Reserve Services

Northwestern University Library

WARNING

CONCERNING COPYRIGHT RESTRICTIONS

The Copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or reproduction.

One of three specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research.

If electronic transmission of reserve material is used for purposes in excess of what constitutes "fair use", that user may be liable for copyright infringement.

This policy is in effect for the following document:

Chihara, Laura and Hesterberg, Tim.
Exploratory Data Analysis (Chapter 2) / from Mathematical Statistics with Resampling and R.
Hoboken, NJ: John Wiley & Sons, 2011. pp. 13-30.

NO FURTHER TRANSMISSION OR DISTRIBUTION OF THIS MATERIAL IS PERMITTED

2

EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) is an approach to examining and describing data to gain insight, discover structure, and detect anomalies and outliers. John Tukey (1915–2000), an American mathematician and statistician who pioneered many of the techniques now used in EDA, stated in his 1977 book *Exploratory Data Analysis* (Tukey (1977)) that "Exploratory data analysis is detective work—numerical detective work—counting detective work—or graphical detective work." In this chapter, we will learn many of the basic techniques and tools for gaining insight into data.

Statistical software packages can easily do the calculations needed for the basic plots and numeric summaries of data. We will use the software package `R`. We will assume that you have gone through the introduction to `R` available at the web site `https://sites.google.com/site/ChiharaHesterberg`.

2.1 BASIC PLOTS

In Chapter 1, we described data on the lengths of flight delays of two major airlines flying from LaGuardia Airport in New York City in 2009. Some basic questions we might ask include how many of these flights were flown by United Airlines and how many by American Airlines? How many flights flown by each of these airlines were delayed more than 30 min?

A *categorical* variable is one that places the observations into groups. For instance, in the `FlightDelays` data set, `Carrier` is a categorical variable (we will also

Mathematical Statistics with Resampling and R, First Edition. By Laura Chihara and Tim Hesterberg.
© 2011 John Wiley & Sons, Inc. Published 2011 by John Wiley & Sons, Inc.

TABLE 2.1 Counts for the `Carrier` Variable

	Carrier	
	American Airlines	United Airways
Number of flights	2906	1123

call this a *factor* variable) with two *levels*, UA and AA. Other data sets might have categorical variables such as `gender` (with two levels, Male or Female) or `size` (with levels Small, Medium, and Large).

A *bar chart* is used to describe the distribution of a categorical (factor) variable. Bars are drawn for each level of the factor variable and the height of the bar is the number of observations in that level. For the `FlightDelays` data, there were 2906 American Airlines flights and 1123 United Airlines flights (Table 2.1). The corresponding bar chart is shown in Figure 2.1.

We might also be interested in investigating the relationship between a carrier and whether or not a flight was delayed more than 30 min. A *contingency table* summarizes the counts in the different categories.

From Table 2.2, we can compute that 13.5% of American Airlines flights were delayed more than 30 min compared to 18.2% of United Airlines flights. Is this difference in percentages *statistically significant*? Could the difference in percentages be due to *natural variability*, or is there a systematic difference between the two airlines? We will address this question in the following chapters.

With a numeric variable, we will be interested in its distribution: What is the range of values? What values are taken on most often? Where is the *center*? What is the *spread*?

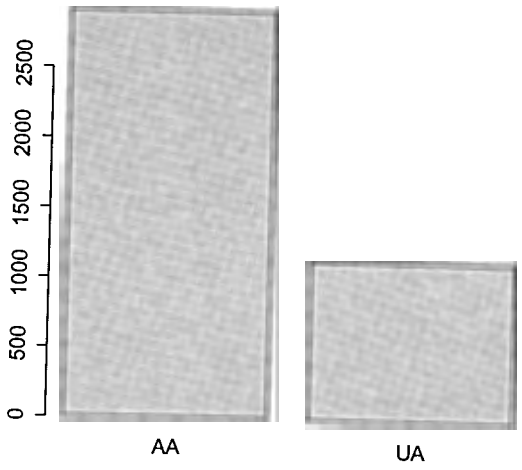

FIGURE 2.1 Bar chart of `Carrier` variable.

