukemia patients. The shaded region represents the probability $P(10 < X < 20)$.*

These are similar ideas; that is, in Figure 4.36, the shaded area represents

- $P(10 < X < 20)$, the **probability** that a single individual (selected from the population) will have a remission time between 10 and 20 months.
- the **proportion** of individuals in the population having a remission time between 10 and 20 months.

UNUSUAL FACT: Continuous random variables are different than discrete random variables. Discrete random variables assign positive probabilities to specific values (see Examples 4.14 and 4.15). On the other hand, continuous random variables assign probability 0 to specific points! *Why?* It has to do with how we assign probabilities for continuous random variables. The area under a density curve, directly above a specific point, is zero! Thus, for the density curve in Figure 4.36, the probability that a remission time for a selected patient **equals** 22 months is $P(X = 22) = 0$.

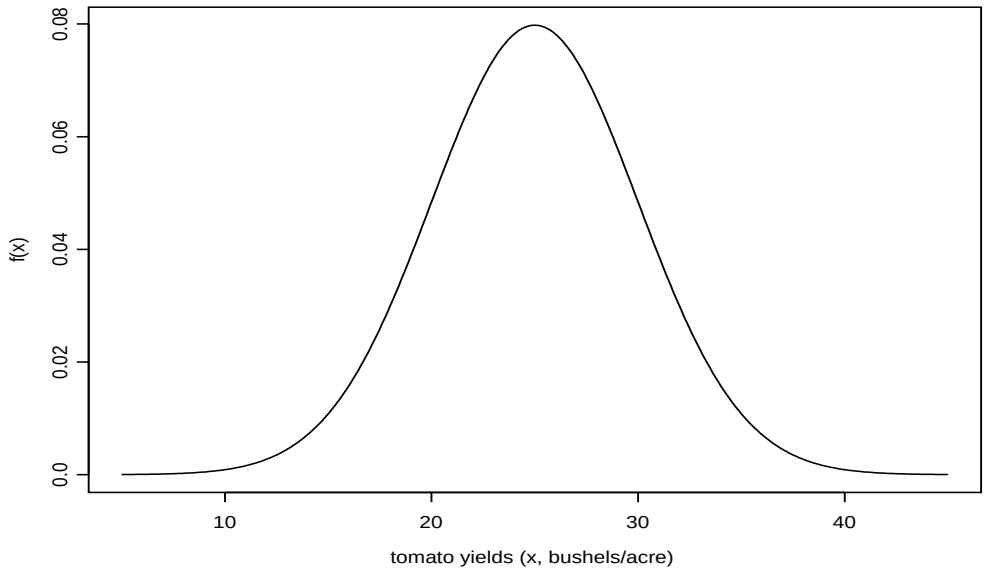


Figure 4.37: *A normal probability distribution with mean* $\mu = 25$ *and standard deviation* $\sigma = 5$. *A model for tomato yields.*

NOTE: The **normal distributions** studied in Chapter 1 (Section 1.3) are examples of probability distributions. Normal density curves were studied extensively in Chapter 1.

RECALL: Suppose that the random variable $X \sim \mathcal{N}(\mu, \sigma)$. Then, the random variable

$$Z = \frac{X - \mu}{\sigma}.$$

has a normal distribution with mean 0 and standard deviation 1 (i.e., a **standard normal distribution**). That is, $Z \sim \mathcal{N}(0, 1)$.

REVIEW QUESTIONS: Let X denote the tomato yields per acre in a certain geographical region. The probability distribution for X is normal and is depicted in Figure 4.37. Compute the following probabilities and draw appropriate shaded pictures:

- $P(X > 27.4)$
- $P(19.3 < X < 30.8)$
- $P(X < 16.2)$.

4.4 Means and variances of random variables

REMARK: Random variables have probability distributions that describe

1. the **values** associated with the random variable
2. the **probabilities** associated with these values.

Graphically, we represent probability distributions with **probability histograms** (in the discrete case) and **density curves** (in the continuous case).

FACT: Random variables have means and variances associated with them! For any random variable X,

- the **mean** of X is denoted by μ_X
- the **variance** of X is denoted by σ_X^2
- the **standard deviation** of X is denoted by σ_X (and is still the positive square root of the variance).

4.4.1 Means: Discrete case

MEAN OF A DISCRETE RANDOM VARIABLE: Suppose that X is a discrete random variable whose distribution is

x	x_1	x_2	x_3	$\cdots$	x_k
Probability	p_1	p_2	p_3	$\cdots$	p_k

To find the **mean** of X, simply multiply each possible value by its probability and add the results; i.e.,

$$\mu_X = x_1p_1 + x_2p_2 + x_3p_3 + \cdots + x_kp_k = \sum_{i=1}^{k} x_ip_i.$$

PHYSICAL INTERPRETATION: The mean μ_X of a discrete random variable X may be thought of as "the balance point" on a probability histogram for X (this is similar to how we interpreted $\overline{x}$ for a sample of data in Chapter 1).

Example 4.17. In Example 4.15, we saw the probability distribution for X, the number of stop lights at which I stop.

x	0	1	2	3	4	5	6	7	8
Probability	0.20	0.20	0.30	0.20	0.04	0.02	0.02	0.01	0.01

What is the mean number of stops I have to make on any given day?

SOLUTION. The mean of X is given by

$$\begin{aligned}
\mu_X &= x_1p_1 + x_2p_2 + x_3p_3 + \cdots + x_9p_9 \\
&= 0(0.20) + 1(0.20) + 2(0.30) + 3(0.20) + 4(0.04) + 5(0.02) \\
&\quad +6(0.02) + 7(0.01) + 8(0.01) \\
&= 1.93.
\end{aligned}$$

Thus, the mean number of stops I make is $\mu_X = 1.93$ stops/day.

INTERPRETATION: The mean number of stops is 1.93; this may cause confusion because 1.93 is not even a value that X can assume! Instead, we may interpret the mean as a **long-run average**. That is, suppose that I recorded the value of X on 1000 consecutive morning commutes to I-77 (that's over 3 years of commutes!). If I computed $\overline{x}$, the average of these 1000 observations, it would be approximately equal to $\mu_X = 1.93$.

LAW OF LARGE NUMBERS: Draw independent observations at random (e.g., SRS) from any population (distribution) with mean μ. The **Law of Large Numbers** says that as n, the number of observations drawn, increases, the sample mean $\overline{x}$ of the observed values approaches the population mean μ and stays arbitrarily close to it. In other words, $\overline{x}$ "converges" to μ.

REMARK: The Law of Large Numbers (LLN) applies to any probability distribution (not just normal distributions). The LLN is illustrated in Figure 4.38.

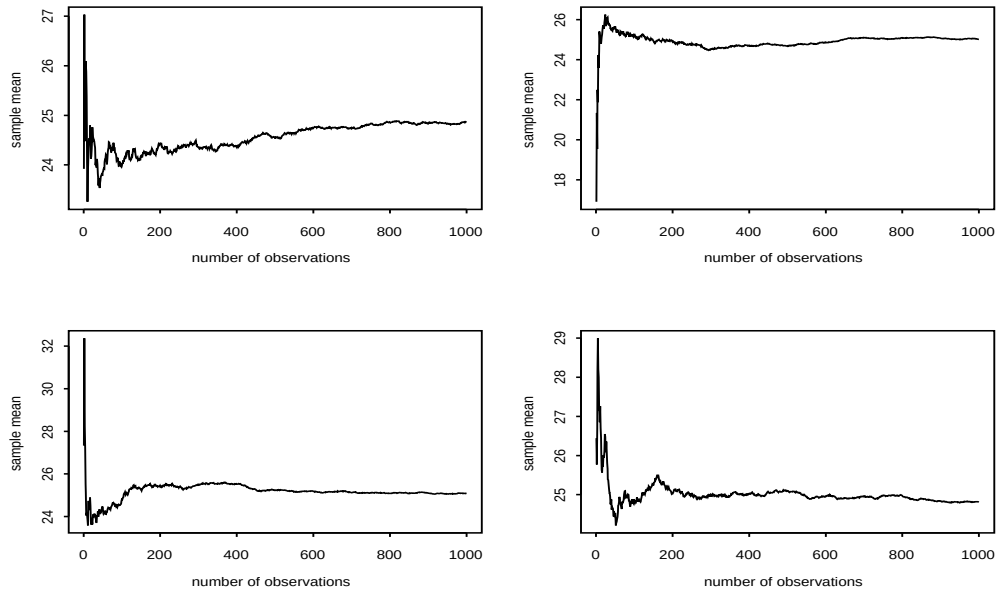


Figure 4.38: *An illustration of the Law of Large Numbers. Each graph depicts the long-run sample mean* $\overline{x}$ *computed from* $1,000$ *observations from a* $\mathcal{N}(25, 5)$ *distribution.*

SCENARIO: Fred and Barney are playing roulette in Las Vegas. For those of you that are not familiar with roulette, don't worry; the important thing to know is that the probability of winning on a given color (red or black) is $p = 18/38$. There are 18 red slots, 18 black slots, and 2 green slots. Fred is playing roulette, always bets on red, and he has lost 10 consecutive times. Barney says "play one more time; you are due for a win!" What is wrong with this reasoning?

Answer. The important statistical issue that Barney is ignoring is that the spins of the roulette wheel are likely **independent**. That is, what comes up in one spin has nothing to do with what comes up in other spins. That Fred has lost the previous 10 spins is irrelevant insofar as his next play. Also, Barney probably has never heard of the Law of Large Numbers. The value $p = 18/38$ is really a **mean**. To see this, let X denote the outcome of Fred's bet on any given spin (i.e., $x = 0$ if Fred loses and $x = 1$ if Fred wins). Then, $x = 1$ occurs with probability $p = 18/38$ and $x = 0$ occurs with probability $1 - p = 20/38$. That is, X obeys the following probability distribution:

x	0 (loss)	1 (win)
Probability	20/38	18/38

Here, the mean of X is

$$\begin{aligned}\mu_X &= x_1p_1 + x_2p_2 \\ &= 0(20/38) + 1(18/38) \\ &= 18/38 = p.\end{aligned}$$

By the Law of Large Numbers, we know that the proportion of Fred's wins, say, $\widehat{p}$, will get close to $p = 18/38$ if he plays the game a large number of times. That is, if he continues to play over and over again, he will win approximately $100(18/38) = 47.4$ percent of the time in the long run; this has little to do with what will happen on the next play.

REMARK: The text describes Barney's erroneous reasoning as the "Law of Small Numbers." Many people incorrectly believe in this; that is, they expect even short sequences of random events to show the kind of average behavior that, in fact, appears only in the long run. Sports commentators are all guilty of this!

4.4.2 Variances: Discrete case

VARIANCE OF A DISCRETE RANDOM VARIABLE: Suppose that X is a discrete random variable whose distribution is

x	x_1	x_2	x_3	$\cdots$	x_k
Probability	p_1	p_2	p_3	$\cdots$	p_k

To find the **variance** of X, we use the following formula:

$$\begin{aligned}\sigma_X^2 &= (x_1 - \mu_X)^2p_1 + (x_2 - \mu_X)^2p_2 + (x_3 - \mu_X)^2p_3 + \cdots + (x_k - \mu_X)^2p_k \\ &= \sum_{i=1}^{k}(x_i - \mu_X)^2p_i.\end{aligned}$$

The **standard deviation** is the positive square root of the variance; i.e., $\sigma_X = \sqrt{\sigma_X^2}$.

Example 4.16. In Example 4.15, we saw the probability distribution for X, the number of stop lights at which I stop.

x	0	1	2	3	4	5	6	7	8
Probability	0.20	0.20	0.30	0.20	0.04	0.02	0.02	0.01	0.01

In Example 4.17, we computed the mean to be $\mu_X = 1.93$. The variance of X is given by

$$\begin{aligned}
\sigma_X^2 &= \sum_{i=1}^{9}(x_i - \mu_X)^2 p_i \\
&= (0-1.93)^2(0.20) + (1-1.93)^2(0.20) + (2-1.93)^2(0.30) \\
&\quad + (3-1.93)^2(0.20) + (4-1.93)^2(0.04) + (5-1.93)^2(0.02) \\
&\quad + (6-1.93)^2(0.02) + (7-1.93)^2(0.01) + (8-1.93)^2(0.01) \\
&= 2.465.
\end{aligned}$$

Thus, the standard deviation of X is equal to $\sigma_X = \sqrt{2.465} = 1.570$ stops/day.

DIFFERENCES: In Chapter 1, we spent some time talking about the values $\overline{x}$ and s, the sample mean and sample standard deviation computed from a sample of data. These values are **statistics** because they are computed from a sample of data. On the other hand, the values μ_X and σ_X are **parameters**. They are not computed from sample data; rather, they are values that are associated with the population (i.e., distribution) of values that are possible. *As a general rule, we use sample statistics to estimate population parameters.* Thus, if I drove to work, say, $n = 25$ days, and computed the sample mean and sample standard deviation of these 25 daily observations, I would have $\overline{x}$ and s. If I didn't know the probability distribution of X, the number of stops per day, I could use these values as **estimates** of μ_X and σ_X, respectively.

MATERIAL TO SKIP: From the text (MM), we are not going to cover the "Rules for Means" section on pages 298-299 and the "Rules for Variances" section on pages 301-304. We are also going to skip Section 4.5 "General Probability Rules."

5 Sampling Distributions

5.1 The binomial distribution

BERNOULLI TRIALS: Many experiments can be envisioned as a sequence of trials. **Bernoulli trials** are trials that have the following characteristics:

(i) each trial results in either a "success" or a "failure" (i.e., there are only 2 possible outcomes per trial),

(ii) there are n trials (where n is fixed in advance),

(iii) the trials are **independent**, and

(iv) the probability of success, denoted as p, $0 < p < 1$, is the same on every trial.

TERMINOLOGY: With a sequence of n Bernoulli trials, define X by

$$X = \text{number of successes (out of } n).$$

Then, X is said to have a **binomial distribution** with parameters n (the number of trials performed) and success probability p. We write

$$X \sim \mathcal{B}(n, p)$$

NOTE: A binomial random variable is **discrete** because it can only assume $n+1$ values.

Example 5.1. Each of the following situations might be modeled as binomial experiments. Are you satisfied with the Bernoulli assumptions in each instance?

(a) Suppose we flip a fair coin 10 times and let X denote the number of tails in 10 flips. Here, $X \sim \mathcal{B}(n = 10, p = 0.5)$.

(b) In a field experiment, forty percent of all plots respond to a certain treatment. I have four plots of land to be treated. If X is the number of plots that respond to the treatment, then $X \sim \mathcal{B}(n = 4, p = 0.4)$.

(c) In a large African city, the prevalence rate for HIV is about 12 percent. Let X denote the number of HIV infecteds in a sample of 500 individuals. Here, $X \sim \mathcal{B}(n = 500, p = 0.12)$.

(d) It is known that screws produced by a certain company do not meet specifications (i.e., are defective) with probability 0.001. Let X denote the number of defectives in a package of 40. Then, $X \sim \mathcal{B}(n = 40, p = 0.001)$.

Example 5.2. Explain why the following are **not** binomial experiments.

(a) I draw 3 cards from an ordinary deck and count X, the number of aces. Drawing is done without replacement.

(b) A couple decides to have children until a girl is born. Let X denote the number of children the couple will have.

(c) In a sample of 5000 individuals, I record the age of each person, denoted as X.

(d) A chemist repeats a solubility test ten times on the same substance. Each test is conducted at a temperature 10 degrees higher than the previous test. Let X denote the number of times the substance dissolves completely.

GOAL: We would like to compute probabilities associated with binomial experiments, so we need to derive a formula that allows us to do this. Recall that X is the number of successes in n Bernoulli trials, and p is the probability of success on any one trial. How can we get **exactly** x successes in n trials? Denoting

$$S = \text{success}$$
$$F = \text{failure}$$

for an individual Bernoulli trial, a possible outcome in the underlying sample space (for n Bernoulli trials) is

$$\underbrace{SSFSFSFFS \cdots FSF}_{n \text{ trials}}$$

Because the individual trials are **independent**, the probability that we get any particular ordering of x successes and $n - x$ failures is $p^x(1-p)^{n-x}$. Now, how many ways are there

to choose x successes from n trials? The answer to this last question is $\binom{n}{x}$, a **binomial coefficient** computed as follows:

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}.$$

Thus, for values of $x = 0, 1, 2, ..., n$, the **probability formula** for X is

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

RECALL: For any positive integer a,

$$a! = a \times (a-1) \times (a-2) \times \cdots \times 2 \times 1.$$

The symbol $a!$ is read "a-factorial." By definition, $0! = 1$. Also, recall that $a^0 = 1$.

Example 5.3. In Example 5.1(b), assume that X, the number of plots that respond to a treatment, follows a binomial distribution with $n = 4$ trials and success probability $p = 0.4$. That is, assume that $X \sim \mathcal{B}(n = 4, p = 0.4)$. What is the probability that exactly 2 plots respond? That is, what is $P(X = 2)$?

SOLUTION. Note that

$$\binom{4}{2} = \frac{4!}{2!\,(4-2)!} = \frac{24}{2 \times 2} = 6.$$

Thus, using the binomial probability formula, we have

$$P(X = 2) = \binom{4}{2} (0.4)^2 (1-0.4)^{4-2} = 6 \times 0.16 \times 0.36 = 0.3456.$$

So, you can see that computing binomial probabilities is quite simple. In fact, we can compute all the individual probabilities associated with this experiment:

$$\begin{aligned}
P(X = 0) &= \tbinom{4}{0}(0.4)^0(1-0.4)^{4-0} = 1 \times (0.4)^0 \times (0.6)^4 = 0.1296 \\
P(X = 1) &= \tbinom{4}{1}(0.4)^1(1-0.4)^{4-1} = 4 \times (0.4)^1 \times (0.6)^3 = 0.3456 \\
P(X = 2) &= \tbinom{4}{2}(0.4)^2(1-0.4)^{4-2} = 6 \times (0.4)^2 \times (0.6)^2 = 0.3456 \\
P(X = 3) &= \tbinom{4}{3}(0.4)^3(1-0.4)^{4-3} = 4 \times (0.4)^3 \times (0.6)^1 = 0.1536 \\
P(X = 4) &= \tbinom{4}{4}(0.4)^4(1-0.4)^{4-4} = 1 \times (0.4)^4 \times (0.6)^0 = 0.0256.
\end{aligned}$$

Note that these probabilities sum to 1 (as they should).

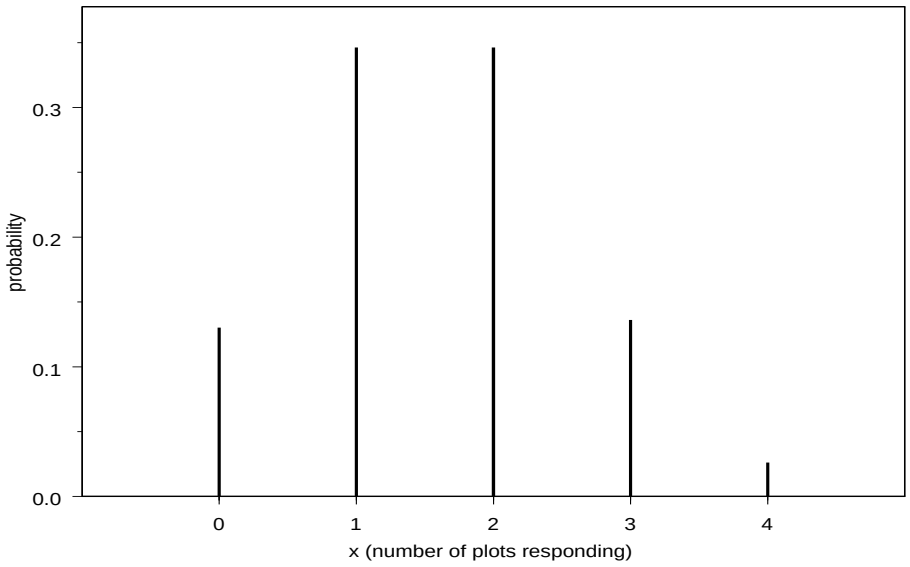


Figure 5.39: *Probability histogram for the number of plots which respond to treatment:* $X \sim \mathcal{B}(n = 4, p = 0.4)$.

ADDITIONAL QUESTIONS: (a) What is the probability that at least one plot responds to treatment? (b) at most one responds? (c) all four respond?

BINOMIAL PROBABILITY TABLE: Table C (MM, pages T6-10) contains the binomial probabilities

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

This table can be used to avoid computing probabilities by hand. *Blank entries in the table correspond to probabilities that are less than* 0.0001. Also, Minitab and SAS (as well as other packages like Excel) can be used as well.

Example 5.4. In a small Phase II clinical trial with 20 patients, let X denote the number of patients that respond to a new skin rash treatment. The physicians assume independence among the patients. Here, $X \sim \mathcal{B}(n = 20, p)$, where p denotes the probability of response to the treatment. For this problem, we'll assume that $p = 0.3$. We want to compute (a) $P(X = 5)$, (b) $P(X \geq 5)$, and (c) $P(X < 3)$.

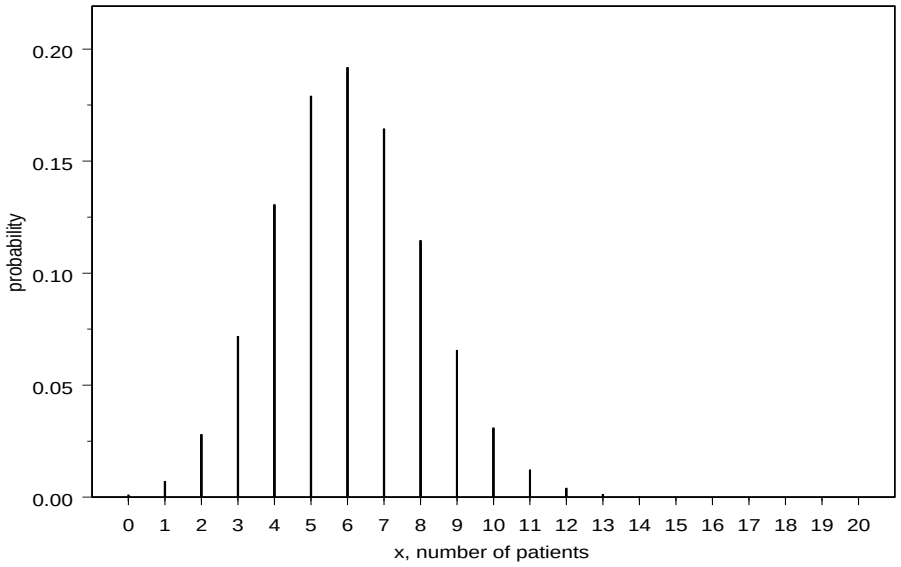


Figure 5.40: *Probability histogram for number of patients:* $X \sim \mathcal{B}(n = 20, p = 0.3)$.

SOLUTIONS. From Table C, the answer to part (a) is

$$P(X = 5) = \binom{20}{5}(0.3)^5(0.7)^{20-5} = 0.1789.$$

For part (b), we could compute

$$P(X \geq 5) = P(X = 5) + P(X = 6) + \cdots + P(X = 20).$$

To make this probability calculation, we would have to use the binomial probability formula 16 times and add up the results! Alternatively, we can use the complement rule.

$$\begin{aligned} P(X \geq 5) &= 1 - P(X \leq 4) \\ &= 1 - \{P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)\} \\ &= 1 - (0.0008 + 0.0068 + 0.0278 + 0.0716 + 0.1304) \\ &= 0.7626. \end{aligned}$$

The probabilities in part (b) were taken from Table C. For part (c), we can also use Table C to compute

$$P(X < 3) = P(X \leq 2) = 0.0008 + 0.0068 + 0.0278 = 0.0354.$$

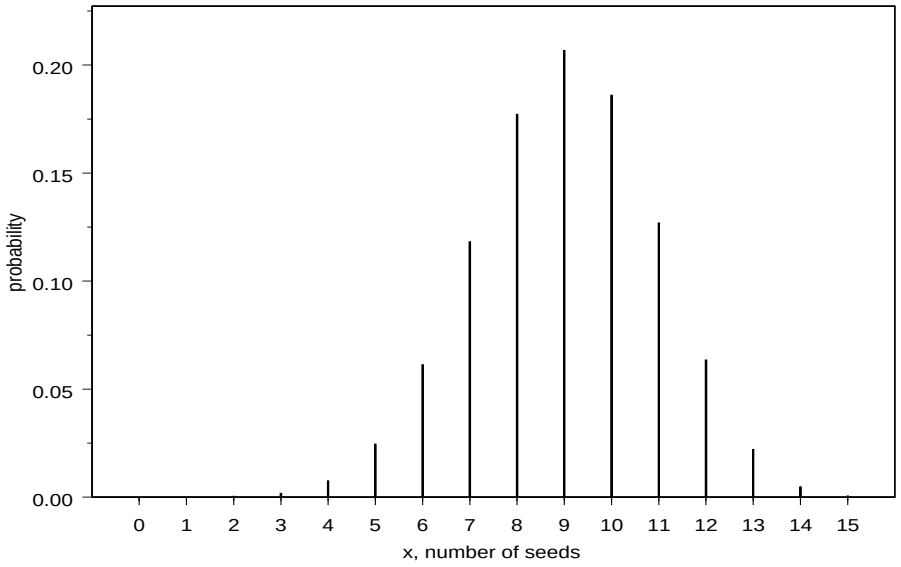


Figure 5.41: *Probability histogram for number of seeds:* $X \sim \mathcal{B}(n = 15, p = 0.6)$.

MEAN AND VARIANCE OF THE BINOMIAL DISTRIBUTION: Mathematics can show that if $X \sim \mathcal{B}(n, p)$, then the mean and standard deviation are given by

$$\begin{aligned} \mu_X &= np \\ \sigma_X &= \sqrt{np(1-p)}. \end{aligned}$$

Example 5.4. Suppose that 15 seeds are planted in identical soils and temperatures, and let X denote the number of seeds that germinate. If 60 percent of all seeds germinate on average and if we assume a $\mathcal{B}(15, 0.6)$ probability model for X, the mean number of seeds that germinate is

$$\mu_X = np = 15(0.6) = 9 \text{ seeds}$$

and the standard deviation is

$$\sigma_X = \sqrt{np(1-p)} = \sqrt{15(0.6)(1-0.6)} \approx 1.9 \text{ seeds}.$$

REMARK: Note that the $\mathcal{B}(n = 15, p = 0.6)$ probability histogram is **approximately symmetric**. As n increases, the $\mathcal{B}(n, p)$ distribution becomes more symmetric.

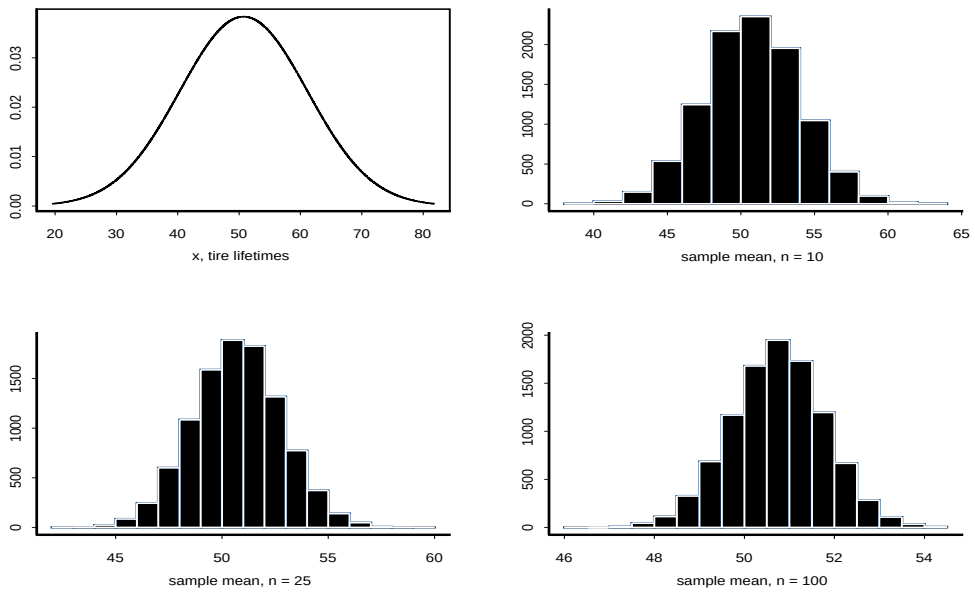


Figure 5.42: *Upper left:* $\mathcal{N}(50.73, 10.41)$ *probability model for tires; Upper right: sampling distribution of* $\overline{x}$ *when* $n = 10$*; Lower left: sampling distribution of* $\overline{x}$ *when* $n = 25$*; Lower right: sampling distribution of* $\overline{x}$ *when* $n = 100$.

5.2 An introduction to sampling distributions

Example 5.5. Many years of research have led to tread designs for automobile tires which offer good traction. The lifetime of Brand A tires, under regular driving conditions, is a variable which we will denote by X (measured in 1000s of miles). The variable X is assumed to follow a $\mathcal{N}(50.73, 10.41)$ probability distribution. This distribution is sketched in Figure 5.42 (upper left). The other three histograms were constructed using simulation. To construct the upper right histogram in Figure 5.42,

- I generated 10,000 samples, each of size $n = 10$, from a $\mathcal{N}(50.73, 10.41)$ distribution.
- I computed $\overline{x}$, the **sample mean**, for each sample.
- I plotted the 10,000 sample means $\overline{x}_1, \overline{x}_2, ..., \overline{x}_{10000}$ in a histogram (this is what you see in the upper right histogram).

NOTE: The other two histograms were computed in the same fashion, except that I used $n = 25$ (lower left) and $n = 100$ (lower right).

DISCUSSION: What we have just done is use simulation to generate the **sampling distribution** of the sample mean $\overline{x}$, when $n = 10$, $n = 25$, and $n = 100$. From these sampling distributions, we make the following observations:

- The sampling distributions are all centered around $\mu = 50.73$, the population mean.
- The variability in $\overline{x}$'s sampling distribution looks to get smaller as n increases.
- The sampling distributions all look normal!

TERMINOLOGY: The **population distribution** of a variable is the distribution of its values for all members of the population. The population distribution is also the probability distribution of the variable when we choose one individual at random from the population. In Example 5.5, the population distribution is $\mathcal{N}(50.73, 10.41)$. This is the distribution for the population of Brand A tires.

TERMINOLOGY: The **sampling distribution** of a statistic is the distribution of values taken by the statistic in repeated sampling using samples of the same size. In Example 5.5, the sampling distribution of $\overline{x}$ has been constructed by simulating samples of individual observations from the $\mathcal{N}(50.73, 10.41)$ population distribution, and then computing $\overline{x}$ for each sample.

OBSERVATION: A statistic from a random sample or randomized experiment is a random variable! This is true because it is computed from a sample of data (which are regarded as realizations of random variables). Thus, the sampling distribution of a statistic is simply its probability distribution!

REMARK: The sampling distribution of a statistic summarizes how the statistic behaves in repeated sampling. This is an important notion to understand because many statistical procedures that we will discuss are based on the idea of "repeated sampling."

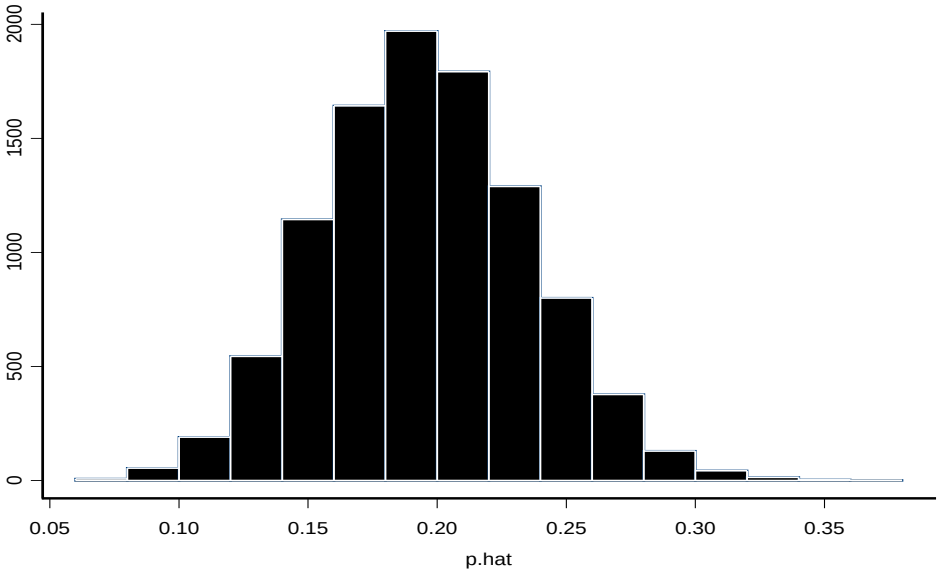


Figure 5.43: *Sampling distribution of* $\widehat{p}$*, when* $n = 100$ *and* $p = 0.2$.

5.3 Sampling distributions of binomial proportions

Example 5.6. (see also Example 3.16). A Columbia-based health club wants to estimate p, the proportion of Columbia residents who enjoy running as a means of cardiovascular exercise. Since p is a numerical measure of the **population** (i.e., Columbia residents), it is a **parameter**. To estimate p, suppose that we take a random sample n residents and record

$$X = \text{ the number of residents that enjoy running (out of } n).$$

Then, the **sample proportion** is simply

$$\widehat{p} = \frac{X}{n}.$$

Thus, the sample proportion is simply the binomial count X divided by the sample size n. We use $\widehat{p}$ (a **statistic**) to estimate the unknown p! In Example 3.16, recall that we simulated the sampling distribution of $\widehat{p}$ under the assumption that $n = 100$ and $p = 0.2$ (see Figure 5.43).

SAMPLING DISTRIBUTION OF THE SAMPLE PROPORTION: Let $\widehat{p} = X/n$ denote the **sample proportion** of successes in a $\mathcal{B}(n, p)$ experiment. Mathematics can be used to show that the mean and standard deviation of $\widehat{p}$ are

$$\begin{aligned} \mu_{\widehat{p}} &= p \\ \sigma_{\widehat{p}} &= \sqrt{\frac{p(1-p)}{n}}. \end{aligned}$$

Furthermore, for large n, mathematics can show that the sampling distribution of $\widehat{p}$ is **approximately normal**! Putting this all together, we can say that

$$\widehat{p} \sim \mathcal{AN}\left(p, \sqrt{\frac{p(1-p)}{n}}\right),$$

for large n. The abbreviation $\mathcal{AN}$ stands for "approximately normal."

REMARK: It is important to remember that this is an approximate result (i.e., $\widehat{p}$ is not perfectly normal). This isn't surprising either; remember that $\widehat{p}$ is a **discrete** random variable because it depends on the binomial count X (which is discrete). What we are saying is that the discrete distribution of $\widehat{p}$ is **approximated** by a smooth normal density curve. The approximation is best

- for larger values of n, and
- for values of p closer to 1/2.

Because of these facts, as a **rule of thumb**, the text recommends that both $np \geq 10$ and $n(1-p) \geq 10$ be satisfied. Under these conditions, you can feel reasonably assured that the normal approximation is satisfactory.

NOTE: Figure 5.45 displays simulated sampling distributions for $\widehat{p}$ for different values of n and p. Two values of p are used. When $p = 0.1$ (top row), the normal approximation is not adequate until $n = 100$ (when $n = 10$ and $n = 40$, the approximation is lousy!). On the other hand, when $p = 0.5$ (bottom row), the normal approximation already looks to be satisfactory when $n = 40$.

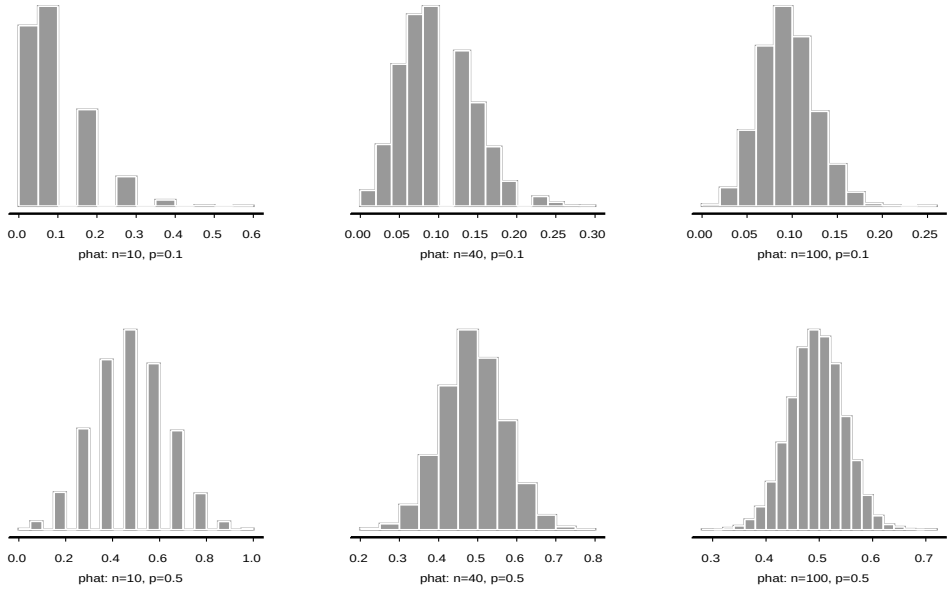


Figure 5.44: *Sampling distributions for the sample proportion $\widehat{p}$ using different n and p.*

Example 5.7. Suppose that in Example 5.6, we plan to take a sample of $n = 100$ residents. If $n = 100$ and $p = 0.2$, what is the probability that 32 or more residents say that they enjoy running as a means of exercise? Do you think this probability will be small or large? (Before you read on, look at Figure 5.43).

SOLUTION. To solve this problem, we could compute an **exact answer** (i.e., one that does not rely on an approximation). If X denotes the number of residents that enjoy running in our sample, and if we assume a binomial model, we know that $X \sim \mathcal{B}(100, 0.2)$. Thus, we could compute

$$\begin{aligned} P(X \geq 32) &= \binom{100}{32}(0.2)^{32}(0.8)^{68} + \binom{100}{33}(0.2)^{33}(0.8)^{67} + \cdots + \binom{100}{100}(0.2)^{100}(0.8)^{0} \\ &= 0.0031 \quad \text{(using Minitab)}. \end{aligned}$$

This is an exact answer, under the binomial model. We could also get an **approximate answer** using the normal approximation. Here, $np = 100(0.2) = 20$ and $n(1-p) = 100(0.8) = 80$, which are both larger than 10. Thus, we know that

$$\widehat{p} \sim \mathcal{AN}\left(0.2, \sqrt{\frac{0.2(1-0.2)}{100}}\right).$$

Using this fact, we can compute

$$\begin{aligned} P(X \geq 32) &= P(\widehat{p} \geq 0.32) \\ &\approx P\left(Z \geq \frac{0.32 - 0.2}{\sqrt{\frac{0.2(1-0.2)}{100}}}\right) \\ &= P(Z \geq 3.0) = 0.0013. \end{aligned}$$

Thus, the approximation is very close to the exact answer! This should convince you that the normal approximation to the sampling distribution of $\widehat{p}$ can be very good. Furthermore, normal probability calculations are often easier than exact lengthy binomial calculations (if you don't have statistical software!).

DISCUSSION: In Example 5.7, suppose that we did observe $\widehat{p} = 0.32$ in our sample. Since $P(\widehat{p} \geq 0.32)$ is so small, what does this suggest? This small probability might lead us to question whether or not p truly is 0.2! Which values of p would be more consistent with the observed $\widehat{p} = 0.32$: values of p **larger** than 0.2 or values of p **smaller** than 0.2?

REMARK: One can use mathematics to show that the $\mathcal{B}(n, p)$ and the $\mathcal{N}\{np, \sqrt{np(1-p)}\}$ distributions are also "close," when the sample size n is large. *Put another way, one could also use the normal approximation when dealing with binomial counts.* We won't pay too much attention to this, however, since a problem involving $X \sim \mathcal{B}(n, p)$ can always be stated as a problem involving $\widehat{p}$. Hence, there is no need to worry about the continuity correction material in MM (p. 347-8). Besides, we can always make exact binomial probability calculations with the help of statistical software (as in Example 5.7).

Example 5.8. Hepatitis C (HCV) is a viral infection that causes cirrhosis and cancer of the liver. Since HCV is transmitted through contact with infectious blood, screening donors is important to prevent further transmission. Currently, the worldwide seroprevalence rate of HCV is around 3%, and the World Health Organization has projected that HCV will be a major burden on the US health care system before the year 2020. A study was performed recently at the Blood Transfusion Service in Xuzhou City, China. The study involved an SRS of $n = 1875$ individuals and each was tested for the HCV antibody. If $p = 0.03$, what is the probability that 70 or more individuals will test positive?

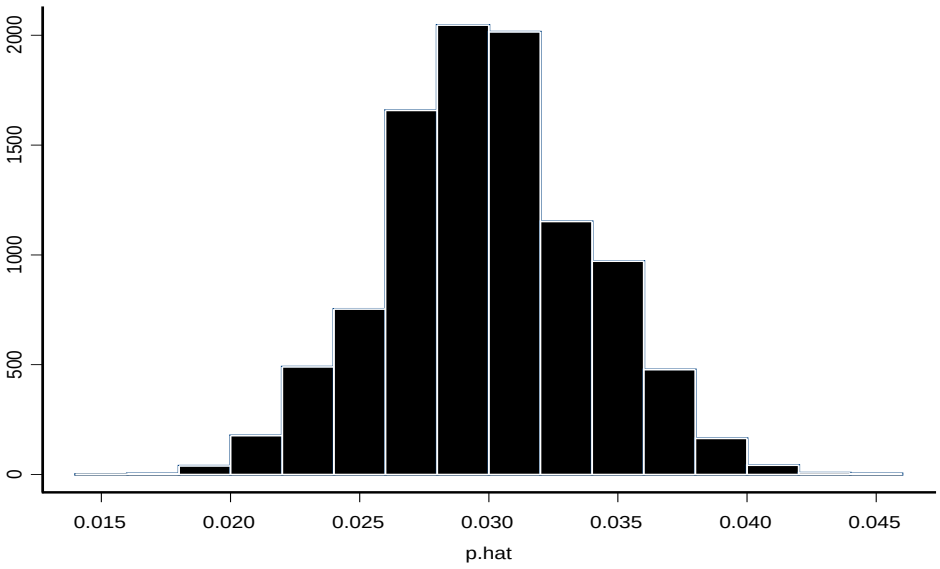


Figure 5.45: *Sampling distribution of* $\widehat{p}$*, when* $n = 1875$ *and* $p = 0.03$.

SOLUTION. If X denotes the number of infecteds, and if we assume a binomial model (is this reasonable?), then $X \sim \mathcal{B}(n = 1875, p = 0.03)$. Here, $np = 1875(0.03) = 56.25$ and $n(1-p) = 1875(0.97) = 1818.75$. Thus, we can feel comfortable with the normal approximation for $\widehat{p}$; that is,

$$\widehat{p} \sim \mathcal{AN}\left(0.03, \sqrt{\frac{0.03(1-0.03)}{1875}}\right).$$

Using this fact, we can compute

$$\begin{aligned} P(X \geq 70) &= P\left(\widehat{p} \geq \frac{70}{1875}\right) \\ &\approx P\left(Z \geq \frac{\frac{70}{1875} - 0.03}{\sqrt{\frac{0.03(1-0.03)}{1875}}}\right) \\ &= P(Z \geq 1.86) = 0.0314. \end{aligned}$$

Again, the event $\{X \geq 70\}$ is not too likely under the assumption that $p = 0.03$. Thus, if we **did** observe 70 or more HCV positives in our sample of $n = 1875$, what might this suggest about individuals living near Xuzhou City? Note that the exact value of $P(X \geq 70)$, using Minitab, is 0.0339 (very close to the approximation!).

5.4 Sampling distributions of sample means

REMARK: Binomial counts and sample proportions are discrete random variables that summarize categorical data. To summarize data for continuous random variables, we use sample means, percentiles, sample standard deviations, etc. These statistics also have sampling distributions! In this subsection, we pay particular attention to the sampling distribution of $\overline{x}$, the **sample mean**.

DISCUSSION: In Example 5.5, we simulated the sampling distribution of $\overline{x}$ using the $\mathcal{N}(50.73, 10.41)$ population distribution. We saw that, in each case (i.e., when $n = 10$, $n = 25$, and when $n = 100$) the sampling distribution of $\overline{x}$ looked to be well-modeled by a normal distribution (*with the same population mean, but with smaller standard deviation*). This is not a surprising fact, in light of the following mathematical fact.

FACTS: Suppose that the statistic $\overline{x}$ is computed using an SRS from a population distribution (not necessarily normal!) with mean μ and standard deviation σ. The mathematical arguments on page p. 360-1 (MM) show that

$$\begin{aligned} \mu_{\overline{x}} &= \mu \\ \sigma_{\overline{x}} &= \frac{\sigma}{\sqrt{n}}. \end{aligned}$$

That is, the mean of the sampling distribution of $\overline{x}$ is equal to the population mean μ and the standard deviation of the sampling distribution of $\overline{x}$ is $1/\sqrt{n}$ times the population standard deviation. In light of the preceding facts, we can state that

- The sample mean $\overline{x}$ is an **unbiased estimator** for the population mean μ.
- The precision with which $\overline{x}$ estimates μ improves as the sample size increases. This is true because the standard deviation of $\overline{x}$ gets smaller as n gets larger!
- Thus, the sample mean $\overline{x}$ is a very good estimate for the population mean μ; it has the desirable properties that an estimator should have; namely, unbiasedness and small variability (for large sample sizes).

BIG RESULT: If a population has a $\mathcal{N}(\mu, \sigma)$ distribution, and the statistic $\overline{x}$ is computed from an SRS from this population distribution, then

$$\overline{x} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Note that this is not an approximate result; i.e., it is an **exact result**. *If the underlying population distribution is $\mathcal{N}(\mu, \sigma)$, the sample mean $\overline{x}$ will also vary according to a normal distribution. This sampling distribution will have the same mean, but will have smaller standard deviation.* The simulated sampling distributions of $\overline{x}$ in Figure 5.42 simply reinforce this result.

Example 5.9. In the interest of pollution control, an experimenter records X, the amount of bacteria per unit volume of water (measured in mg/cm^3). The population distribution for X is assumed to be $\mathcal{N}(48, 10)$.

(a) What is the probability that a single water specimen's bacteria amount will exceed 50 mg/cm^3?

SOLUTION. Here, we use the population distribution $\mathcal{N}(48, 10)$ to compute

$$\begin{aligned} P(X \geq 50) &= P\left(Z \geq \frac{50-48}{10}\right) \\ &= P(Z \geq 0.2) = 0.4207. \end{aligned}$$

(b) Suppose that the experimenter takes an SRS of $n = 100$ water specimens. What is the probability that the **sample mean** $\overline{x}$ will exceed 50 mg/cm^3?

SOLUTION. Here, we need to use the sampling distribution of the sample mean $\overline{x}$. Since the population distribution is normal, we know that

$$\overline{x} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \sim \mathcal{N}\left(48, \frac{10}{\sqrt{100}}\right) \sim \mathcal{N}\left(48, 1\right).$$

Thus,

$$\begin{aligned} P(\overline{x} \geq 50) &= P\left(Z \geq \frac{50-48}{1}\right) \\ &= P(Z \geq 2.00) = 0.0228. \end{aligned}$$

Thus, we see that $P(X \geq 50)$ and $P(\overline{x} \geq 50)$ are very different probabilities!

CURIOSITY: What happens when we sample data from a **nonnormal** population distribution? Does $\overline{x}$ still have a normal distribution?

Example 5.10. Animal scientists are interested in the proximate mechanisms animals use to guide movements. The ultimate bases for movements are related to animal adaptations to different environments and the development of behaviors that bring them to those environments. Denote by X the distance (in meters) that an animal moves from its birth site to the first territorial vacancy. For the banner-tailed kangaroo rat, the population distribution of X has mean $\mu = 10$ and standard deviation $\sigma = \sqrt{10}$. The density curve that summarizes this population distribution is given in Figure 5.46 (upper left). This density curve is sometimes called an **exponential distribution**.

INVESTIGATION: Suppose that I take an SRS of n banner-tailed kangaroo rats and record X for each rat. How will the sample mean $\overline{x}$ behave in repeated sampling? That is, what values of $\overline{x}$ can I expect to see when sampling from this exponential distribution with mean $\mu = 10$? To investigate this, we again use a **simulation study**.

- Generate 10,000 samples, each of size n, from an exponential population distribution with mean $\mu = 10$ and standard deviation $\sigma = \sqrt{10}$.
- Compute $\overline{x}$, the **sample mean**, for each sample.
- Plot the 10,000 sample means $\overline{x}_1, \overline{x}_2, ..., \overline{x}_{10000}$ in a histogram. This histogram will be the approximate sampling distribution of $\overline{x}$.

OBSERVATIONS: We first note that the exponential distribution is heavily skewed right! However, from Figure 5.46, we note that

- the sampling distribution for $\overline{x}$, when $n = 5$ is still skewed right, but it already is taking a unimodal shape.
- the sampling distribution for $\overline{x}$, when $n = 25$ is almost symmetric. Sharp eyes might be able to detect a slight skew to the right.
- the sampling distribution for $\overline{x}$, when $n = 100$ is nearly perfectly normal in shape.

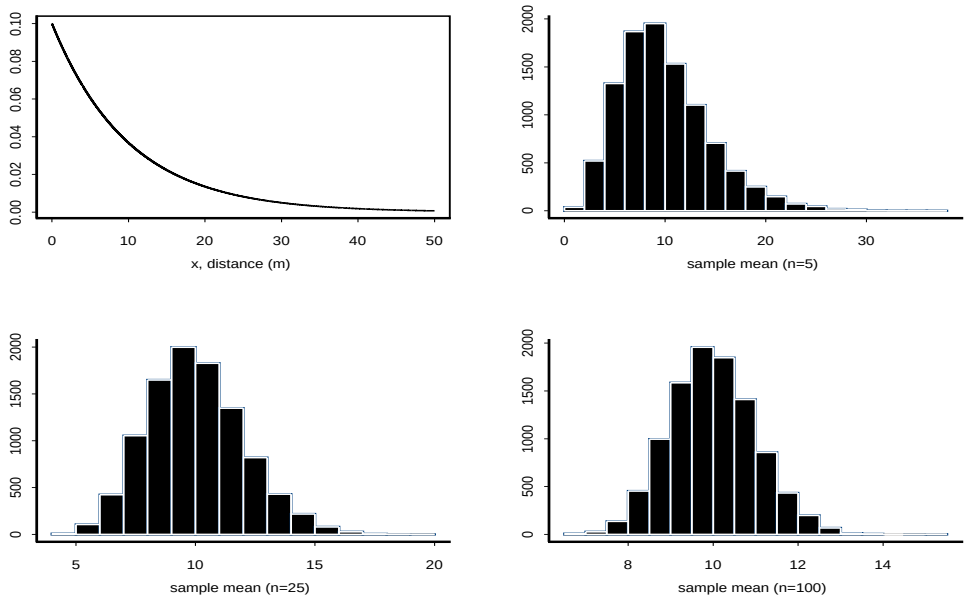


Figure 5.46: *Upper left: population distribution of distance traveled to first territorial vacancy for banner-tailed kangaroo rats; Upper right: sampling distribution of* $\overline{x}$ *when* $n = 5$*; Lower left: sampling distribution of* $\overline{x}$ *when* $n = 25$*; Lower right: sampling distribution of* $\overline{x}$ *when* $n = 100$.

REMARK: What we have just witnessed in Example 5.10 is an application of the following important result. It is hard to overstate the importance of this result in statistics.

CENTRAL LIMIT THEOREM: Draw an SRS of size n from **any population distribution** with mean μ and standard deviation σ. When n is large, the sampling distribution of $\overline{x}$ is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$; that is,

$$\overline{x} \sim \mathcal{AN}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

The real novelty in this result is that the sample mean will be approximately normal for large sample sizes, even if the original population distribution is not! Remember that $\mathcal{AN}$ is an abbreviation for **approximately normal**.

HOW GOOD IS THE APPROXIMATION?: Since the Central Limit Theorem (CLT) only offers an **approximate** sampling distribution for $\overline{x}$, one might naturally wonder

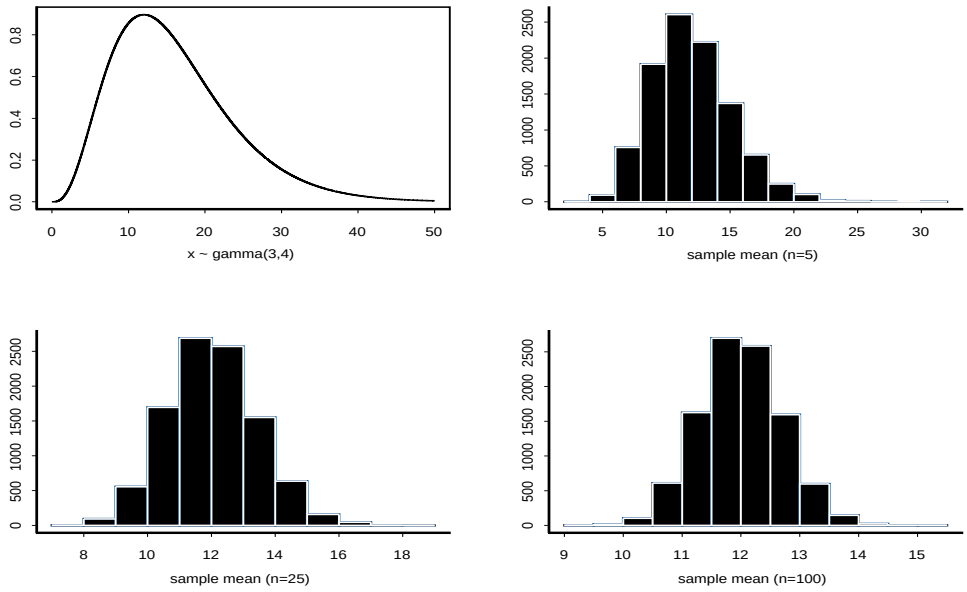


Figure 5.47: *Upper left: gamma population distribution with* $\mu = 12$ *and* $\sigma = \sqrt{48}$*; Upper right: sampling distribution of* $\overline{x}$ *when* $n = 5$*; Lower left: sampling distribution of* $\overline{x}$ *when* $n = 25$*; Lower right: sampling distribution of* $\overline{x}$ *when* $n = 100$.

exactly how good the approximation is. In general, the goodness of the approximation **jointly** depends on

- the sample size, n, and
- the skewness in the underlying population distribution.

JOINT BEHAVIOR: For heavily skewed population distributions (such as the exponential), we need the sample size to be larger for the CLT to "work." On the other hand, for population distributions that are not too skewed, the CLT will "work" even when smaller values of n are used. To illustrate, consider the population distribution in Figure 5.47 (upper left). This is a **gamma distribution** with mean $\mu = 12$ and standard deviation $\sigma = \sqrt{48}$. You should see that with only $n = 5$ (upper right), the normal approximation to the sampling distribution of $\overline{x}$ is already very good! At $n = 25$, it is already almost perfect (compare these findings with the exponential distribution in Figure 5.46).

QUESTION: *So, how large does n have to be for the CLT to "work?"* Unfortunately, there is no minimum sample size n that fits in every situation! As a general rule, larger sample sizes are needed when the population distribution is more skewed; smaller sample sizes are needed when the population distribution is less skewed.

Example 5.11. There are many breeds of potatoes that are sold and studied throughout the United States and all over the world. For all varieties, it is important to conduct carefully designed experiments in order to study the marketable maturity and adaptability with emphasis upon general appearance, susceptibility to infection, and external and internal defects which affect specific markets. For one specific variety of potatoes, *Cherry Red*, an experiment was carried out using 40 plots of land. The plots were fairly identical in every way in terms of soil composition, amount of precipitation, etc. The population distribution of yields from last year's harvest was estimated to have mean $\mu = 158.2$ and standard deviation $\sigma = 14.9$ (bushels/plot). Suppose that this year's average yield (in the forty plots) was $\overline{x} = 155.8$. Would you consider this unusual?

SOLUTION. We can answer this question by computing $P(\overline{x} \leq 155.8)$ under the assumption that $\mu = 158.2$ and $\sigma = 14.9$. From the CLT, we know that

$$\overline{x} \sim \mathcal{AN}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \sim \mathcal{AN}\left(158.2, \frac{14.9}{\sqrt{40}}\right) \sim \mathcal{AN}\left(158.2, 2.356\right).$$

Thus, we have that

$$\begin{aligned} P(\overline{x} \leq 155.8) &\approx P\left(Z \leq \frac{155.8 - 158.2}{2.356}\right) \\ &= P(Z \leq -1.02) = 0.1539. \end{aligned}$$

This probability is not that small! In the light of this, would you consider $\overline{x} = 155.8$ to be all that unusual?

FINAL NOTE: That the sample proportion

$$\widehat{p} \sim \mathcal{AN}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

for large n is really just an application of the Central Limit Theorem! This is true because the sample proportion $\widehat{p}$ is really just an **average** of values that are 0s and 1s. For more detail, see p. 365 (MM).

6 Introduction to Statistical Inference

6.1 Introduction

We are now ready to formally start the second part of this course; namely, we start our introduction to **statistical inference**. As we have learned, this part of the analysis aims to answer the question, *"What do the observed data in my sample suggest about the population of individuals?"*

TWO MAIN AREAS OF STATISTICAL INFERENCE:

- estimation of one or more population parameters (e.g., **confidence intervals**)
- **hypothesis tests** for population parameters.

PREVIEW: We wish to learn about a single population (or multiple populations) based on the observations in a random sample from an observational study or from a randomized experiment. The observed data in our sample arise by chance, because of natural variability (e.g., biological, etc.) in the population of interest. Thus, the statements that we make about the population will be **probabilistic** in nature. *From our sample data, we will never get to make deterministic statements about our population.*

Example 6.1. Here are some examples of the types of problems that we will consider throughout the rest of the course.

- A veterinarian is studying the effect of lead-poisoning on a species of Canadian geese. She wants to estimate μ, the mean plasma-glucose level (measured in mg/100 mL plasma) for the population of infected geese.
- An aluminum company is experimenting with a new design for electrolytic cells. A major design objective is maximize a cell's mean service life. Is the new design superior to the industry standard?

- In a large scale Phase III clinical trial involving patients with advanced lung cancer, the goal is to determine the efficacy of a new drug. Ultimately, we would like to know whether this drug will extend the life of lung cancer patients as compared to current available therapies.

- A tribologist is studying the effects of different lubrications in the design of a rocket engine. He is interested in estimating the probability of observing a cracked bolt in a particular mechanical assembly in the population of engines in the fleet.

- In laboratory work, it is desirable to run careful checks on the variability of readings produced on standard samples. In a study of the amount of calcium in drinking water undertaken as part of a water quality assessment, multiple observations are taken on the same sample to investigate whether or not there was too much variability in the measurement system.

- The Montana Agricultural Experiment Station continues to develop wheat varieties widely adopted by Montana producers. In an experiment with four varieties, researchers are interested if there is significant variability among the wheat yields.

6.2 Confidence intervals for a population mean when σ is known

REMARK: We start our discussion of confidence intervals in a rather unrealistic setting. To be specific, we would like to construct a confidence interval for the population mean μ when σ (the population standard deviation) is **known**.

REMARK: Why is this an unrealistic situation? Recall that σ is a **parameter**; that is, it is a measure that summarizes the entire population. Rarely (i.e., almost never) will population parameters be known. However, there are instances where we might have a "reliable guess" for σ (perhaps from past data or knowledge). In this situation, the methods that we discuss this chapter are applicable. In the next chapter, we discuss the more realistic situation of when σ is not known.

RECALL: With data from a random sample, we have seen that the sample mean $\overline{x}$ is an **unbiased estimate** for the population mean μ. However, we know that the value of $\overline{x}$ changes from sample to sample whereas μ does not change. That is, the estimate $\overline{x}$ is a "likely value" of μ, but that is all. Recall that if our data are normally distributed, then

$$\overline{x} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Since $\overline{x}$ can assume many different values (in theory, an infinite number of values), instead of reporting a single value as an estimate for a parameter, let's report an *interval of plausible values* for the parameter. That is, let's report an interval that is likely to contain the unknown parameter μ. Of course, the term "likely" means that probability is going to come into play.

TERMINOLOGY: A **confidence interval** for a parameter is an interval of likely (plausible) values for the parameter.

DERIVATION: To find a confidence interval for the population mean μ, we start with the sampling distribution of $\overline{x}$. In particular, we know that from a random sample of normally distributed data, the statistic

$$\overline{x} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Thus, we know that, by standardizing,

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1),$$

and

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

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentile from the standard normal distribution; i.e., $z_{\alpha/2}$ satisfies

$$P\left(Z < -z_{\alpha/2}\right) = P\left(Z > z_{\alpha/2}\right) = \alpha/2.$$

For example, if $\alpha = 0.05$, then $\alpha/2 = 0.025$ and $z_{0.025} = 1.96$ (from Table A; see also Figure 6.48).

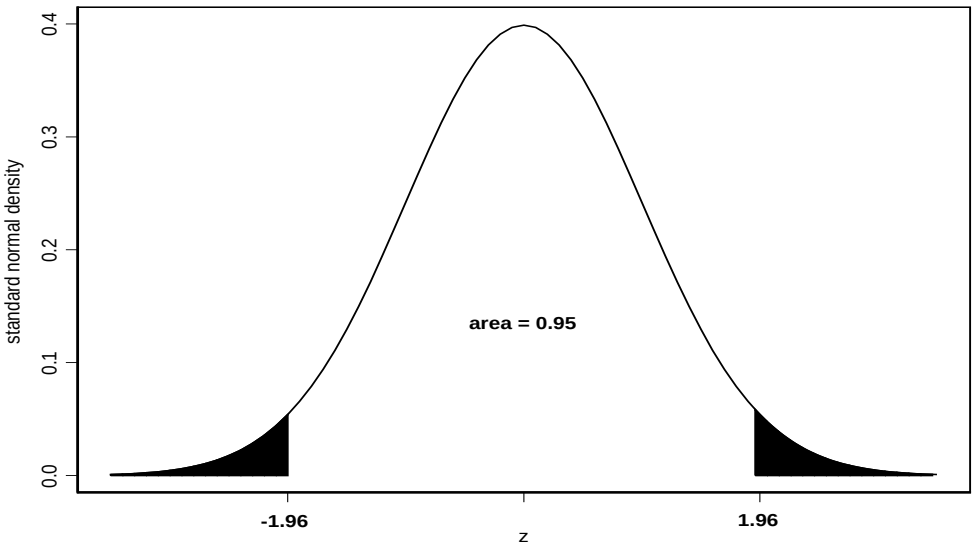


Figure 6.48: *Standard normal density with unshaded area equal to* 0.95.

With $\alpha = 0.05$ (see Figure 6.48), we have that

$$\begin{aligned}
0.95 &= P\left(-1.96 < \frac{\overline{x}-\mu}{\sigma/\sqrt{n}} < 1.96\right) \\
&= P\left(-1.96\frac{\sigma}{\sqrt{n}} < \overline{x}-\mu < 1.96\frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(1.96\frac{\sigma}{\sqrt{n}} > \mu-\overline{x} > -1.96\frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(\overline{x}+1.96\frac{\sigma}{\sqrt{n}} > \mu > \overline{x}-1.96\frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(\overline{x}-1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{x}+1.96\frac{\sigma}{\sqrt{n}}\right).
\end{aligned}$$

Thus, we have "trapped" the unknown population mean μ in between the random endpoints $\overline{x} - 1.96\sigma/\sqrt{n}$ and $\overline{x} + 1.96\sigma/\sqrt{n}$ with probability 0.95 (recall that σ is known).

TERMINOLOGY: We call

$$\left(\overline{x}-1.96\frac{\sigma}{\sqrt{n}},\ \overline{x}+1.96\frac{\sigma}{\sqrt{n}}\right).$$

a **95 percent confidence interval** for μ.

Example 6.2. The dissolved oxygen (DO) content (measured in mg/L) is recorded for a sample of $n = 6$ water specimens in a certain geographic location. Here are the data $x_1, x_2, ..., x_6$:

```
2.62    2.65    2.79    2.83    2.91    3.57
```

From past knowledge, a reliable guess of the population standard deviation is $\sigma = 0.3$ mg/L. Assuming that the DO contents are well-represented by a normal distribution, we would like to find a 95 percent confidence interval for μ, the population mean DO concentration (i.e., of all possible water specimens from this location). The sample mean is given by

$$\overline{x} = \frac{17.37}{6} = 2.90\text{mg/L}.$$

Thus, the 95 percent confidence interval for μ is

$$\left(2.90 - 1.96 \times \frac{0.3}{\sqrt{6}},\ 2.90 + 1.96 \times \frac{0.3}{\sqrt{6}}\right) \quad \text{or} \quad (2.66, 3.14).$$

Thus, we are 95 percent confident that the population mean DO concentration μ is between 2.66 and 3.14 mg/L.

BEHAVIOR: The probability associated with the confidence interval has to do with the **random endpoints**. The interval

$$\left(\overline{x} - 1.96\frac{\sigma}{\sqrt{n}},\ \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

is random (because $\overline{x}$ varies from sample to sample), not the parameter μ. The values of the endpoint change as $\overline{x}$ change! See Figure 6.3 on p. 386 (MM).

- *If we are allowed the freedom to take repeated random samples, each of size n, then* 95 *percent of the intervals constructed would contain the unknown parameter* μ. Hence, we are making a statement about the process of selecting samples and the methods used. Of course, this has to do with the quality of the data collection!
- We say that "we are **95 percent confident** that the interval contains the mean μ." In Example 6.1, we are 95 percent confident that the **population mean** DO concentration level μ is between 2.66 and 3.14 mg/L.

- It is not technically correct to say that "the probability that μ is between 2.66 and 3.14 mg/L is 0.95." Remember, μ is a fixed quantity; it is either in the the interval (2.66, 3.14) or it is not.

- The confidence interval is an interval of likely values for μ, the population mean (a parameter). In Example 6.1, some students would erroneously conclude that 95 percent of the individual DO measurements will be between 2.66 and 3.14 mg/L! To compute the proportion of **individual measurements** between 2.66 and 3.14 mg/L, we would have to use the population distribution of DO content; not the sampling distribution of $\overline{x}$!

REMARK: There is nothing special about using a **95 percent** confidence interval. We can use any **confidence level**; the only thing that will change is the value of $z_{\alpha/2}$ in Figure 6.48. Popular confidence levels used in practice are 99 percent, 95 percent (the most common), 90 percent, and even 80 percent. The larger the confidence level, the "more confident" we are that the resulting interval will contain the population mean μ.

TERMINOLOGY: A **level C confidence interval** for a parameter is an interval computed from sample data by a method that has probability C of producing an interval containing the true value of the parameter. The value C is called the **confidence level**. In decimal form, $C = 0.99$, $C = 0.95$, $C = 0.90$, $C = 0.80$, etc.

GENERAL FORM: The general form of a level $1-\alpha$ confidence interval for a population mean μ is

$$\left(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

To compute this interval, we need four things:

- the sample mean $\overline{x}$,

- the standard normal percentile $z_{\alpha/2}$,

- the population standard deviation σ, and

- the sample size n.

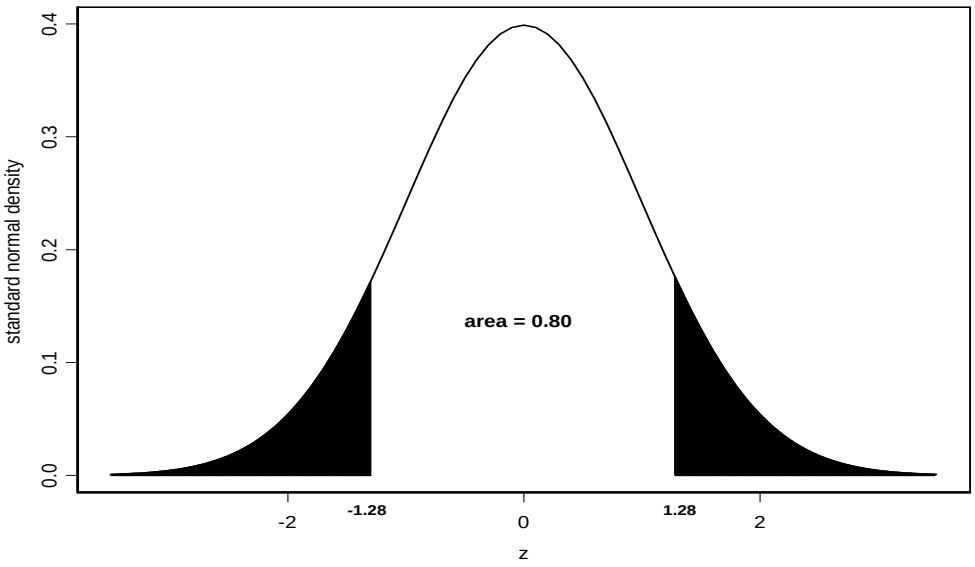


Figure 6.49: *Standard normal density with unshaded area equal to* 0.80.

FINDING $z_{\alpha/2}$: The value $z_{\alpha/2}$ is the upper $\alpha/2$ percentile of the standard normal distribution. Here are the values of $z_{\alpha/2}$ (found using Table A) for commonly used confidence levels. Make sure you understand where these values come from!!

α	0.20	0.10	0.05	0.01
$C = 1 - \alpha$	0.80	0.90	0.95	0.99
$z_{\alpha/2}$	1.28	1.65	1.96	2.58

EXERCISE: What would the value of $z_{\alpha/2}$ be for a 92 percent confidence interval? an 84 percent confidence interval? a 77 percent confidence interval?

Example 6.2. In Example 6.1, an 80 percent confidence interval for the population mean DO content μ is

$$\left(2.90 - 1.28 \times \frac{0.3}{\sqrt{6}},\ 2.90 + 1.28 \times \frac{0.3}{\sqrt{6}}\right) \quad \text{or} \quad (2.74, 3.06) \text{ mg/L}.$$

Thus, we are 80 percent confident that the population mean DO concentration μ is between 2.74 and 3.06 mg/L. Note that this is a shorter interval than the 95 percent

confidence interval (2.66, 3.14). Intuitively this makes sense; namely, the higher the confidence level, the lengthier the interval.

NONNORMAL DATA: When we derived the general form of a level $1-\alpha$ confidence interval for the population mean μ, we used the fact that

$$\overline{x} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

This is only (exactly) true if the data $x_1, x_2, ..., x_n$ are normally distributed. *However, what if the population distribution is nonnormal?* Even if the underlying population distribution is nonnormal, we know (from the Central Limit Theorem) that

$$\overline{x} \sim \mathcal{AN}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

when the sample size n is large. In this situation, we call

$$\left(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

an **approximate** level $1-\alpha$ confidence interval.

Example 6.3. In Example 1.4, we examined the shelf life data from $n = 25$ beverage cans in an industrial experiment. The data (measured in days) are listed below.

262	188	234	203	212	212	301	225	241	211	231	227	217
252	206	281	251	219	268	231	279	243	241	290	249	

In Chapter 1, we informally decided that a normal distribution was not a bad model for these data (although there was some debate). We computed the range to be

$$R = x_{max} - x_{min} = 301 - 188 = 113.$$

We can use the range to formulate a guess of σ. To see how, recall that the distribution of shelf lives was approximately symmetric; Thus, by the Empirical Rule, almost all the data should be within 3σ of the mean. This allows us to say that

$$R \approx 6\sigma \Longrightarrow \sigma = 113/6 \approx 18.8.$$

From the 25 cans, recall that we computed $\overline{x} = 238.96$. Thus, an approximate confidence interval for μ, the population mean shelf life, is given by

$$\left(238.96 - 1.96 \times \frac{18.8}{\sqrt{25}},\ 238.96 + 1.96 \times \frac{18.8}{\sqrt{25}}\right) \quad \text{or} \quad (231.59, 246.33).$$

Thus, we are approximately 95 percent confident that the population mean shelf life μ is between 231.59 and 246.33 days.

MARGIN OF ERROR: For a level $1 - \alpha$ confidence interval for a population mean μ, the **margin of error** is defined as

$$m = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

This is the quantity that we add and subtract to the sample mean $\overline{x}$ so that the confidence interval for μ can be "thought of" as

$$(\overline{x} - m,\ \overline{x} + m).$$

Recall that this interval is exact when the population distribution is normal and is approximately correct for large n in other cases. Clearly, the **length** of the interval then is simply

$$L = 2m = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

This expression for the length of a confidence interval for μ is helpful. We can use it to study how the length is related to the confidence level $1 - \alpha$, the sample size n, and the population standard deviation σ.

- As the confidence level increases, so does $z_{\alpha/2}$ (look at Figure 6.50; then look at Figure 6.48). This, in turn, increases the length of the confidence interval (this makes sense: the more confident we want to be, the larger the interval must be).
- As the sample size increases, the length of the interval decreases (note that n is in the denominator of L). This also makes sense: the more data (information) we have from the population, the more precise our interval estimate will be. Shorter intervals are desirable; they are more precise than larger intervals.

- The population standard deviation σ is a parameter; we can not think about "changing its value." However, if we were to focus on a particular **subpopulation**, the value of σ associated with this smaller population may be smaller. For example, suppose that in Example 6.3, we provided a guess of the population standard deviation to be $\sigma \approx 18.8$. If we were to focus on a smaller population of cans, say, only those cans produced by a particular manufacturer, it may be that the variability in the fill weights may be smaller (because, inherently, the cans themselves are more similar!). If we can focus on a particular subpopulation with a smaller σ, we can reduce the length of our interval. Of course, then our inference is only applicable to this smaller subpopulation.

SAMPLE SIZE DETERMINATIONS: Before one launches into a research investigation where data are to be collected, inevitably one starts with the simple question: *"How many observations do I need?"* The problem is, the answer is often not simple! The answer almost always depends on the resources available (e.g., time, money, space, etc.) and statistical issues like **confidence** and **power**. In single population problems, we usually can come up with a **sample size formula** which incorporates these statistical issues. Recall that the margin of error for a confidence interval for μ (when σ is known) is given by

$$m = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

This is an equation we can solve for n; in particular,

$$n = \left(\frac{z_{\alpha/2}\sigma}{m}\right)^2.$$

Thus, if we specify our confidence level $1-\alpha$, an estimate of the population standard deviation σ, and a desired margin of error m, we can find the sample size needed.

Example 6.4. In a biomedical experiment, we would like to estimate the mean lifetime of healthy rats (μ, measured in days) that are given a high dose of toxic substance. This may be done in an early phase clinical trial by researchers trying to find a maximum tolerable dose for humans. Suppose that we would like to write a 99 percent confidence

interval for μ with length no more than 4 days. The researchers have provided a guess of $\sigma \approx 8$ days. How many rats should we use for the experiment?

SOLUTION: With a confidence level of $1 - \alpha = 0.99$, our value of $z_{\alpha/2}$ is

$$z_{\alpha/2} = z_{0.01/2} = z_{0.005} = 2.58.$$

Also, if the length of the desired interval is 4 days, our margin of error is $m = 2$ days. The sample size needed is

$$n = \left(\frac{2.58 \times 8}{2}\right)^2 \approx 107.$$

Thus, we would need $n = 107$ rats to achieve these goals.

CURIOSITY: If collecting 107 rats is not feasible, we might think about weakening our requirements (after all, 99 percent confidence is high, and the margin of error is tight). Suppose that we used a 90 percent confidence level instead with margin of error $m = 5$ days. Then, the desired sample size would be

$$n = \left(\frac{1.65 \times 8}{5}\right)^2 \approx 7.$$

This is an easier experiment to carry out now (we need only 7 rats!). However, we have paid a certain price: less confidence and less precision in our interval estimate).

CAUTIONS REGARDING CONFIDENCE INTERVALS: Moore and McCabe (p. 393) offer the following cautions when using confidence intervals for inference:

- Data that we use to construct a confidence interval should be (or should be close to) a simple random sample. If the sample is biased, so will be the results. Poor data collection techniques inundated with nonsampling errors will produce poor results too!
- Confidence interval formulas for μ are different if you use different sampling designs (e.g., stratified, cluster, systematic, etc.). We won't discuss these.
- Outliers almost always affect the analysis. You need to be careful in checking for them and decide what to do about them.

- When the sample size n is small, and when the population distribution is highly nonnormal, the confidence interval formula we use for μ is probably a bad formula. This is true because the normal approximation to the sampling distribution for $\overline{x}$ is probably a bad approximation. The text offers an $n \geq 15$ guideline for most population distributions. That is, if your sample size is larger than or equal to 15, you're probably fine. However, this is only a guideline. If the underlying population distribution is very skewed, this may or may not be a good guideline (you might need n to be larger). Of course, it goes both ways; namely, if the underlying distribution is very symmetric, then this guideline might actually be too restrictive!

6.3 Hypothesis tests for the population mean when σ is known

Example 6.5. In a laboratory experiment to investigate the influence of different doses of vitamin A on weight gain over a three week period, $n = 5$ rats received a standard dose of vitamin A. The following weight increases (mg) were observed:

35 49 51 43 27

For now, we will assume that these data arise from a population distribution which is normal. A **reliable guess** of the population standard deviation is $\sigma \approx 14$ mg. The sample mean of these data is $\overline{x} = 41$ mg.

REMARK: Often, we take a sample of observations with a specific **research question** in mind. For example, consider the data on weight gains of rats treated with vitamin A discussed in Example 6.5. Suppose that we know from several years of experience that the mean weight gain of rats this age and during a three week period when they are **not** treated with vitamin A is 27.8 mg.

SPECIFIC QUESTION: If we treat rats of this age and type with vitamin A, how does this affect 3-week weight gain? That is, if we could administer the standard dose of vitamin A to the entire population of rats of this age and type, would the **population**

mean weight gain change from what it would be if no vitamin A was given? The sample results seem to suggest so, but whether or not this increase is representative of the population is an entirely different question!

STRATEGY: Since we can not observe the entire population of rats, we observe a sample of such rats, treat them with vitamin A, and view the sample as being randomly drawn from the population of rats treated with vitamin A. This population has an unknown mean weight gain, denoted by μ. Clearly, our question of interest may be regarded as a question about the parameter μ.

(i) Either $H_0 : \mu = 27.8$ mg is true; that is, the vitamin A treatment does not affect weight gain, and the mean is equal to what it would be if no vitamin A was given.

(ii) Or $H_1 : \mu \neq 27.8$ mg is true; that is, vitamin A does have an affect on the mean weight gain.

TERMINOLOGY: We call H_0 the **null hypothesis** and call H_1 the **alternative hypothesis**. A formal statistical inference procedure for deciding between H_0 and H_1 is called a **hypothesis test**.

APPROACH: Suppose that, in truth, the application of vitamin A has **no effect** on weight gain and that μ really is 27.8 mg. For the particular sample we observed, $\overline{x} = 41.0$ mg (based on $n = 5$). So, the key question becomes *"How likely is it that we would see a sample mean of* $\overline{x} = 41.0$ *mg if the population mean really is* $\mu = 27.8$ *mg?"*

- If $\overline{x} = 41.0$ mg is likely when H_0 is true, then we would not discount H_0 as a plausible explanation. That is, we would **not reject** H_0.
- If $\overline{x} = 41.0$ mg is not likely when H_0 is true, then this would cause us to think that H_0 is not a plausible explanation. That is, we would **reject** H_0 as an explanation.

THE FORMAL METHOD: As you might expect, we characterize the notion of "likely" and "not likely" in terms of probability. To do this, we "pretend" that H_0 is true and assess the probability of seeing the $\overline{x}$ value we observed in our particular sample.

- If this probability is small, then we reject H_0.
- If this probability is not small, then we do not reject H_0.

IN GENERAL: Consider the general situation where we desire to test

$$H_0 : \mu = \mu_0$$
versus
$$H_1 : \mu \neq \mu_0,$$

where μ_0 is the value of interest in the experiment or observational study. If we assume that H_0 is true, then $\mu = \mu_0$. Recall that when $x_1, x_2, ..., x_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ population, we know that **when H_0 is true**, the quantity

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Hence, if the null hypothesis is true, z follows a standard normal distribution. The quantity z is sometimes called a **one-sample z statistic**.

INTUITIVELY: A "likely value" of $\overline{x}$ is one for which $\overline{x}$ is close to μ_0, or equivalently, one for which the value of the z statistic is close to zero. That is, if H_0 is true, we would expect to see

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \text{ close to 0.}$$

On the other hand, an unlikely value of $\overline{x}$ would be one for which $\overline{x}$ is not close to μ_0, or equivalently, one for which the z statistic is far away from zero (**in either direction**).

STRATEGY: To formalize the notion of "unlikely," suppose that we decide to reject H_0 if the probability of seeing the value of $\overline{x}$ that we saw is less than some small value α, say, $\alpha = 0.05$. That is, for probabilities smaller than α, we feel that our sample **evidence** is strong enough to reject H_0.

REALIZATION: Values of the z statistic that are larger than $z_{\alpha/2}$ or smaller than $-z_{\alpha/2}$ (i.e., in the shaded regions in Figure 6.50) are unlikely in the sense that the chance of

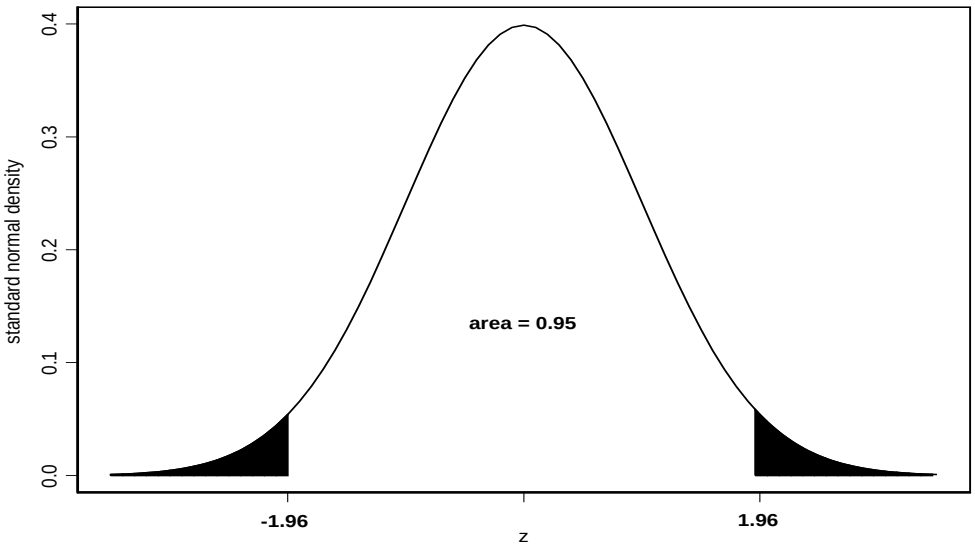


Figure 6.50: *Standard normal density shaded region equal to* $\alpha = 0.05$*. If the value of* z *falls in the shaded region, our sample evidence against* $H_0 : \mu = \mu_0$ *is strong.*

seeing them is less than α, the cut-off probability for "unlikeliness" that we have specified. In Figure 6.50, the level of $\alpha = 0.05$.

TERMINOLOGY: We call the shaded areas in Figure 6.50 the **rejection region** for the hypothesis test of $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.

- If the value of the z statistic falls in the rejection region, then **we reject** H_0. That is, the evidence in our sample shows that the null hypothesized value of the mean, μ_0, is **not** a likely value for the parameter μ.
- If the z statistic does not fall in one of the shaded regions, then we **do not reject** H_0. That is, there is not enough evidence to refute the conjecture that $\mu = \mu_0$.
- In either case, note that we are making a statement about H_0. We either have sufficient evidence to reject it, or we do not.

Example 6.6. Using the rat-weight gain data from Example 6.5, we wish to test, at level $\alpha = 0.05$,

$$H_0 : \mu = 27.8$$

versus

$$H_1 : \mu \neq 27.8,$$

with our $n = 5$ experimental values. Recall that our guess for $\sigma \approx 14$. The z statistic is equal to

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{41.0 - 27.8}{14/\sqrt{5}} = 2.11.$$

Taking $\alpha = 0.05$, we have that $z_{0.025} = 1.96$. Thus, the rejection region for the test is values of z that are larger than 1.96 or smaller than -1.96 (see Figure 6.50).

CONCLUSION: Comparing the z statistic 2.11 to the **critical value** $z_{0.025} = 1.96$, we see that $2.11 > 1.96$ (see Figure 6.50). Hence, we reject H_0 at the $\alpha = 0.05$ level since the z statistic falls in the rejection region. The evidence in our sample is strong enough to discount $\mu_0 = 27.8$ as plausible; i.e., to conclude that *vitamin A does have an effect on the mean weight gain.*

TERMINOLOGY: The z statistic

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

is an example of a **test statistic**. A test statistic is computed from the sample and is used as a basis for deciding between H_0 and H_1. *If the test statistic falls in the α-level rejection region for a test, we reject H_0 at level α.*

6.3.1 The significance level

REALIZATION: In the hypothesis test of $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$, we used a cut-off probability level of α. We realized that this probability determined the size of the rejection region. The value α is called the **significance level** for the test. Formally, because we perform the test assuming that H_0 is true, α denotes the probability of rejecting a true H_0.

REJECTING H_0: When do we reject H_0? There are two scenarios:

(i) H_0 really is **not true**, and this caused the unlikely value of z that we saw.

(ii) H_0 is, in fact, **true**, but it turned out that we ended up with an "unusual" sample that caused us to reject H_0 nonetheless.

TYPE I ERROR: The situation in (ii) is a error. That is, we have made an incorrect judgement between H_0 and H_1. Unfortunately, because we are dealing with chance mechanisms in the sampling procedure, it is possible to make this type of mistake. A mistake like that in (ii) is called a **Type I Error**. The probability of committing a Type I Error is equal to α, the **significance level** for the test.

TERMINOLOGY: When we reject H_0, we say formally that we "reject H_0 at the level of significance α," or that "we have a statistically significant result at level α." Thus, we have clearly stated what criterion we have used to determine what is "unlikely." If we do not state the level of significance, other people have no sense of how stringent or lenient we were in our determination! An observed value of the test statistic leading to the rejection of H_0 is said to be **statistically significant** at level α.

6.3.2 One and two-sided tests

TWO-SIDED TESTS: For the rat weight gain study in Example 6.5, we considered the following hypotheses:

$$H_0 : \mu = 27.8$$

versus

$$H_1 : \mu \neq 27.8.$$

The alternative hypothesis here is an example of a **two-sided alternative**; hence, this is called a **two-sided test**. The reason for this is that the alternative simply specifies a deviation from the null hypothesized value μ_0 but does not specify the direction of

that deviation. Values of the z statistic far enough away from 0 in either direction will ultimately lead to the rejection of H_0.

ONE-SIDED TESTS: In some applications, the researcher may be interested in testing that the population mean μ is **no larger** or **no smaller** than a certain pre-specified value. For example, suppose that we are hopeful that vitamin A not only has some sort of effect on weight gain, but, in fact, it causes rats to **gain** more weight than they would if they were untreated. In this instance, we would be interested in the following test:

$$H_0 : \mu = 27.8$$
versus
$$H_1 : \mu > 27.8.$$

That is, it might be of more interest to specify a different type of alternative hypothesis. As we now see, the principles underlying the approach are similar to those in the two-sided case, but the procedure is modified slightly to accommodate the particular direction of a departure from H_0 in which we are interested (here, that the population mean is larger than 27.8). The alternative hypothesis $H_1 : \mu > 27.8$ is called a **one-sided alternative** and the test above is called a **one-sided test**.

INTUITION: If H_1 is really true, we would expect to see a value of $\overline{x}$ larger than 27.8, or, equivalently, a value of

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

larger than 0. With a two-sided test, we only cared about the z statistic being large in magnitude, regardless of direction. But, if our interest is in this one-sided alternative, *we now care about the direction.*

IN GENERAL: Just as before, consider the general set of hypotheses:

$$H_0 : \mu = \mu_0$$
versus
$$H_1 : \mu > \mu_0.$$

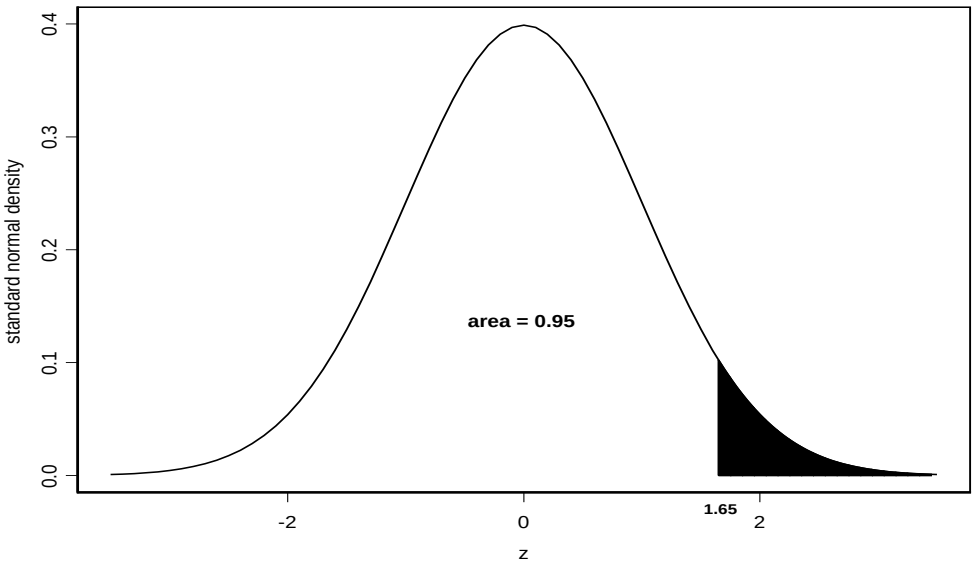


Figure 6.51: *Standard normal density with shaded area equal to* $\alpha = 0.05$. *This is the rejection region for a one-sided upper tail test. If the value of* z *falls in the shaded region, our sample evidence against* H_0 *is strong.*

If we assume that the null hypothesis H_0 is true (i.e., $\mu = \mu_0$), then the one-sample z statistic

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

still follows a standard normal distribution. However, we are now interested only in the situation where $\overline{x}$ is **significantly larger** than μ_0; i.e., where the z statistic is significantly larger than 0. We know that there is a value z_α such that

$$P(z > z_\alpha) = \alpha.$$

Graphically, the shaded area in Figure 6.51 has area (probability) equal to $\alpha = 0.05$. Note here that the entire probability α is concentrated in the **right tail**, because we are only interested in large, positive values of the z statistic. With $\alpha = 0.05$, values of the z statistic greater than $z_{0.05} = 1.65$ are "unlikely" in this sense.

Example 6.7. Using the rat-weight gain data from Example 6.5, we wish to test, at level $\alpha = 0.05$,

$$H_0 : \mu = 27.8$$
versus
$$H_1 : \mu > 27.8.$$

The test statistic is the same as before; i.e.,

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{41.0 - 27.8}{14/\sqrt{5}} = 2.11.$$

CONCLUSION: From Table A (MM), with $\alpha = 0.05$, we have $z_{0.05} = 1.65$ (see also Figure 6.51). Comparing the value of the z statistic to this tabled value, we see that $2.11 > 1.65$. We thus reject H_0 at the $\alpha = 0.05$ level; that is, the evidence in the sample is strong enough to conclude that the mean weight gain is greater than $\mu_0 = 27.8$ mg; i.e., *that vitamin A increases weight gain on average.*

QUESTION: Is there enough evidence to reject H_0 at the $\alpha = 0.01$ level? at the $\alpha = 0.001$ level? Why or why not? Note that $z_{0.01} = 2.33$ and $z_{0.001} = 3.08$ (using Table A).

6.3.3 A closer look at the rejection region

TWO-SIDED REJECTION REGION: The test of hypotheses of the form

$$H_0 : \mu = \mu_0$$
versus
$$H_1 : \mu \neq \mu_0$$

is a **two-sided test**. The alternative hypothesis specifies that μ is different from μ_0 but may be on either side of it. With tests of this form, we know that the rejection region is located in **both tails** of the standard normal distribution, with total shaded area equal to α (the significance level for the test).

ONE-SIDED REJECTION REGION: Similarly, tests of hypotheses of the form

$$H_0 : \mu = \mu_0$$
versus
$$H_1 : \mu > \mu_0$$

and

$$H_0 : \mu = \mu_0$$
versus
$$H_1 : \mu < \mu_0$$

are **one-sided** hypothesis tests. The alternative in which we are interested lies to one side of the null hypothesized value. For the first test, the rejection region is located in the **upper tail** of the standard normal distribution. For the second test, the rejection region is located in the **lower tail** of the standard normal distribution. In either case, the area of the rejection region equals α, the significance level for the test.

6.3.4 Choosing the significance level

CURIOSITY: How does one decide on an appropriate value for α? Recall that we mentioned a particular type of mistake that we might make, that of a **Type I Error**. Because we perform the hypothesis test under the assumption that H_0 is true, this means that the probability we reject a true H_0 is equal to α. Choosing α, thus, has to do with how serious a mistake a Type I Error might be in the particular application.

Example 6.8. Suppose the question of interest concerns the efficacy of a costly new drug for the treatment of advanced glaucoma in humans, and the new drug has potentially dangerous side effects (e.g., permanent loss of sight, etc.). Suppose a study is conducted where sufferers of advanced glaucoma are randomly assigned to receive either the standard treatment or the new drug (a **clinical trial**), and suppose the random variable of interest is the period of prolongation of sight. It is known that the standard drug prolongs sight

for μ_0 months. We hope that the new drug is more effective in the sense that it increases the prolongation of sight, in which case it may be worth its additional expense and risk of side effects. We thus would consider testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$, where μ denotes the mean prolongation period under treatment of the **new** drug.

SCENARIO: Suppose that after analyzing the data, unbeknownst to us, our sample of patients leads us to commit a Type I Error; that is, we reject a true H_0 and claim that the new drug is more effective than the standard drug when, in reality, it isn't. Because the new drug is so costly and carries the possibility of serious side effects, this could be a very serious mistake. In this instance, patients would be paying more with the risk of dangerous side effects for no real gain over the standard treatment. In a situation like this, it is intuitively clear that we would like α to be very small, so that the chance of claiming a new treatment is superior, when it really isn't, is small. That is, we would like to be **conservative**. In situations where the consequences of a Type I Error are not so serious, there is no reason to take α to be so small. That is, we might choose a more **anticonservative** significance level. This is often done in preliminary investigations.

TYPE II ERROR: Committing a Type I Error is not the only type of mistake that we can make in a hypothesis test. The sample data might be "unusual" in such a way that we end up not rejecting H_0 when H_1 is really true. This type of mistake is called a **Type II Error**. Because a Type II Error is also a mistake, we would like the probability of committing such an error, say, β, to also be small. In many situations, committing a Type II Error is not as serious as committing a Type II Error. In Example 6.8, if we commit a Type II Error, we infer that the new drug is not effective when it really is. Although this, too, is undesirable, as we are discarding a potentially better treatment, we are no worse off than before we conducted the test, whereas if we commit a Type I Error, we will unduly expose patients to unnecessary costs and risks for no gain.

SUMMARY:

- **Type I Error**: Reject H_0 when it is true.
- **Type II Error**: Not rejecting H_0 when H_1 is true.

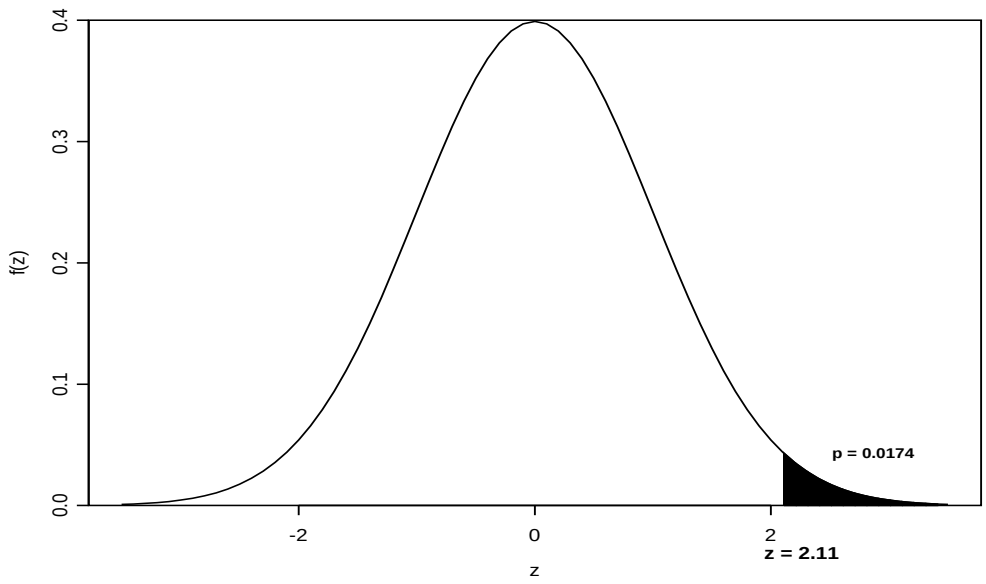


Figure 6.52: *Standard normal density with the area to the right of* $z = 2.11$. *The shaded area equals* 0.0174, *the probability value for the one-sided test in Example* 6.7.

6.3.5 Probability values

INVESTIGATION: In Example 6.7, we considered the one-sided test

$$H_0 : \mu = 27.8$$
versus
$$H_1 : \mu > 27.8.$$

Recall that our one-sample z statistic was $z = 2.11$. We saw that when $\alpha = 0.05$, we rejected H_0 since $z > z_{0.05} = 1.65$. However, when $\alpha = 0.01$, we do not reject H_0 since $z < z_{0.01} = 2.33$.

REALIZATION: From the above statements, we know that there must exist some α **between** 0.01 and 0.05 where our test statistic $z = 2.11$ and critical value z_α, will be the same. The value of α where this occurs is called the probability value for the test.

TERMINOLOGY: The smallest value of α for which H_0 is rejected is called the **probability value** of the test. We often abbreviate this as P **value**.

CALCULATION: In Example 6.7, the probability value for the test is the area to the **right** of $z = 2.11$ under the standard normal distribution; i.e.,

$$P(z > 2.11) = 0.0174.$$

Note that how we compute the probability value is consistent with the alternative hypothesis $H_1 : \mu > 27.8$. We reject H_0 for large positive values of z; thus, the probability value is the area in the **right** tail of the distribution; see Figure 6.52.

MAIN POINT: *In any hypothesis test, we can always make our decision by comparing the probability value to our significance level* α. In particular,

- If the probability value is smaller than α, we reject H_0.
- If the probability value is not smaller than α, we do not reject H_0.

RULES FOR COMPUTING PROBABILITY VALUES: If we have a **one-sided** hypothesis test of the form

$$H_0 : \mu = \mu_0$$
versus
$$H_1 : \mu > \mu_0,$$

the probability value for the test is given by the **area to the right** of z under the standard normal distribution. If we have a **one-sided** hypothesis test of the form

$$H_0 : \mu = \mu_0$$
versus
$$H_1 : \mu < \mu_0,$$

the probability value for the test is given by the **area to the left** of z under the standard normal distribution. If we have a **two-sided** hypothesis test

$$H_0 : \mu = \mu_0$$
versus
$$H_1 : \mu \neq \mu_0,$$

the probability value for the test is given by the **area to the right** of $|z|$ plus the **area to the left** of $-|z|$ under the standard normal distribution.

IN ANY CASE: If the probability value is smaller than α, we reject H_0.

Example 6.8. It is thought that the body temperature of intertidal crabs exposed to air is less than the ambient temperature. Body temperatures were obtained from a random sample of $n = 8$ such crabs exposed to an ambient temperature of 25.4 degrees C.

25.8	24.6	26.1	24.9	25.1	25.3	24.0	24.5

Assume that the body temperatures are approximately normally distributed and let μ denote the mean body temperature for the population of intertidal crabs exposed to an ambient temperature of 25.4 degrees C. Then, we wish to test, say, at the conservative $\alpha = 0.01$ level, $H_0 : \mu = 25.4$ versus $H_1 : \mu < 25.4$.

ANALYSIS: From past experience, a reliable guess of $\sigma \approx 0.7$ degrees C. Simple calculations show that $\overline{x} = 25.0$ so that the one-sample z statistic is

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{25.0 - 25.4}{0.7/\sqrt{8}} = -1.62.$$

To compute the P value, note that $P(z < -1.62) = 0.0526$, which is not less than 0.01. Thus, we do not have enough evidence against H_0. That is, there is not enough evidence in the sample, at the $\alpha = 0.01$ level of significance, to suggest that the mean body temperature of intertidal crabs exposed to air at 25.4 degrees C is, indeed, less than 25.4.

6.3.6 Decision rules in hypothesis testing

SUMMARY: Summarizing everything we have talked about so far, in a hypothesis testing situation, we reject H_0 if

1. the test statistic z falls in the rejection region for the test. This is called the **rejection region approach** to testing.
2. the probability value for the test is smaller than α. This is called the **probability value approach** to testing.

These two **decision rules** are equivalent.

- With the rejection region approach, we think about the value of the test statistic. If it is "large," then it is an "unlikely" value under the null hypothesis. Of course, "large" depends on the probability α we have chose to define "unlikely."
- With the probability value approach, we think directly about the probability of seeing something as weird or weirder than what we saw in our experiment. If this probability is "small" (i.e., smaller than α), then the test statistic we saw was "unlikely."
- A "large" test statistic and a "small" probability value are equivalent.
- An advantage to working with probability values is that we calculate the probability of seeing what we saw; this is useful for thinking about just how "strong" the evidence in the data really is. Smaller probability values mean "more unlikely." Hence, since we compute the probability value under the assumption that H_0 is true, it should be clear that the smaller probability is, the more unlikely it is that H_0 is true. That is, the smaller the probability value, the more evidence against H_0. Large probability values are not evidence against H_0.

6.3.7 The general outline of a hypothesis test

SUMMARY: We now summarize the steps in conducting a hypothesis test for a single population mean μ (when σ is known). The same principles that we discuss here can be generalized to other testing situations.

1. *Determine the question of interest.* This is the first and foremost issue. No experiment should be conducted unless the scientific questions are well-formulated.

2. *Express the question of interest in terms of two hypotheses involving the population mean μ.*

   - H_0: the **null hypothesis** − the hypothesis of **no effect**. If μ_0 is a specified value, $H_0 : \mu = \mu_0$.
   - H_1: the **alternative hypothesis** − the condition that we suspect or hope is true. Depending on the nature of the problem, H_1 will be two-sided, i.e., $H_1 : \mu \neq \mu_0$, or one-sided; i.e., $H_1 : \mu > \mu_0$ or $H_1 : \mu < \mu_0$.

3. *Choose the significance level*, α, to be a small value, like 0.05. The particular situation (i.e., the severity of committing a Type I Error) will dictate its value.

4. *Conduct the experiment, collect the data, and calculate the test statistic.*

5. *Perform the hypothesis test*, either rejecting or not rejecting H_0 in favor of H_1. This can be done two different ways, as we have seen:

   - **The rejection region approach**. Reject the null hypothesis if the test statistic falls in the rejection region for the test.
   - **The probability value approach**. Reject the null hypothesis if the probability value for the test is smaller than α.

Example 6.9. In a certain region, a forester wants to determine if μ, the mean age of a certain species of tree, is significantly different than 35 years. Using a carbon dating procedure, the following ages were observed for a simple random sample of $n = 16$ trees:

```
32.8   25.2   39.7   30.8   30.5   26.7   20.9   17.6
37.9   28.6   13.8   42.8   36.1   29.4   21.0   18.7
```

From previous studies a reliable guess of the population standard deviation is $\sigma = 5$ years. We would like to test, at the $\alpha = 0.10$ level, $H_0 : \mu = 32$ versus $H_1 : \mu \neq 32$ (a two-sided alternative). To perform this test, I used **Minitab**; here is the output:

```
Test of mu = 32 vs not = 32
The assumed standard deviation = 5

Variable   N     Mean   StDev  SE Mean        90% CI             Z      P
years     16  28.2813  8.4181   1.2500  (26.2252, 30.3373)  -2.98  0.003
```

CONCLUSION: Because the probability value is small; i.e., smaller than $\alpha = 0.10$, we have sufficient evidence to reject H_0 using our sample results. That is, our data suggest that the mean age of the trees under investigation is different than 32 years. *As a side note, observe that the* 90 *percent confidence interval for* μ *does not include* 32*, the null-hypothesized value of* μ*!!*

6.3.8 Relationship with confidence intervals

EQUIVALENCE: There is an elegant duality between hypothesis tests and confidence intervals. We have already seen that a confidence interval for a single population mean is based on the probability statement

$$P\left(-z_{\alpha/2} \leq \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

As we have seen, a **two-sided** hypothesis test is based on a probability statement of the form

$$P\left(\left|\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}\right| \geq z_{\alpha/2}\right) = \alpha.$$

Comparing the two statements, a little algebra shows that they are actually the same! Thus, choosing a "small" level of significance level α in a hypothesis test is equivalent to choosing a "large" confidence level $1-\alpha$ for a confidence interval. Furthermore, with the same choice of α, we may

- reject $H_0 : \mu = \mu_0$ in favor of the two-sided alternative, at the significance α, if the $100(1-\alpha)$ percent confidence interval does **not** include μ_0.

Thus, the experimenter can perform a **two-sided test** at level α just by looking at whether or not the null hypothesized value μ_0 falls in the $100(1-\alpha)$ percent confidence interval. *We can not use our confidence intervals to conduct one-sided tests.*

OBSERVATION: In Example 6.9, we see that $\mu_0 = 32$ is not contained in the 90 percent confidence interval for μ. Thus, we may reject $H_0 : \mu = 32$, in favor of the two-sided alternative, at the $\alpha = 0.10$ level.

6.4 Some general comments on hypothesis testing

Although we have only discussed hypothesis tests for a single population mean μ, the ideas outlined below will hold in many other hypothesis testing situations.

- *We do not even begin to collect data until the question of interest has been established.* There are good reasons for doing this. For one, it makes the experimenter specify, up front, the goal of the experiment and research question of interest. Also, if one were to start performing the experiment and then pose the question of interest, preliminary experimental results may sway the researcher into making biased judgements about the real question at hand. For example, if one starts off with a two-sided alternative hypothesis, but changes his mind after seeing some data to use a one-sided alternative, this may introduce bias into the experiment and the data not yet collected.

- Refrain from using statements like "Accept H_0" and "H_0 is true." These statements can be grossly misleading. If we do reject H_0, we are saying that the sample evidence is sufficiently strong to suggest that H_0 is *probably* not true. On the other hand, if we do not reject H_0, we do not because the sample does not contain enough evidence against H_0. Hypothesis tests are set up so that we assume H_0 to be true and then try to refute it based on the experimental data. If we can not refute H_0, this doesn't mean that H_0 is true, only that we couldn't reject it.

- To illustrate the preceding remark, consider an Amercian courtroom trial analogy. The defendant, before the trial starts, is assumed to be "not guilty." This is the null hypothesis. The alternative hypothesis is that the defendant is "guilty." This can be viewed as a hypothesis test:

$$H_0 : \text{defendant is not guilty}$$
$$\text{versus}$$
$$H_1 : \text{defendant is guilty.}$$

  Based on the testimony (data), we make a decision between H_0 and H_1. If the evidence is overwhelming (beyond a reasonable doubt) against the defendant, we reject H_0 in favor of H_1 and classify the defendant as "guilty." If the evidence presented in the case is not beyond a reasonable doubt against the defendant, we stick with our original assumption that the defendant is "not guilty." This does not mean that the defendant is innocent! It only means that there wasn't enough evidence to sway our opinion from "not-guilty" to "guilty." By the way, what would Type I and Type II Errors be in this analogy?

- The significance level and rejection region are not cast in stone. The results of hypothesis tests should **not** be viewed with absolute yes/no interpretation, but rather as guidelines for aiding us in interpreting experimental results and deciding what to do next.

- The assumption of normality may be a bad assumption. If the original population distribution is severely non-normal, this could affect the results since the one sample

z statistic is no longer exactly standard normal. As long as the departure from normality is not great, we should be fine.

- It has become popular in reports and journals in many applied disciplines (which routinely use these statistical tests) to set $\alpha = 0.05$ regardless of the problem at hand, then strive rigorously to "find a probability value less than 0.05." From a practical standpoint, there probably isn't a lot of difference between a probability value of 0.049 and a probability value of 0.051. Thus, probability values need to interpreted with these cautionary remarks in mind.

- Many statistical packages, such as SAS and Minitab, report only probability values. This way, a test at any level of significance level (α level) may be performed by comparing the probability value reported in the output to the pre-specified α.

- Changing the value of α after the experiment has been performed or during the experiment is a terrible misuse of a statistics. For example, if, before the experiment is performed, the researcher took $\alpha = 0.01$ (conservative), but after getting a probability value of 0.04, the researcher changes to $\alpha = 0.05$ just to get a statistically significant result. Theoretical arguments can show that this practice pretty much negates all information conveyed in the test, and hence, the experiment itself.

A LOOK AHEAD: We have spent a great deal of time learning the mechanics of hypothesis tests for a single population mean μ (in the situation where σ is known). In introducing hypothesis tests for a single population mean μ, we have discussed many ideas (e.g., significance level, probability values, rejection region, null and alternative hypotheses, one and two-sided tests, Type I and II errors, philosophies of testing, relationship with confidence intervals, etc.). The basic premise behind these underlying ideas will be **the same** for the rest of the course, regardless of what population parameter(s) we are interested in. The theory and philosophy we have discussed will carry over to other hypothesis testing situations.

OMISSION: For now, we will skip Section 6.4 in MM.

7 Inference for Distributions

7.1 Introduction

In the last chapter, we learned that hypothesis tests and confidence intervals for the **population mean** μ, based on an SRS, were constructed using the fact that

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

You will recall that a $100(1-\alpha)$ percent confidence interval for μ was given by

$$\left(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

and the one-sample z statistic used to test $H_0 : \mu = \mu_0$ was given by

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}.$$

For these procedures to make sense, we required that the population standard deviation σ was **known**.

REALITY: In most applications, σ will not be known. Thus, if we want to write a confidence interval or perform a hypothesis test for the population mean μ, we should not use the methods outlined in the previous chapter.

SOLUTION: When σ is not known, however, we can do the almost obvious thing; namely, we can use the sample standard deviation s as an **estimate** for σ. Recall that

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Using s as an estimate of σ we can create the following quantity:

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}.$$

However, unlike z, t does not have a $\mathcal{N}(0,1)$ distribution!! This follows because we are using an estimate s for the population parameter σ. Put another way, z and t have different **sampling distributions**. Because of this, we need to become comfortable with a new density curve.

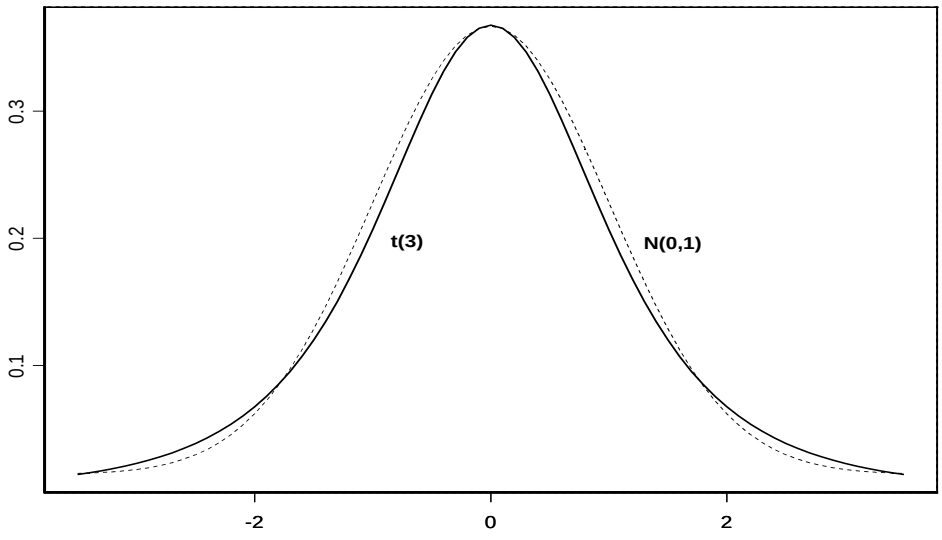


Figure 7.53: *The t_3 distribution* (solid) *and the $\mathcal{N}(0,1)$ distribution* (dotted).

7.2 One-sample t procedures

BIG RESULT: Suppose that an SRS of size n is drawn from a $\mathcal{N}(\mu, \sigma)$ population distribution. Then, the random variable

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

has a t **distribution** with $n-1$ degrees of freedom. This density curve is denoted by t_{n-1}. Figure 7.53 displays the t_3 density curve along with the $\mathcal{N}(0,1)$ density curve.

FACTS ABOUT THE t DISTRIBUTION:

- All t distributions are continuous and **symmetric** about 0.
- The t family is indexed by a degree of freedom value k (an integer).
- As k increases, the t_k distribution "approaches" the standard normal distribution. When $k > 30$, the two distributions are nearly identical.

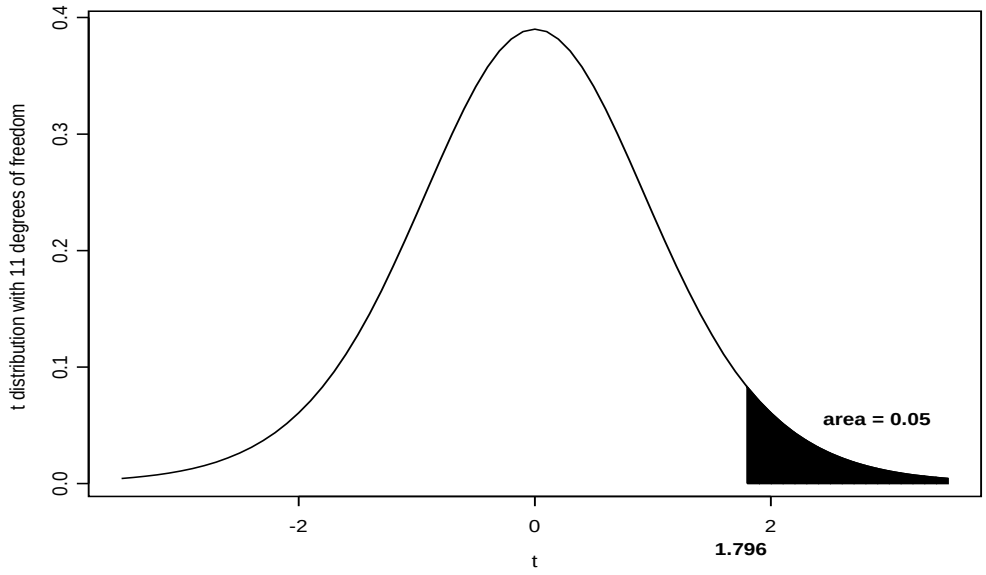


Figure 7.54: *The t_{11} density curve and 95th percentile $t_{11,0.05} = 1.796$.*

- When compared to the standard normal distribution, the t distribution, in general, is less peaked, and has more mass in the tails.
- Table D (MM) lists probabilities and percentiles for the t distributions. Of course, statistical software packages should be used in practice.

NOTATION: The **upper α percentile** of a t_k distribution is the value $t_{k,\alpha}$ which satisfies

$$P(t < -t_{k,\alpha}) = P(t > t_{k,\alpha}) = \alpha.$$

In Figure 7.54, we see that the 95th percentile of the t_{11} distribution is $t_{11,0.05} = 1.796$ (see also Table D). By symmetry, the 5th percentile is $-t_{11,0.05} = -1.796$.

QUESTIONS:

- What is the 99th percentile of the t_{11} distribution? the 1st percentile?
- What is the 90th percentile of the t_{19} distribution? the 10th percentile?
- What is the 92nd percentile of the t_6 distribution? the 8th percentile?

7.2.1 One-sample t confidence intervals

ONE SAMPLE t CONFIDENCE INTERVAL: Suppose that an SRS is drawn from a $\mathcal{N}(\mu, \sigma)$ population. A $100(1-\alpha)$ percent **confidence interval** for μ is given by

$$\left(\overline{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}},\ \overline{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right).$$

This interval is exact when the population distribution is normal and is approximately correct for large n in other cases.

REMARKS: Comparing the t interval to the z interval, we see that they are identical in form; that is, each interval is of the form $\overline{x} \pm m$, where m denotes the **margin of error**. For the t interval, the margin of error is

$$m = t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}.$$

For the z interval, the margin of error was

$$m = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

REMINDER: When the population standard deviation σ is not known, we use the t interval!

TERMINOLOGY: In the t confidence interval, the margin of error can be written as

$$m = \underbrace{t_{n-1,\alpha/2}}_{t \text{ percentile from Table D}} \times \underbrace{\frac{s}{\sqrt{n}}}_{\text{standard error}}.$$

The quantity $s/\sqrt{n}$ is called the **standard error** of the estimate $\overline{x}$. To see where this comes from, recall that from a random sample of normally distributed data, the statistic

$$\overline{x} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Thus, the standard error $s/\sqrt{n}$ is an estimate of the standard deviation of $\overline{x}$. The standard error of an estimate (such as $\overline{x}$) is an important quantity. It summarizes numerically the amount of **variation** in an estimate. *Standard errors of estimates are commonly reported in statistical analysis.*

Example 7.1. In an agricultural experiment, a random sample of $n = 10$ plots produces the following yields (measured in kg per plot). The plots were treated identically in the planting, growing, and harvest phases. The goal is to obtain a 95 percent confidence interval for μ, the population mean yield. Here are the sample yields:

```
23.2   20.1   18.8   19.3   24.6   27.1   33.7   24.7   32.4   17.3
```

From these data, we compute $\overline{x} = 24.1$ and $s = 5.6$. Also, with $n = 10$, the degrees of freedom is $n - 1 = 9$, and $t_{n-1,\alpha/2} = t_{9,0.025} = 2.262$ (Table D). The 95 percent confidence interval is

$$\left(24.1 - 2.262 \times \frac{5.6}{\sqrt{10}},\ 24.1 + 2.262 \times \frac{5.6}{\sqrt{10}}\right) \quad \text{or} \quad (20.1, 28.1) \text{ kg/plot}$$

Thus, based on these data, we are 95 percent confident that the population mean yield μ is between 20.1 and 28.1 kg/plot.

EXERCISE: Using the data in Example 7.1, find a 90 percent confidence interval for μ. Also, find a 99 percent confidence interval.

7.2.2 One-sample t tests

RECALL: In that last chapter, we saw that the one-sample z statistic, used to test $H_0 : \mu = \mu_0$, was given by

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}.$$

Like the one-sample z confidence interval, we required that σ be **known** in advance. When σ is not known in advance, we can use a **one-sample t test**.

ONE-SAMPLE t TEST: Suppose that an SRS is drawn from a $\mathcal{N}(\mu, \sigma)$ population, where both μ and σ are unknown. To test $H_0 : \mu = \mu_0$, we use the one-sample t statistic

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}.$$

When H_0 is true, the **one-sample t statistic** varies according to a t_{n-1} sampling distribution. Thus, rejection regions are located in the tails of this density curve.

TWO-SIDED REJECTION REGION: The **two-sided test**

$$H_0 : \mu = \mu_0$$
versus
$$H_1 : \mu \neq \mu_0$$

has a rejection region located in both tails of the t_{n-1} distribution. The total area of the rejection region equals α (the significance level for the test). By convention, each tail has area $\alpha/2$.

ONE-SIDED REJECTION REGION: **One-sided tests** of the form

$$H_0 : \mu = \mu_0$$
versus
$$H_1 : \mu > \mu_0$$

and

$$H_0 : \mu = \mu_0$$
versus
$$H_1 : \mu < \mu_0$$

have rejection regions located in only one tail of the t_{n-1} distribution. The test with $H_1 : \mu > \mu_0$ has its rejection region in the **upper tail**; the test with $H_1 : \mu < \mu_0$ has rejection region in the **lower tail**. The total area of the rejection region equals α.

NOTE: **Probability values** are computed in the same manner they were before (in the last chapter). The only difference is that now we are finding areas under the t_{n-1} density curve (instead of finding areas under the $\mathcal{N}(0,1)$ density curve). To be more explicit,

- $H_1 : \mu > \mu_0$: Probability value = area to the right of t
- $H_1 : \mu < \mu_0$: Probability value = area to the left of t
- $H_1 : \mu \neq \mu_0$: Probability value = twice that of a one-sided test.

Example 7.2. Starting salaries for a school's recent graduates is an important factor when assessing the quality of the program. At a particular college, a random sample of $n = 15$ students was taken; here were the starting salaries (in thousands of dollars) for those students in the sample:

```
37.1   49.8   36.5   26.3   40.1   18.3   45.2   51.3   39.3   49.9
                44.1   64.3   39.3   41.0   56.3
```

The dean of the college, in an attempt to boost the ratings of his program, claims that "his school's average starting salary exceeds \$40,000." Based on these data, is there evidence to support his claim?

ANALYSIS: Statistically, we are interested in testing

$$H_0 : \mu = 40$$

versus

$$H_1 : \mu > 40,$$

where μ denotes the population mean starting income for new graduates in this school's program. To conduct the test, we will use $\alpha = 0.05$. Here is the Minitab output:

```
Test of mu = 40 vs > 40
Variable   N     Mean    StDev  SE Mean      T      P
salary    15  42.5883  11.3433   2.9288   0.88  0.196
```

The t statistic provided above is computed by

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{42.59 - 40}{11.34/\sqrt{15}} \approx 0.88.$$

The probability value is computed by finding the area to the **right** of $t = 0.88$ under the t_{14} density curve (see Figure 7.55). Note that $P = 0.196$ is much larger than $\alpha = 0.05$. This is not a significant result (i.e., we do not have enough evidence to reject H_0).

CONCLUSION: At the $\alpha = 0.05$ level, we do not have evidence to conclude that the mean starting income is larger than $40,000$ dollars per year. That is, we cannot support the dean's assertion.

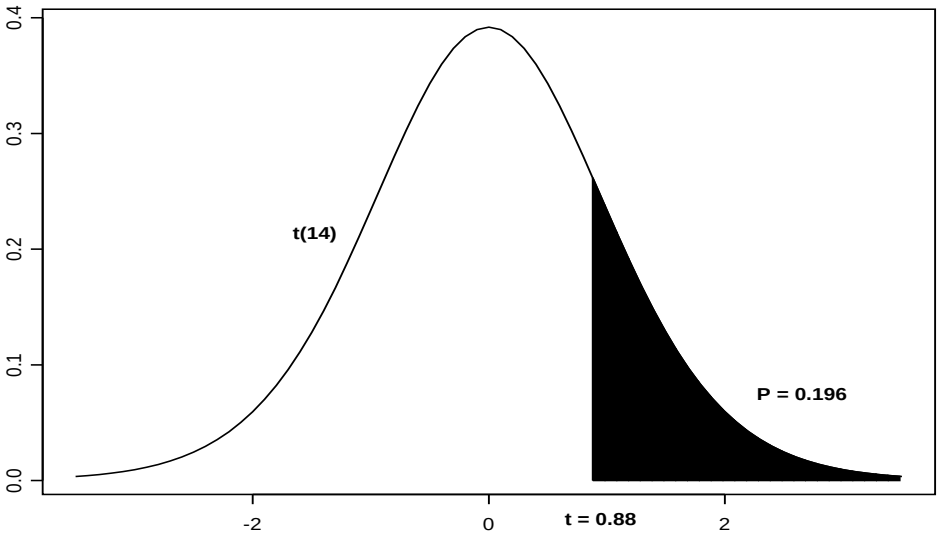


Figure 7.55: *The t_{14} density curve, test statistic, and probability value for Example* 7.2.

Example 7.3. A rental car company is monitoring the drop-off/pick-up times at the Kansas City International Airport. The times listed below (measured in minutes) are the round-trip times to circle the airport (passengers are dropped off and picked up at two terminals during the round trip). Management believes that the average time to circle the airport is about 12.5 minutes. To test this claim, management collects a sample of $n = 20$ times selected at random over a one-week period. Here are the data:

19.7	14.7	19.2	15.1	10.4	12.1	13.1	13.1	19.6	33.3
13.7	10.2	10.6	19.1	10.8	21.0	24.0	26.0	12.3	14.4

ANALYSIS: Since it is unknown whether or not the mean is larger or smaller than 12.5 minutes, management decides to test $H_0 : \mu = 12.5$ versus $H_1 : \mu \neq 12.5$ at the $\alpha = 0.05$ level, where μ denotes the population mean waiting time. Here is the Minitab output:

```
Test of mu = 12.5 vs not = 12.5
Variable   N     Mean   StDev  SE Mean        95% CI           T      P
time      20  16.6291  6.0840   1.3604  (13.7816, 19.4765)  3.04  0.007
```

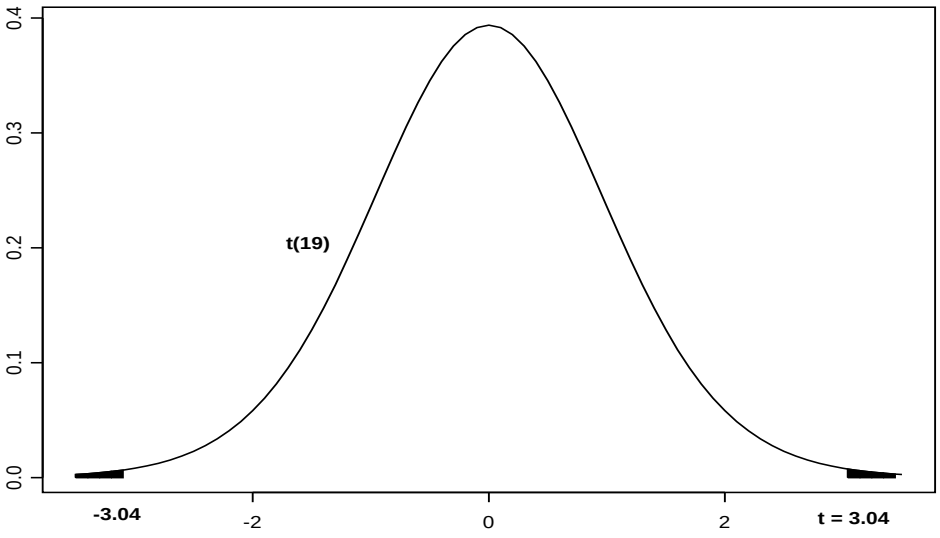


Figure 7.56: *The t_{19} density curve and test statistic for Example* 7.3. *The probability value* (shaded region) *is* $P = 0.007$, *twice the area to the right of* $t = 3.04$.

Note that we can make our decision in different ways:

- The 95 percent **confidence interval** is $(13.78, 19.48)$ minutes. Note that this does not include the null hypothesized value $\mu_0 = 12.5$ minutes. Thus, we would reject $H_0 : \mu = 12.5$ at the five percent level.
- The **probability value** for the test is $P = 0.007$ (see Figure 7.56). This is much smaller than the significance level $\alpha = 0.05$. Thus, we would reject $H_0 : \mu = 12.5$ at the five percent level.
- The **rejection region** for the test consists of values of t larger than $t_{19,0.025} = 2.093$ and values of t smaller than $-t_{19,0.025} = -2.093$. Since our test statistic t is in the rejection region (i.e., t is larger than 2.093), we would reject $H_0 : \mu = 12.5$ at the five percent level.

CONCLUSION: At the five percent level, we have sufficient evidence to conclude that the mean waiting time is different than 12.5 minutes.

Table 7.10: *Systolic blood pressure data.*

Subject	Before	After	Difference
1	120	128	−8
2	124	131	−7
3	130	131	−1
4	118	127	−9
5	140	132	8
6	128	125	3
7	140	141	−1
8	135	137	−2
9	126	118	8
10	130	132	−2
11	126	129	−3
12	127	135	−8

7.2.3 Matched-pairs t test

RECALL: In Chapter 3, we talked about an experimental design where each subject received two different treatments (in an order determined at random) and provided a response to each treatment; we called this a **matched-pairs design**. We now learn how to analyze data from such designs. Succinctly put, matched-pairs designs are analyzed by looking at the differences in response from the two treatments.

Example 7.4. In Example 3.8, we considered an experiment involving middle-aged men; in particular, we wanted to determine whether or not a certain stimulus (e.g., drug, exercise regimen, etc.) produces an effect on the systolic blood pressure (SBP). Table 7.10 contains the before and after SBP readings for the $n = 12$ middle-aged men in the experiment. We would like to see if these data are statistically significant at the $\alpha = 0.05$ level.

APPROACH: In matched-pairs experiments, to determine whether or not the treatments differ, we use a one-sample t test on the **data differences**.

HYPOTHESIS TEST: To test whether or not there is a difference in the matched-pairs treatment means, we can test

$$H_0 : \mu = 0$$

versus

$$H_1 : \mu \neq 0,$$

where μ denotes the **mean treatment difference**. The null hypothesis says that there is no difference between the mean treatment response, while H_1 specifies that there is, in fact, a difference (without regard to direction). Of course, one-sided tests would be carried out in the obvious way. Also, one could easily construct a one-sample t confidence interval for μ. Even though there are two treatments, we carry out a one-sample analysis because we are looking at the data differences.

MINITAB: For Example 7.4, here is the Minitab output:

```
Test of mu = 0 vs not = 0
Variable     N     Mean   StDev  SE Mean       95% CI             T      P
difference  12  -1.8333  5.8284   1.6825  (-5.5365, 1.8698)  -1.09  0.299
```

ANALYSIS: We have $n = 12$ data differences (one for each man). From the output, we see that $\overline{x} = -1.8333$ and $s = 5.8284$ so that

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{-1.8333 - 0}{5.8284/\sqrt{12}} = -1.09.$$

The probability value is $P = 0.299$; i.e., the area to the left of $t = -1.09$ plus the area to the right of $t = 1.09$ under the t_{11} density curve. Because this probability is not small (i.e., certainly not smaller than $\alpha = 0.05$) we do not reject H_0. See Figure 7.57.

CONFIDENCE INTERVAL: We see that the 95 percent confidence interval for μ, based on this experiment, is $(-5.54, 1.87)$. This interval includes $\mu_0 = 0$.

CONCLUSION: At the five percent level, we do not have evidence to say that the stimulus changes the mean SBP level in this population of middle-aged men.

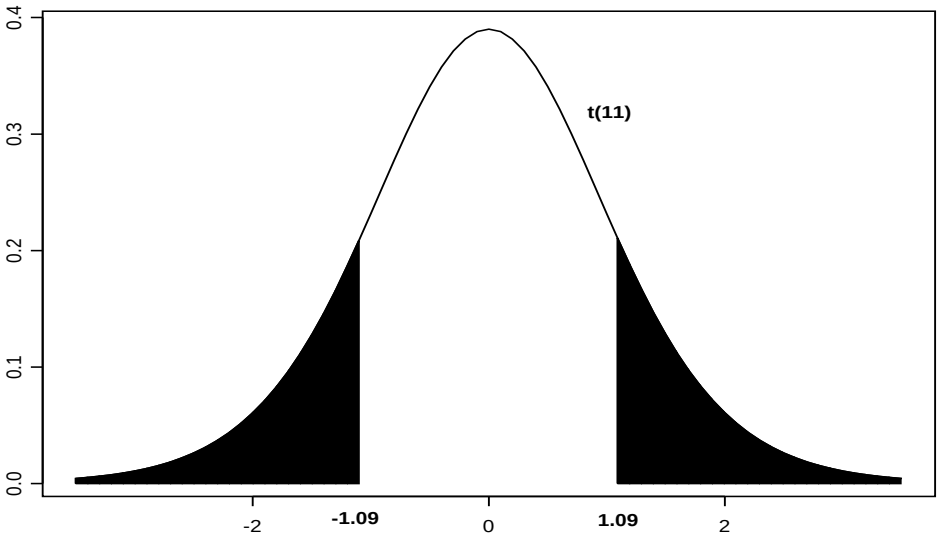


Figure 7.57: *The t_{11} density curve for Example* 7.4. *The probability value* (shaded region) *is twice the area to the right of $t = -1.09$. This probability value equals $P = 0.299$.*

7.3 Robustness of the t procedures

TERMINOLOGY: A statistical inference procedure is called **robust** if the probability calculations are not affected by a departure from the assumptions made.

IMPORTANT: The one-sample t procedures are based on the population distribution being normal. However, these procedures are robust to departures from normality. This means that even if the population distribution (from which we obtain our data) is non-normal, we can still use the t procedures. The text provides the following guidelines:

- $n < 15$: Use t procedures only if the population distribution appears normal and there are no outliers.
- $15 \leq n \leq 40$: Be careful about using t procedures if there is strong skewness and/or outliers present.
- $n > 40$: t procedures should be fine regardless of the population distribution shape.

NONPARAMETRIC TESTS: When normality is in doubt and the sample size is not large, it might be possible to use a **nonparametric** testing procedure. The term "nonparametric" means "distribution-free." The nice thing about nonparametric methods is that they do not require distributional assumptions (such as normality). We will not discuss nonparametric tests in this course. For more information, see Chapter 15 (MM).

7.4 Two-sample t procedures

7.4.1 Introduction

A COMMON PROBLEM: One of the most common statistical problems is that of **comparing** two treatment (e.g., treatment versus control) or **two** stratum **means**. For example, which of two HIV drugs longer delays the onset of AIDS? Are there salary differences between the two genders among university professors? Which type of football (air-filled versus helium filled) travels the farthest? Do rats living in a germ-free environment gain more weight than rats living in an unprotected environment?

How can we answer these questions?

PROCEDURE: Take **two** simple random samples of experimental units (e.g., plants, patients, professors, plots, rats, etc.). Each unit in the first sample receives treatment 1; each in the other receives treatment 2. We would like to make a statement about the difference in responses to the treatments based on this setup.

Example 7.5. Suppose that we wish to compare the effects of two concentrations of a toxic agent on weight loss in rats. We select a random sample of rats from the population and then randomly assign each rat to receive either concentration 1 or concentration 2. The variable of interest is

$$x = \text{weight loss for rat.}$$

Until the rats receive the treatments, we assume them all to have arisen from a common population where $x \sim \mathcal{N}(\mu, \sigma)$. Once the treatments are administered, however, the two samples become **different**. One convenient way to view this is to think of two populations:

$$\begin{aligned} \text{Population 1:} &\quad \text{individuals under treatment 1} \\ \text{Population 2:} &\quad \text{individuals under treatment 2.} \end{aligned}$$

That is, populations 1 and 2 may be thought of as the original population with all possible rats treated with treatment 1 and 2, respectively. We may thus regard our samples as being **randomly selected** from these two populations. Because of the nature of the data, it is further reasonable to think about two random variables x_1 and x_2, one corresponding to each population, and to think of them as being normally distributed:

$$\begin{aligned} \text{Population 1:} &\quad x_1 \sim \mathcal{N}(\mu_1, \sigma_1) \\ \text{Population 2:} &\quad x_2 \sim \mathcal{N}(\mu_2, \sigma_2). \end{aligned}$$

NOTATION: Because we are now thinking about **two independent populations**, we must adjust our notation accordingly so we may talk about the two different random variables and the observations on each of them. Write x_{ij} to denote the jth unit receiving the ith treatment; that is, the jth value observed on the random variable x_i. With this definition, we may thus view our data as follows:

$$\begin{aligned} x_{11}, x_{12}, ..., x_{1n_1} &\quad n_1 = \#\text{ units in sample from population 1} \\ x_{21}, x_{22}, ..., x_{2n_2} &\quad n_2 = \#\text{ units in sample from population 2} \end{aligned}$$

FRAMEWORK: In this framework, we may now cast our question as follows:

difference in mean response for two treatments $\Longrightarrow$ is μ_1 different than μ_2?

More formally, then, we look at the difference $\mu_1 - \mu_2$:

$$\begin{aligned} \mu_1 - \mu_2 = 0 &\quad \Rightarrow \quad \text{there is no difference} \\ \mu_1 - \mu_2 \neq 0 &\quad \Rightarrow \quad \text{there is a difference.} \end{aligned}$$

OBVIOUS STRATEGY: We base our investigation of this population mean difference on the data from the two samples by estimating the difference. In particular, we can compute $\overline{x}_1 - \overline{x}_2$, the **difference in the sample means**.

THEORETICAL RESULTS: It may be shown mathematically that if both population distributions (one for each treatment) are normally distributed, then the random variable $\overline{x}_1 - \overline{x}_2$ satisfies

$$\overline{x}_1 - \overline{x}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right).$$

That is, the **sampling distribution** of $\overline{x}_1 - \overline{x}_2$ is normal with mean $\mu_1 - \mu_2$ and standard deviation

$$\sigma_{\overline{x}_1 - \overline{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

It also follows, by direct standardization, that

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0, 1).$$

IMPLICATIONS: From the last result, we can make the following observations:

- A $100(1-\alpha)$ percent **confidence interval** for $\mu_1 - \mu_2$ is given by

$$\left[(\overline{x}_1 - \overline{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ (\overline{x}_1 - \overline{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right].$$

- To test $H_0 : \mu_1 - \mu_2 = 0$ versus $H_1 : \mu_1 - \mu_2 \neq 0$ at the α significance level, we can use the **two-sample z statistic**

$$z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

  Note that this test statistic is computed assuming that H_0 is true; i.e., $\mu_1 - \mu_2 = 0$. Thus, values of $z > z_{\alpha/2}$ and $z < -z_{\alpha/2}$ lead to rejecting H_0 in favor of H_1. Of course, **one-sided tests** could be carried out as well. **Probability values** could also be used (e.g., areas to the right/left of z).

PROBLEM: As in the one-sample case, it will rarely be true that we know the population standard deviations σ_1 and σ_2. If we do not know these, then the two-sample confidence interval and two-sample z statistic above are really not useful!

SOLUTION: As in the one-sample setting, we do the almost obvious thing. If σ_1 and σ_2 are not known, *we estimate them*! To estimate σ_1, we use the **sample standard deviation** from sample 1; i.e.,

$$s_1 = \sqrt{\frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (x_{1j} - \overline{x}_1)^2},$$

where $\overline{x}_1$ is the sample mean of the data from sample 1. Similarly, to estimate σ_2, we use the **sample standard deviation** from sample 2; i.e.,

$$s_2 = \sqrt{\frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (x_{2j} - \overline{x}_2)^2},$$

where $\overline{x}_2$ is the sample mean of the data from sample 2.

IMPORTANT FACT: Suppose that $x_{11}, x_{12}, ..., x_{1n_1}$ is an SRS of size n_1 from a $\mathcal{N}(\mu_1, \sigma_1)$ population, that $x_{21}, x_{22}, ..., x_{2n_2}$ is an SRS of size n_2 from a $\mathcal{N}(\mu_2, \sigma_2)$ population, and that the two samples are **independent** of each other. Then, the quantity

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

has (approximately) a t_k sampling distribution, where

$$k \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}.$$

If k is not an integer in the formula above, we can simply round k to the nearest integer. This expression is sometimes called **Satterwaite's formula** for the degrees of freedom. That $t \sim t_k$ is not an exact result; rather it is an **approximation**. The approximation is best when the population distributions are normal. For nonnormal populations, this approximation improves with larger sample sizes n_1 and n_2.

REMARK: The authors of your textbook note that computing k by hand is frightful. In the light of this, the authors recommend that, instead of computing k via Satterwaite's formula, you use

$$k = \text{the smaller of } n_1 - 1 \text{ and } n_2 - 1.$$

I think this is a silly recommendation (although I understand why they suggest it). *In practice, software packages are going to compute k anyway, and there is no reason not to use software!*

7.4.2 Two-sample t confidence intervals

RECALL: When σ_1 and σ_2 were known, we learned that a $100(1-\alpha)$ percent confidence interval for $\mu_1 - \mu_2$ was given by

$$\left[(\overline{x}_1 - \overline{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ (\overline{x}_1 - \overline{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right].$$

When σ_1 and σ_2 are **not known**, a $100(1-\alpha)$ percent confidence interval for $\mu_1 - \mu_2$ becomes

$$\left[(\overline{x}_1 - \overline{x}_2) - t_{k,\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},\ (\overline{x}_1 - \overline{x}_2) + t_{k,\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right],$$

where the degrees of freedom k is computed using Satterwaite's formula. This is called a **two-sample t confidence interval** for $\mu_1 - \mu_2$.

INTERPRETATION: The two-sample t confidence interval gives plausible values for the difference $\mu_1 - \mu_2$.

- If this interval includes 0, then $\mu_1 - \mu_2 = 0$ is a plausible value for the difference, and there is no reason to doubt that the means are truly different.
- If this interval does not include 0, then $\mu_1 - \mu_2 = 0$ is a not plausible value for the difference. That is, it looks as though the means are truly different!
- Thus, we can use the confidence interval to make a decision about the difference!

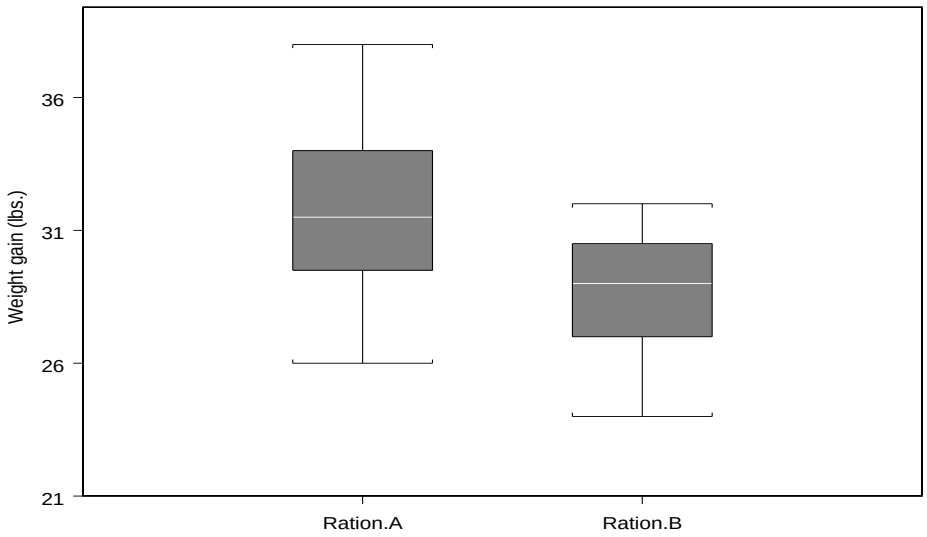


Figure 7.58: *Boxplots of pig weight gains by ration in Example* 7.6.

Example 7.6. Agricultural researchers are interested in two types of rations, 1 and 2, being fed to pigs. An experiment was conducted with 24 pigs randomly selected from a specific population; then

- $n_1 = 12$ pigs were randomly assigned to Ration 1, then
- the other $n_2 = 12$ were assigned to Ration 2.

The goal of the experiment was to determine whether there is a mean difference in the weight gains (lbs) for pigs fed the two different rations. Because of this, researchers are interested in $\mu_1 - \mu_2$, the **mean difference**. The data (weight gains, in lbs) from the experiment are below. Boxplots for the data appear in Figure 7.58.

```
Ration 1(A): 31  34  29  26  32  35  38  34  30  29  32  31
Ration 2(B): 26  24  28  29  30  29  32  26  31  29  32  28
```

Assuming that these data are well modeled by normal distributions, we would like to find a 95 percent confidence interval for $\mu_1 - \mu_2$ based on this experiment.

MINITAB: Here is the output from Minitab for Example 7.6:

```
Two-sample T for Ration 1 vs Ration 2

    N   Mean  StDev  SE Mean
1  12  31.75   3.19     0.92
2  12  28.67   2.46     0.71

Difference = mu (1) - mu (2)
Estimate for difference: 3.08333
95% CI for difference: (0.65479, 5.51187)
T-Test of difference = 0 (vs not =): T = 2.65 P-Value = 0.015 DF = 20
```

INTERPRETATION: From the output, we see that $\overline{x}_1 = 31.75$, $\overline{x}_2 = 28.67$, $s_1 = 3.19$, $s_2 = 2.46$, and $k \approx 20$ (this is Satterwaite's degrees of freedom approximation). From Table D, we see that $t_{20,0.025} = 2.086$. **For right now**, we only focus on the 95 percent confidence interval for $\mu_1 - \mu_2$; this interval is

$$(\overline{x}_1 - \overline{x}_2) \pm t_{k,\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \Longrightarrow (31.75 - 28.67) \pm 2.086\sqrt{\frac{(3.19)^2}{12} + \frac{(2.46)^2}{12}},$$

or $(0.65, 5.51)$. That is, we are 95 percent confident that the mean difference $\mu_1 - \mu_2$ is between 0.65 and 5.51 lbs. Indeed, the evidence suggests that the two rations are different with respect to weight gain.

7.4.3 Two-sample t hypothesis tests

GOAL: Based on random samples from normal populations, we would like to develop a hypothesis testing strategy for

$$H_0 : \mu_1 - \mu_2 = 0$$

versus

$$H_1 : \mu_1 - \mu_2 \neq 0,$$

where μ_1 is the mean of population 1 and μ_2 is the mean of population 2. We do not assume that the population standard deviations σ_1 and σ_2 are known. One-sided tests may also be of interest; these can be performed in the usual way.

RECALL: Suppose that $x_{11}, x_{12}, ..., x_{1n_1}$ is an SRS of size n_1 from a $\mathcal{N}(\mu_1, \sigma_1)$ population, that $x_{21}, x_{22}, ..., x_{2n_2}$ is an SRS of size n_2 from a $\mathcal{N}(\mu_2, \sigma_2)$ population, and that the two samples are **independent** of each other. Then, the quantity

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

has (approximately) a t_k sampling distribution, where

$$k \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}.$$

Thus, when $H_0 : \mu_1 - \mu_2 = 0$ is true, the **two-sample** t **statistic** becomes

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

Note the two-sample t statistic is computed assuming that $H_0 : \mu_1 - \mu_2 = 0$ is **true**. When H_0 is true, this statistic varies (approximately) according to a t_k sampling distribution, where k is computed using Satterwaite's formula. Thus, we can do hypothesis tests (one or two-sided tests) by computing t and comparing it to the t_k distribution! Values of t that are "unlikely" lead to the rejection of H_0.

Example 7.7. Two identical footballs, one air-filled and one helium-filled, were used outdoors on a windless day at The Ohio State University's athletic complex. Each football was kicked 39 times and the two footballs were alternated with each kick. The experimenter recorded the distance traveled by each ball. Rather than give you the raw data, we'll look at the summary statistics (**sample** means and variances). Boxplots appear in Figure 7.59.

Treatment	n_i	$\overline{x}_i$	s_i^2
Air	39	25.92	21.97
Helium	39	26.38	38.61

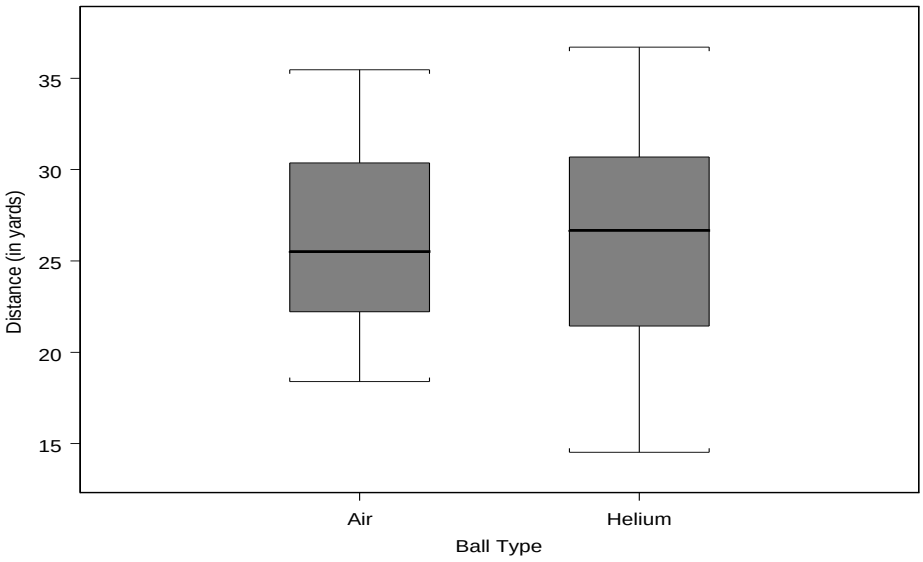


Figure 7.59: *Ohio State football data.*

Let μ_1 denote the **population** mean distance kicked using air and let μ_2 denote the **population** mean distance kicked using helium. The belief is that those balls kicked with helium, on average, will be **longer** than those balls filled with air, so that $\mu_1 - \mu_2 < 0$ (this is the researcher's hypothesis). Thus, we are interested in testing, at $\alpha = 0.05$, say,

$$H_0 : \mu_1 - \mu_2 = 0$$

versus

$$H_1 : \mu_1 - \mu_2 < 0.$$

The two-sample t test statistic is given by

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{25.92 - 26.38}{\sqrt{\frac{21.97}{39} + \frac{38.61}{39}}} = -0.37.$$

Is this an unusual value? To make this decision, we compare t to a t_k distribution, where

$$k \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left(\frac{21.97}{39} + \frac{38.61}{39}\right)^2}{\frac{\left(\frac{21.97}{39}\right)^2}{38} + \frac{\left(\frac{38.61}{39}\right)^2}{38}} \approx 71.$$

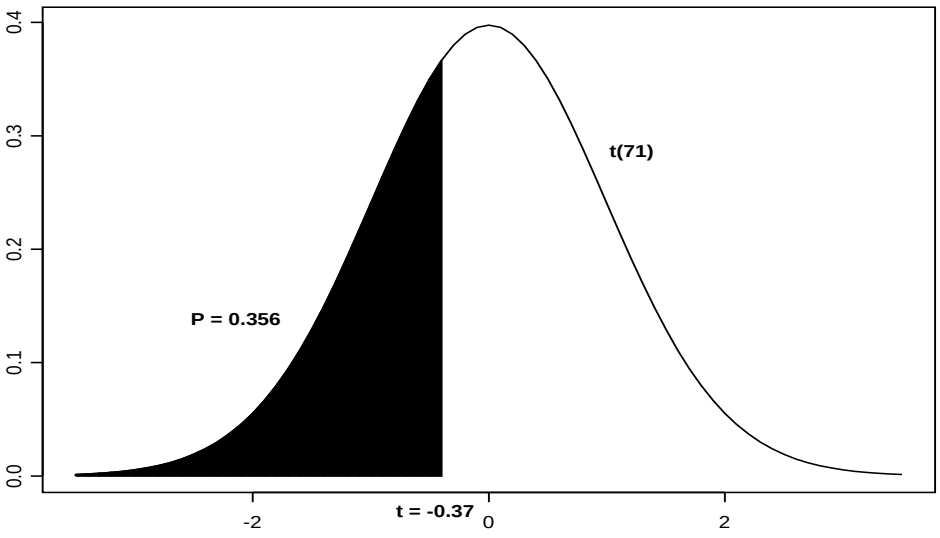


Figure 7.60: *The t_{71} density curve, test statistic, and probability value for Example* 7.7.

Table D (MM) doesn't have a row for $k = 71$ degrees of freedom, but Minitab provides $-t_{71,0.05} = -1.67$. Note that this value is close to $-z_{0.05} = -1.65$ (why is this?)

DECISION: Since $t = -0.37$ is not less than $-t_{71,0.05} = -1.67$, we do not reject H_0. That is, there is insufficient evidence to conclude that those balls kicked with helium, on average, will travel farther than those balls filled with air. The **probability value** for this test is $P = 0.356$, the area to the left of $t = -0.37$ under the t_{71} density curve; see Figure 7.60. We do not have a statistically significant result at any reasonable α level!!

ROBUSTNESS AGAIN: Like the one-sample t procedure, the two-sample t procedures are **robust** as well. That is, our populations need not be exactly normal for the two-sample t confidence intervals and hypothesis tests to provide accurate results, as long as there are no **serious departures** from normality and the sample sizes are not too small. A good exploratory analysis should always be done beforehand to check whether or not the normality assumption seems reasonable. If it does not, then the results you obtain from these procedures may not be valid.

7.4.4 Two-sample pooled t procedures

REMARK: The two-sample t procedures that we have discussed so far (in Sections 7.4.2 and 7.4.3) are always usable when the underlying distributions are normal (or approximately normal because of robustness). However, it turns out that when the true population standard deviations are **equal**; i.e., when $\sigma_1 = \sigma_2$, there is a different approach to take when making inferential statements about $\mu_1 - \mu_2$ (in the form of a confidence interval or hypothesis test). This approach is called the **pooled approach**.

NOTE: There are many instances where it is reasonable to assume that the two populations have a common standard deviation; i.e., $\sigma_1 = \sigma_2 = \sigma$. One interpretation is that the treatment application affects **only the mean** of the response, but not the variability. If there is strong empirical evidence that $\sigma_1 \neq \sigma_2$, do not use these pooled procedures! Formal hypothesis tests do exist for testing $\sigma_1 = \sigma_2$, but your text does not recommend using them (see Section 7.3, MM; we'll skip this section). As it turns out, hypothesis tests for $\sigma_1 = \sigma_2$ are not robust to nonnormality; thus, the tests can give unreliable results when population distributions are not exactly normal.

RULE OF THUMB: Do not use the pooled two-sample procedures if the **ratio** of the largest sample standard deviation to the smallest sample standard deviation exceeds 2.

A POOLED ESTIMATE: If both populations have a common standard deviation, then both populations have a common variance; i.e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$, and both of the sample variances s_1^2 and s_2^2 are estimating this common value of σ^2. Hence, in computing an estimate for σ^2, we will **pool** observations together and form a **weighted average** of the two sample variances, s_1^2 and s_2^2. That is, our estimate of σ^2 becomes

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

This is called the **pooled estimate** of σ^2, the assumed common variance.

IMPORTANT FACT: Suppose that $x_{11}, x_{12}, ..., x_{1n_1}$ is an SRS of size n_1 from a $\mathcal{N}(\mu_1, \sigma_1)$ population, that $x_{21}, x_{22}, ..., x_{2n_2}$ is an SRS of size n_2 from a $\mathcal{N}(\mu_2, \sigma_2)$ population, and

that the two samples are **independent** of each other. When $\sigma_1 = \sigma_2$ (i.e., the population standard deviations are equal), then the quantity

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has a $t_{n_1+n_2-2}$ sampling distribution. This is an exact result; i.e., it is not approximation (as long as the population standard deviations are truly equal). The quantity $s_p = \sqrt{s_p^2}$, where s_p^2 is the **pooled sample variance** given above.

POOLED CONFIDENCE INTERVAL FOR $\mu_1 - \mu_2$: From the last result, we may conclude that a $100(1-\alpha)$ percent confidence interval for $\mu_1 - \mu_2$ is given by

$$\left[(\overline{x}_1 - \overline{x}_2) - t_{n_1+n_2-2,\alpha/2} \times s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ (\overline{x}_1 - \overline{x}_2) + t_{n_1+n_2-2,\alpha/2} \times s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right].$$

This is an exact interval when the population distributions are normal and when $\sigma_1 = \sigma_2 = \sigma$ (when the population standard deviations are truly equal). This interval is robust to departures from normality, but it is not robust to departures from the equal population standard deviation assumption!

Example 7.8. A botanist is interested in comparing the growth response of dwarf pea stems to different levels of the hormone indoleacetic acid (IAA). Using 16-day-old pea plants, the botanist obtains 5-millimeter sections and floats these sections on solutions with different hormone concentrations to observe the effect of the hormone on the growth of the pea stem. Let x_1 and x_2 denote, respectively, the independent growths that can be attributed to the hormone during the first 26 hours after sectioning for $\frac{1}{2}10^{-4}$ and 10^{-4} levels of concentration of IAA (measured in mm). Previous studies show that both x_1 and x_2 are **approximately normal with equal variances**; that is, we are assuming that $\sigma_1^2 = \sigma_2^2 = \sigma^2$. The botanist would like to determine if there is a difference in means for the two treatments. To do this, she would like to compute a 90 percent confidence interval for $\mu_1 - \mu_2$, the difference in the population mean growth response.

Here are the data from the experiment; boxplots for the data are in Figure 7.61.

Treatment 1: `0.8 1.8 1.0 0.1 0.9 1.7 1.0 1.4 0.9 1.2 0.5`

Treatment 2: `1.0 0.8 1.6 2.6 1.3 1.1 2.4 1.8 2.5 1.4 1.9 2.0 1.2`

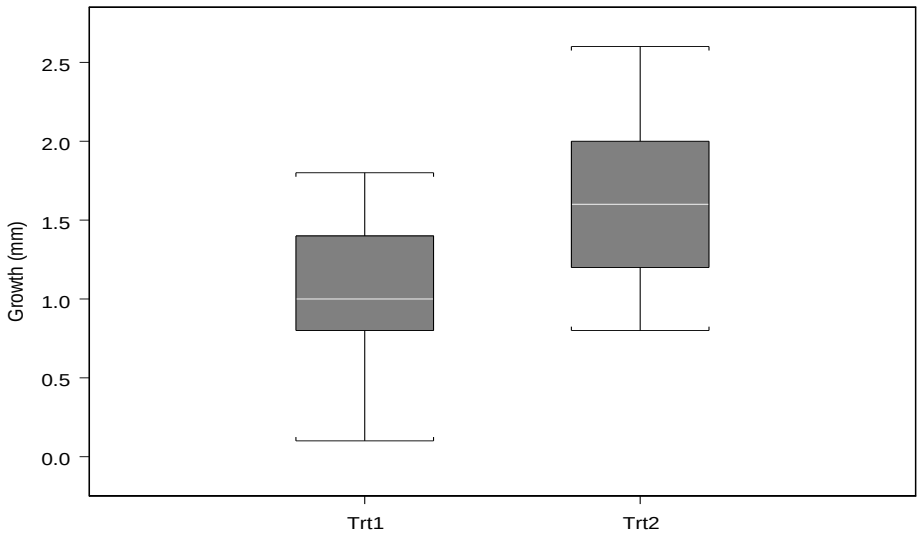


Figure 7.61: *Boxplots of stem growths by treatment.*

MINITAB: Here is the Minitab output for the analysis:

```
Two-sample T for Trt 1 vs Trt 2

        N   Mean  StDev  SE Mean
Trt 1  11  1.027  0.494     0.15
Trt 2  13  1.662  0.594     0.16

Difference = mu (Trt 1) - mu (Trt 2)
Estimate for difference:  -0.634266
90% CI for difference:  (-1.021684, -0.246847)
T-Test of difference = 0 (vs not =): T = -2.81  P-Value = 0.010  DF = 22
Both use Pooled StDev = 0.5507
```

ANALYSIS: From the output, we see $\overline{x}_1 = 1.027$, $\overline{x}_2 = 1.662$, $s_1 = 0.494$, and $s_2 = 0.594$. The degrees of freedom associated with this experiment is $n_1 + n_2 - 2 = 11 + 13 - 2 = 22$.

Our estimate of the common variance σ^2 is given by the pooled estimate

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{10(0.494)^2 + 12(0.594)^2}{11+13-2} = (0.5507)^2 \approx 0.3033.$$

From Table D, we see that $t_{22,0.05} = 1.715$. **For right now**, we only focus on the 90 percent confidence interval for $\mu_1 - \mu_2$; this interval is

$$(\overline{x}_1 - \overline{x}_2) \pm t_{n_1+n_2-2,\alpha/2} \times s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \Longrightarrow (1.027 - 1.662) \pm 1.715 \times 0.5507\sqrt{\frac{1}{11}+\frac{1}{13}},$$

or $(-1.022, -0.247)$ mm. That is, we are 90 percent confident that the mean difference $\mu_1 - \mu_2$ is between -1.022 and -0.247 mm. Since this interval does not include zero, the evidence suggests that the two levels of IAA have a different with respect to growth.

POOLED HYPOTHESIS TESTS FOR $\mu_1 - \mu_2$: When the researcher assumes that $\sigma_1 = \sigma_2 = \sigma$ (a common population standard deviation), the form of the two-sample t statistic changes as well. We are interested in testing $H_0 : \mu_1 - \mu_2 = 0$ versus one or two-sided alternatives.

RECALL: Suppose that $x_{11}, x_{12}, ..., x_{1n_1}$ is an SRS of size n_1 from a $\mathcal{N}(\mu_1, \sigma_1)$ population, that $x_{21}, x_{22}, ..., x_{2n_2}$ is an SRS of size n_2 from a $\mathcal{N}(\mu_2, \sigma_2)$ population, and that the two samples are **independent** of each other. When $\sigma_1 = \sigma_2$ (i.e., the population standard deviations are equal), then the quantity

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

has a $t_{n_1-n_2-2}$ sampling distribution. Recall that $s_p = \sqrt{s_p^2}$, where s_p^2 is the pooled sample variance.

NOTE: Observe that when $H_0 : \mu_1 - \mu_2 = 0$, the quantity above becomes

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}.$$

This is called the **two-sample pooled t statistic**. When H_0 is true, this statistic varies according to a $t_{n_1+n_2-2}$ sampling distribution. Thus, we can make a decision about H_0 by comparing t to this density curve.

Example 7.9. In an experiment that compared a standard fertilizer and a modified fertilizer for tomato plants, a gardener randomized 5 plants to receive the standard fertilizer (Treatment 1) and 6 plants to receive the modified fertilizer (Treatment 2). To minimize variation, the plants were all planted in one row. The gardener's hypothesis was that the modified fertilizer **was superior** to that of the standard fertilizer (i.e., that the modified fertilizer produced a higher mean yield). At the $\alpha = 0.05$ significance level, the researcher would like to test

$$H_0 : \mu_1 - \mu_2 = 0$$

versus

$$H_1 : \mu_1 - \mu_2 < 0.$$

We assume that yields arise from two independent normal populations with a **common standard deviation**. To test his claim, the following data are yields that were observed in the experiment:

Standard (1):	26.7	17.9	21.1	15.5	18.8	
Modified (2):	28.6	25.7	29.5	19.1	19.7	24.3

MINITAB: Since we are assuming that the population standard deviations are equal, we use a **pooled analysis**. Here is the Minitab output for this experiment:

```
Two-sample T for Fert.1 vs Fert.2

        N   Mean  StDev  SE Mean
Fert.1  5  20.00   4.25      1.9
Fert.2  6  24.48   4.37      1.8

Difference = mu (Fert.1) - mu (Fert.2)
Estimate for difference:  -4.48333
T-Test of difference = 0 (vs <): T-Value = -1.72  P-Value = 0.060  DF = 9
Both use Pooled StDev = 4.3165
```

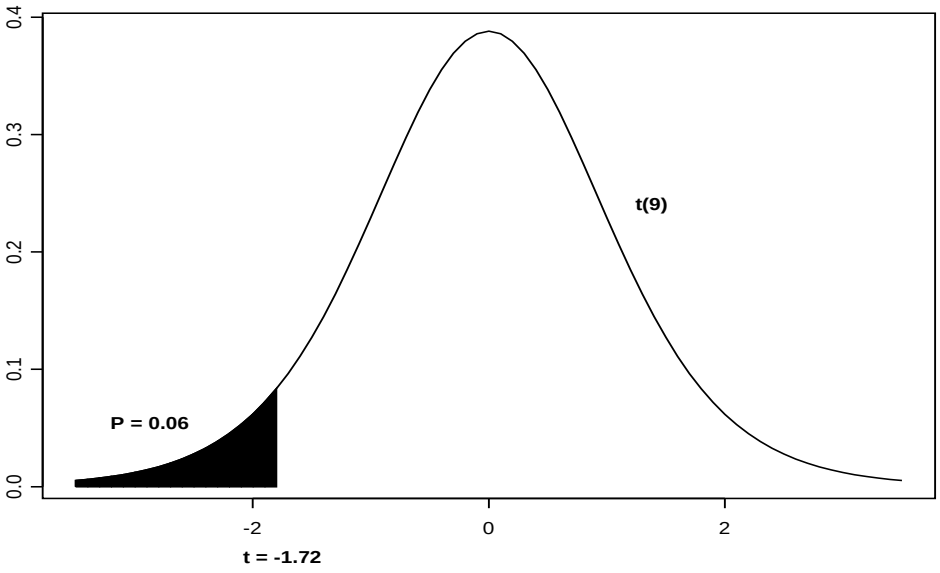


Figure 7.62: *The t_9 density curve, test statistic, and probability value for Example* 7.9.

ANALYSIS: From the output, we see $\overline{x}_1 = 20.00$, $\overline{x}_2 = 24.48$, $s_1 = 4.25$, and $s_2 = 4.37$. The degrees of freedom associated with this experiment is $n_1 + n_2 - 2 = 5 + 6 - 2 = 9$. Our estimate of the common variance σ^2 is given by the pooled estimate

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{4(4.25)^2 + 5(4.37)^2}{5 + 6 - 2} = (4.3165)^2 \approx 18.6322.$$

From Table D, we see that $-t_{9,0.05} = -1.833$. The two-sample pooled t statistic is given by

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{20.00 - 24.48}{4.3165\sqrt{\frac{1}{5} + \frac{1}{6}}} = -1.72.$$

Note that $t = -1.72$ does not fall in the rejection region since t is not less than $-t_{9,0.05} = -1.833$ (however, it is very close). The probability value is $P = 0.06$ (see Figure 7.62).

CONCLUSION: Officially, we do not have a statistically significant result at the $\alpha = 0.05$ level. That is, there is not enough evidence to conclude that the modified fertilizer is superior to the standard fertilizer. *There is evidence that the modified fertilizer outperforms the standard; just not enough at the $\alpha = 0.05$ level.* I call this a **borderline result**.

8 Inference for Proportions

REMARK: In Chapters 6 and 7, we have discussed inference procedures for **means** of **quantitative** variables using one or two samples. We now switch gears and discuss analogous procedures for **categorical** data. With categorical data, it is common to deal with **proportions**.

RECALL: We denote the **population proportion** by p. For example, p might denote the proportion of test plants infected with a certain virus, the proportion of students who graduate in 4 years or less, the proportion of the voting population who favor a certain candidate, or the proportion of individuals with a certain genetic condition. As we will now see, to make probability statements about p, we return to our use of the standard normal distribution.

8.1 Inference for a single population proportion

RECALL: Suppose that the number of successes X is observed in a binomial investigation; that is,

$$X \sim \mathcal{B}(n, p),$$

and let $\widehat{p} = X/n$ denote the **sample proportion** of successes. In Chapter 5, we argued (via simulation) that the approximate sampling distribution of $\widehat{p}$ was normal; in fact, mathematics shows that

$$\widehat{p} \sim \mathcal{AN}\left(p, \sqrt{\frac{p(1-p)}{n}}\right),$$

for large n (i.e., for large sample sizes). Recall that the abbreviation $\mathcal{AN}$ stands for **approximately normal**. *For small sample sizes, this approximation may not be adequate!* In the light of this result, we know that

$$z = \frac{\widehat{p} - p}{\sqrt{p(1-p)/n}} \sim \mathcal{AN}(0, 1),$$

for large sample sizes; that is, z has an approximate standard normal distribution. This provides the mathematical basis for the **one-sample z procedures for proportions**.

8.1.1 Confidence intervals

WALD INTERVAL: From the last result, we know that there exists a value $z_{\alpha/2}$ such that

$$P\left(-z_{\alpha/2} \leq \frac{\widehat{p}-p}{\sqrt{p(1-p)/n}} \leq z_{\alpha/2}\right) = 1-\alpha.$$

For example, if $\alpha = 0.05$ (i.e., 95 percent confidence), then $z_{\alpha/2} = z_{0.025} = 1.96$. If we perform some straightforward algebra on the event in the last probability equation and estimate $\sqrt{p(1-p)/n}$ with the **standard error**

$$\text{SE}_{\widehat{p}} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},$$

we obtain a $100(1-\alpha)$ **percent confidence interval** for p. This interval is given by

$$\left[\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},\ \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right].$$

We call this the **Wald confidence interval** for p. You will note that this interval has the **same form** as all of our other confidence intervals!! Namely, the form of the interval is

$$\text{estimate } \pm \text{ margin of error.}$$

Here, the estimate is the **sample proportion** $\widehat{p}$ and the **margin of error** is given by

$$m = z_{\alpha/2} \times \underbrace{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}}_{\text{SE}_{\widehat{p}}}.$$

WALD RULES OF THUMB: The Wald interval for p is based on the approximate sampling distribution result for $\widehat{p}$ given on the last page. *Remember that this result may not hold in small samples!* Thus, your authors have made the following **recommendations** regarding when to use this interval; namely, use the Wald interval when

- you want to get a 90, 95, or 99 percent confidence interval (or anything in between 90-99 percent), and
- both of the quantities $n\widehat{p}$ and $n(1-\widehat{p})$ are larger than 15.

Example 8.1. Apple trees in a large orchard were sprayed in order to control moth injuries to apples growing on the trees. A random sample of $n = 1000$ apples was taken from the orchard, 150 of which were injured. Thus, the sample proportion of injured apples was $\widehat{p} = 150/1000 = 0.15$ and a 95 percent Wald confidence interval for p, the population proportion of injured apples, is

$$\left[0.15 - 1.96\sqrt{\frac{0.15(1-0.15)}{1000}},\ 0.15 + 1.96\sqrt{\frac{0.15(1-0.15)}{1000}}\ \right],$$

or (0.128,0.172). That is, we are 95 percent confident that the true proportion of injured apples, p, is somewhere between 12.8 and 17.2 percent. We should check the Wald Rules of Thumb in this example:

- this is a 95 percent confidence interval and both of the quantities $n\widehat{p} = 1000(0.15) = 150$ and $n(1-\widehat{p}) = 1000(0.85) = 850$ are larger than 15.

REMARK: Checking the Wald interval Rules of Thumb is important. Why? If these guidelines are not met, the Wald interval may not be a very good interval!! Extensive simulation studies have shown that the Wald interval can be very **anticonservative**; that is, the interval produces a true confidence level much less than what you think your getting. In fact, it is not uncommon for a "95 percent" confidence interval to have a true confidence level around 0.85-0.90 or even lower!! This is why your authors have proposed the Rules of Thumb for checking the Wald interval's accuracy and appropriateness. *Summarizing, the Wald interval can perform very poorly when sample sizes are small.* This has prompted statisticians to consider other confidence interval expressions for p. The one we present now is a slight modification of the Wald interval and it is clearly superior, especially in small-sample size situations.

AGRESTI-COULL INTERVAL: The form of the Agresti-Coull (AC) interval is much like that of the Wald interval, except that a slight adjustment to the sample proportion is made. The Wald interval uses

$$\widehat{p} = \frac{X}{n} \quad \text{whereas the AC interval uses} \quad \widetilde{p} = \frac{X+2}{n+4}.$$

The estimate $\widetilde{p}$ is called the **plus-four estimate**. Note that this estimate is formed by adding two successes $(X + 2)$ and two failures, so that the total sample size is $n + 4$. The resulting confidence interval then is computed in the same way as the Wald interval; namely, the interval is given by

$$\left[\widetilde{p} - z_{\alpha/2}\sqrt{\frac{\widetilde{p}(1-\widetilde{p})}{n+4}},\ \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widetilde{p}(1-\widetilde{p})}{n+4}}\right].$$

This is called a $100(1-\alpha)$ **percent Agresti-Coull confidence interval** for p.

AGRESTI-COULL RULES OF THUMB: Use the AC interval when

- you want to get a 90, 95, or 99 percent confidence interval (or anything in between 90-99 percent), and
- the sample size n is 10 or more.

NOTE: Use the AC interval (instead of the Wald interval) especially when sample sizes are small!

Example 8.2. At a local supermarket, the owner has received 32 checks from out of state customers. Of these, 2 of the checks "bounced" (i.e., this means that the customer's bank account had insufficient funds to cover the amount on the check). For insurance purposes, the store owner would like to estimate the proportion of bad checks received monthly using a 90 percent confidence interval. Because the sample size is small, it is best to use the Agresti-Coull interval. Here, the plus-four estimate of p is given by

$$\widetilde{p} = \frac{X+2}{n+4} = \frac{2+2}{32+4} = \frac{4}{36} \approx 0.111.$$

Recalling that with 90 percent confidence, $z_{\alpha/2} = z_{0.10/2} = z_{0.05} = 1.65$ (Table A), the 90 percent AC interval is given by

$$\left[0.111 - 1.65 \times \sqrt{\frac{0.111(1-0.111)}{36}},\ 0.111 + 1.65 \times \sqrt{\frac{0.111(1-0.111)}{36}}\right],$$

or (0.025,0.197). Thus, we are 90 percent confident that the true proportion of bad checks is between 2.5 and 19.7 percent.

8.1.2 Hypothesis tests

REMARK: We can also perform **hypothesis tests** regarding p, the population proportion. As with the Wald and AC confidence intervals, we will use the standard normal distribution to gauge which test statistics are unlikely under H_0.

TEST STATISTIC: Suppose that we observe the number of successes X in a binomial investigation; that is, $X \sim \mathcal{B}(n, p)$. To test the hypothesis $H_0 : p = p_0$, we compute the **one-sample z statistic for proportions**

$$z = \frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}.$$

Note that this test statistic is computed assuming that $H_0 : p = p_0$ is true. When H_0 is true, $z \sim \mathcal{AN}(0, 1)$; that is, z has an approximate standard normal distribution. Thus, unlikely values of z are located in the tails of this distribution! Of course, the significance level α and the direction of H_1 tells us which values of z are unlikely under H_0.

- H_0 is rejected at level α in favor of $H_1 : p > p_0$ when $z > z_\alpha$; i.e., z falls in the upper tail.
- H_0 is rejected at level α in favor of $H_1 : p < p_0$ when $z < -z_\alpha$; i.e., z falls in the lower tail.
- H_0 is rejected at level α in favor of the two-sided alternative $H_1 : p \neq p_0$ when $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$.

NOTE: **Probability values** are computed in the same manner! To be more explicit, to test H_0 versus H_1, the probability values are computed as areas under the standard normal curve; i.e.,

- $H_1 : p > p_0$: Probability value = area to the right of z
- $H_1 : p < p_0$: Probability value = area to the left of z
- $H_1 : p \neq p_0$: Probability value = twice that of a one-sided test.

RULES OF THUMB: To perform hypothesis tests for p, the appropriate Rules of Thumb are that both of the quantities np_0 and $n(1-p_0)$ are larger than 10.

Example 8.3. Dimenhydrinate, also known by the trade names Dramamine and Gravol, is an over-the-counter drug used to prevent motion sickness. A random sample of $n = 100$ Navy servicemen was given Dramamine to control seasickness while out at sea. From previous studies, it was estimated that about 25 percent of all men experienced some form of seasickness when not treated. To evaluate the effectiveness of Dramamine, it is desired to test, at $\alpha = 0.05$,

$$H_0 : p = 0.25$$
$$\text{versus}$$
$$H_1 : p < 0.25,$$

where p denotes the proportion of men who would experience seasickness when treated with Dramamine. Of the 100 servicemen, 20 of them did, in fact, experience seasickness. Thus, $\widehat{p} = \frac{20}{100} = 0.20$, and our one-sample z statistic is

$$z = \frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.20 - 0.25}{\sqrt{0.25(1-0.25)/100}} = -1.15.$$

Is this an unlikely value of the test statistic? We can make our decision using either the rejection region or probability value approach.

- **Rejection region approach.** Since our test statistic $z = -1.15$ is less than $-z_{0.05} = -1.65$, we do not reject H_0.
- **Probability value approach.** The probability value is the area to the left of $z = -1.15$; this area is 0.1251 (see Figure 8.63). Since the probability value is not smaller than $\alpha = 0.05$, we do not reject H_0.

CONCLUSION: At the five percent significance level, there is not enough evidence to support the hypothesis that Dramamine reduces the rate of seasickness.

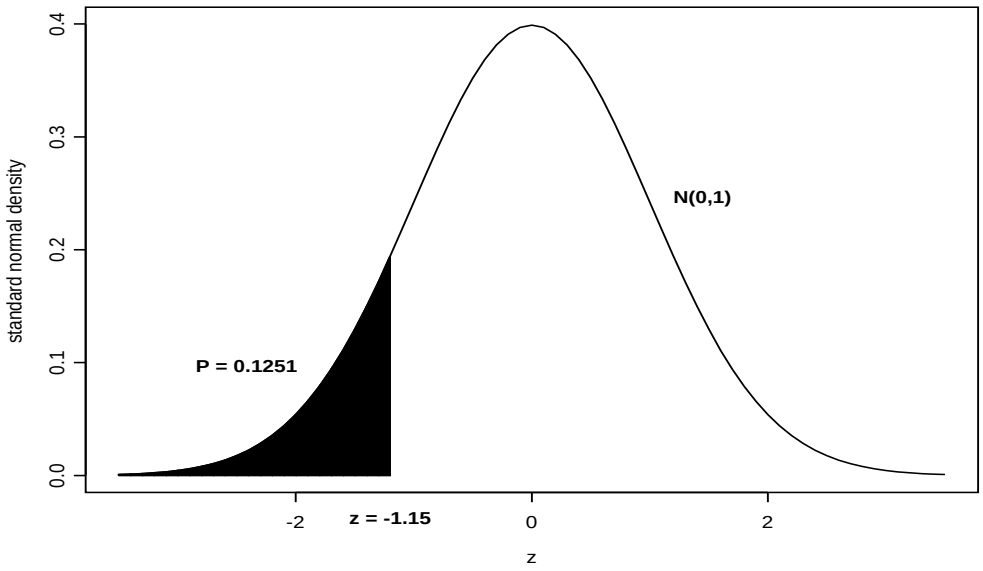


Figure 8.63: *The standard normal density curve, test statistic, and probability value for Example* 8.3.

8.1.3 Sample size determinations

RECALL: In Chapter 6, we discussed a sample size formula appropriate for estimating a population mean using a $100(1-\alpha)$ percent confidence interval with a certain prescribed margin of error. We now discuss the analogous problem with proportions. We will focus explicitly on the Wald interval.

CHOOSING A SAMPLE SIZE: To determine an appropriate sample size for estimating p, we need to specify the margin of error

$$m = z_{\alpha/2} \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

However, you will note that m depends on $\widehat{p}$, which, in turn, depends on n! This is a small problem, but we can overcome the problem by replacing $\widehat{p}$ with p^*, a **guess** for the value of p. Doing this, the last expression becomes

$$m = z_{\alpha/2} \times \sqrt{\frac{p^*(1-p^*)}{n}}.$$

Solving this last equation for n, we get

$$n = \left(\frac{z_{\alpha/2}}{m}\right)^2 p^*(1-p^*).$$

This is the desired sample size to find a $100(1-\alpha)$ percent confidence interval for p with a prescribed margin of error equal to m.

CONSERVATIVE APPROACH: If there is no sensible guess for p available, use $p^* = 0.5$. In this situation, the resulting value for n will be as large as possible. Put another way, using $p^* = 0.5$ gives the most **conservative** solution (i.e., the largest sample size).

Example 8.4. In a Phase II clinical trial, it is posited that the proportion of patients responding to a certain drug is $p^* = 0.4$. To engage in a larger Phase III trial, the researchers would like to know how many patients they should recruit into the study. Their resulting 95 percent confidence interval for p, the true population proportion of patients responding to the drug, should have a margin of error no greater than $m = 0.03$. What sample size do they need for the Phase III trial?

SOLUTION. Here, we have $m = 0.03$, $p^* = 0.4$, and $z_{\alpha/2} = z_{0.05/2} = 1.96$ (Table A). The desired sample size is

$$n = \left(\frac{z_{\alpha/2}}{m}\right)^2 p^*(1-p^*) = \left(\frac{1.96}{0.03}\right)^2 0.4(1-0.4) = 1025.$$

Thus, their Phase III trial should recruit 1025 patients.

8.2 Comparing two proportions

COMPARING PROPORTIONS: Analogously to the problem of comparing two population means (Chapter 7; Section 7.3), we would also like to **compare two population proportions**, say, p_1 and p_2. For example, do male and female voters differ on their support of a political candidate? Is the response rate higher for a new drug than the control drug? Is the the proportion of nonresponse different for two credit card mailings? We can use the following methods to answer these types of questions.

FRAMING THE PROBLEM: Our conceptualization of the problem is similar to that for comparing two means. We now have two independent binomial investigations:

$$X_1 \sim \mathcal{B}(n_1, p_1)$$
$$X_2 \sim \mathcal{B}(n_2, p_2).$$

From the investigations, our two **sample proportions** are $\widehat{p}_1 = X_1/n_1$ and $\widehat{p}_2 = X_2/n_2$. Clearly the problem involves the difference of these proportions; i.e.,

$$D = \widehat{p}_1 - \widehat{p}_2.$$

If the true population proportions p_1 and p_2 were, in fact, equal, then we would expect to see values of $\widehat{p}_1 - \widehat{p}_2$ close to zero. We need to know how this statistic varies in repeated sampling.

8.2.1 Confidence intervals

MATHEMATICAL RESULTS: Choose independent SRSs from two populations with proportions p_1 and p_2, respectively. The estimator

$$\widehat{p}_1 - \widehat{p}_2 \sim \mathcal{AN}\left(p_1 - p_2,\ \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right).$$

That is, the sampling distribution of $\widehat{p}_1 - \widehat{p}_2$ is approximately normal with mean $p_1 - p_2$ and standard deviation

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

Replacing p_1 and p_2 with their estimates $\widehat{p}_1$ and $\widehat{p}_2$, respectively, in the last expression gives the **standard error** of $D = \widehat{p}_1 - \widehat{p}_2$. That is,

$$\text{SE}_D = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}.$$

TWO SAMPLE WALD INTERVAL: An **approximate** $100(1-\alpha)$ **percent confidence interval** for $p_1 - p_2$, based on two independent random samples, is given by

$$(\widehat{p}_1 - \widehat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}.$$

We call this the **two-sample Wald interval** for $p_1 - p_2$. Note that this interval has the same form as all of our other confidence intervals!! Namely, the form of the interval is

$$\text{estimate} \ \pm \ \text{margin of error.}$$

Here, the estimate is $D = \widehat{p}_1 - \widehat{p}_2$ and the **margin of error** is given by

$$z_{\alpha/2} \times \text{SE}_D = z_{\alpha/2} \times \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}.$$

RULES OF THUMB: Much like the one-sample problem considered in Section 8.1, there are guidelines to use with the two-sample Wald interval. Your authors recommend to use the Wald interval only when

- you want to get a 90, 95, or 99 percent confidence interval (or anything in between 90-99 percent), and
- the quantities $n_1\widehat{p}_1$, $n_1(1-\widehat{p}_1)$, $n_2\widehat{p}_2$, and $n_2(1-\widehat{p}_2)$ are all larger than 10.

Example 8.5. An experimental type of chicken feed, Ration 1, contains an unusually large amount of a certain feed ingredient that enables farmers to raise heavier chickens. However, farmers are warned that the new feed may be too strong and that the mortality rate may be higher than that with the usual feed. One farmer wished to compare the mortality rate of chickens fed Ration 1 with the mortality rate of chickens fed the current best-selling feed, Ration 2. Denote by p_1 and p_2 the population mortality rates (proportions) for Ration 1 and Ration 2, respectively. Researchers would like to get a 95 percent confidence interval for $p_1 - p_2$. Two hundred chickens were randomly assigned to each ration; of those fed Ration 1, 24 died within one week; of those fed Ration 2, 16 died within one week.

- Sample 1: 200 chickens fed Ration 1 $\Longrightarrow \widehat{p}_1 = 24/200 = 0.12$
- Sample 2: 200 chickens fed Ration 2 $\Longrightarrow \widehat{p}_2 = 16/200 = 0.08$.

Thus, the difference $D = \widehat{p}_1 - \widehat{p}_2 = 0.12 - 0.08 = 0.04$, and an approximate 95 percent Wald confidence interval for the true difference $p_1 - p_2$ (based on this experiment) is

$$\left[0.04 - 1.96\sqrt{\frac{0.12(0.88)}{200} + \frac{0.08(0.92)}{200}},\ 0.04 + 1.96\sqrt{\frac{0.12(0.88)}{200} + \frac{0.08(0.92)}{200}}\right],$$

or $(-0.02, 0.10)$. Thus, we are 95 percent confident that the true difference in mortality rates is between -0.02 and 0.10. Note that this interval includes zero, so we can not say that the mortality rates are necessarily different for the two rations (at the five percent level). It is easy to see that the Rules of Thumb are satisfied in this problem.

AGRESTI-CAFFO INTERVAL: Just as in the one-sample problem (Section 8.1), the Wald interval has some serious flaws when the sample sizes are small. An alternative interval is available for small sample sizes; it is called the **Agresti-Caffo interval**. The interval is based, again, on adding 2 successes and 2 failures. However, because we have two samples now, we add 1 success and 1 failure to each. To be more specific, we compute

$$\widetilde{p}_1 = \frac{X_1 + 1}{n_1 + 2} \quad \text{and} \quad \widetilde{p}_2 = \frac{X_2 + 1}{n_2 + 2}.$$

The $100(1-\alpha)$ **percent Agresti-Caffo (AC) confidence interval** for $p_1 - p_2$ is given by

$$(\widetilde{p}_1 - \widetilde{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\widetilde{p}_1(1-\widetilde{p}_1)}{n_1 + 2} + \frac{\widetilde{p}_2(1-\widetilde{p}_2)}{n_2 + 2}}.$$

Your authors recommend this method when confidence levels are between 90-99 percent and when both sample sizes n_1 and n_2 are at least 5.

Example 8.6. Who is a better breaker of his opponent's serve: Andre Agassi or Roger Federer? In a recent match between the two tennis stars, Agassi converted 3 of 13 break opportunities and Federer converted 6 of 8. For all of you non-tennis aficionados, a "break" of serve occurs when one player wins a game from his opponent while the opponent is serving. Denote by p_1 the true proportion of breaks for Agassi and p_2 the true proportion of breaks for Federer. Our plus-four estimates for p_1 and p_2 are given by

$$\widetilde{p}_1 = \frac{3+1}{13+2} \approx 0.27 \quad \text{and} \quad \widetilde{p}_2 = \frac{6+1}{8+2} = 0.70.$$

A 95 percent confidence interval for $p_1 - p_2$ is given by

$$(0.27 - 0.70) \pm 1.96\sqrt{\frac{0.27(1-0.27)}{13+2} + \frac{0.70(1-0.70)}{8+2}},$$

or $(-0.79, -0.07)$. Thus, we are 95 percent confident that the true difference $p_1 - p_2$ is between -0.79 and -0.07. Because the interval does not include 0, this suggests that there is a difference between the rates at which these two players break in the other's service games.

8.2.2 Hypothesis tests

HYPOTHESIS TEST FOR TWO PROPORTIONS: Analogously to performing a hypothesis test for two means (Section 7.3), we can also compare two proportions using a hypothesis test. To be precise, we would like to test

$$H_0 : p_1 - p_2 = 0$$

versus

$$H_1 : p_1 - p_2 \neq 0.$$

Of course, one-sided tests are available and are conducted in the usual way. Our **two-sample z statistic for proportions** is given by

$$z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

where

$$\widehat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

is the overall sample proportion of successes in the two samples. This estimate of p is called the **pooled estimate** because it combines the information from both samples. When $H_0 : p_1 - p_2 = 0$ is true, this statistic has an approximate standard normal distribution. Thus, values of z in the tails of this distribution are considered unlikely.

RULES OF THUMB: This two-sample test should perform adequately as long as the quantities $n_1\widehat{p}_1$, $n_1(1-\widehat{p}_1)$, $n_2\widehat{p}_2$, and $n_2(1-\widehat{p}_2)$ are all larger than 5.

Example 8.7. A tribologist is interested in testing the effect of two lubricants used in a manufacturing assembly found in rocket engines. Lubricant 1 is the standard lubricant that is currently used. Lubricant 2 is an experimental lubricant which is much more expensive. Does Lubricant 2 reduce the proportion of cracked bolts? A random sample of $n = 450$ is used, with 250 bolts receiving Lubricant 1 and the other 200 receiving Lubricant 2. After simulating rocket engine conditions, it is noted what proportion of the bolts have cracked. It is thus desired to test, at the conservative $\alpha = 0.01$ level,

$$H_0 : p_1 - p_2 = 0$$
$$\text{versus}$$
$$H_1 : p_1 - p_2 > 0,$$

where p_1 denotes the proportion of bolts that would crack using lubricant 1 and p_2 denotes the proportion of bolts that would crack using lubricant 2. After the experiment is over, it is noted that 21 of the bolts cracked using lubricant 1 ($\widehat{p}_1 = 21/250 = 0.084$), and 8 with lubricant 2 ($\widehat{p}_2 = 8/200 = 0.040$). The pooled estimate of p is given by

$$\widehat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{21 + 8}{250 + 200} = 0.064,$$

and the value of the two-sample z test statistic is

$$z = \frac{0.084 - 0.040}{\sqrt{0.064(1 - 0.064)\left(\frac{1}{250} + \frac{1}{200}\right)}} \approx 1.89.$$

The probability value for the test, the area to the right of $z = 1.89$ on the standard normal distribution, is 0.0294 (Table A). This is a small probability value! However, it is not small enough to be deemed significant at the $\alpha = 0.01$ level. Alternatively, you could have noted that the critical value here is $z_{0.01} = 2.33$. Our test statistic $z = 1.89$ does not exceed this value.

CONCLUSION: There is some evidence that Lubricant 2 does reduce the proportion of cracked bolts in this assembly, but not enough evidence to be statistically significant at the conservative $\alpha = 0.01$ level.