TABLE 2.2 Counts of Delayed Flights Grouped by Carrier

Carrier	Delayed More Than 30 Min?		
	No	Yes	Total
American Airlines	2513	393	2906
United Airlines	919	204	1123

For the flight delays data set, although we can inspect the distribution of the lengths of the delays with a table by partitioning the values into nonoverlapping intervals (Table 2.3), a visual representation is often more informative.

A *histogram* corresponding to Table 2.3 is shown in Figure 2.2. Note that the height of each bar reflects the frequency of flights whose delays fall in the corresponding interval. For example, 722 flights departed on time or earlier than scheduled, while 249 flights were delayed by at most 50 min. Some software will give users the option to create bar heights equal to proportions or percentages.

We describe this distribution as *right skewed*. Most of the flights departed on time (or were early) and the counts of late departures decrease as time increases.

Average January temperatures in the state of Washington follow a *left-skewed distribution* (Figure 2.3): in most years, average temperate fell in the 30–35 °F interval, and the number of years in which temperatures were less than 30 °F decreases as temperature decreases.

Remark The exact choice of subintervals to use is discretionary. Different software packages utilize various algorithms for determining the length of the subintervals; also, some software packages may use subintervals of the form $[a, b)$ instead of $(a, b]$. ||

For small data sets, a *dot plot* is an easy graph to create by hand. A dot represents one observation and is placed above the value it represents. The number of dots in a column represents the frequency of that value.

The dot plot for the data 4, 4.5, 4.5, 5, 5, 5, 6, 6, 6.5, 7, 7, 7 is shown in Figure 2.4.

TABLE 2.3 Distribution of Length of Flight Delays for United Airlines

Time Interval	Number of Flights
(−50, 0]	722
(0, 50]	249
(50, 100]	86
(100, 150]	39
(150, 200]	14
(200, 250]	7
(250, 300]	3
(350, 400]	2
(400, 450]	1

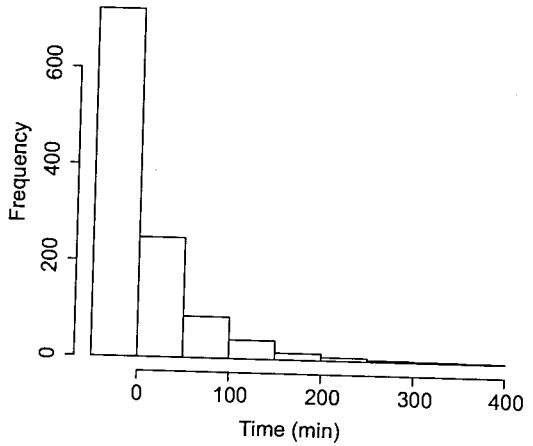

FIGURE 2.2 Histogram of lengths of flight delays for United Airlines. The distribution is right skewed.

2.2 NUMERIC SUMMARIES

It is often useful to have numerical descriptions of variables. Unfortunately, the old adage "a picture is worth a thousand words" cuts both ways—doing without a picture limits what we can say without thousands of words. So we focus on key characteristics—center, spread, and sometimes shape.

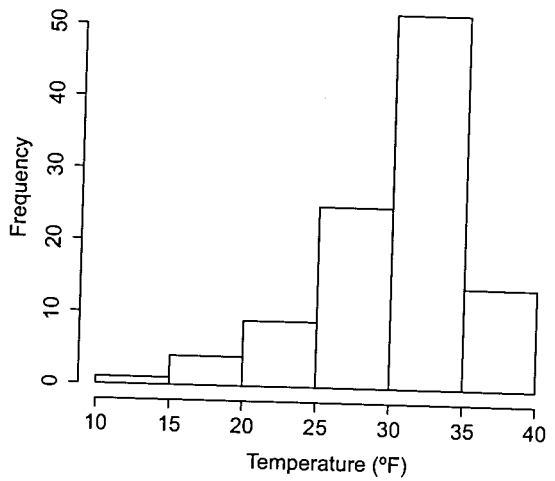

FIGURE 2.3 Histogram of average January temperatures in Washington state (1895–1999). The distribution is left skewed.

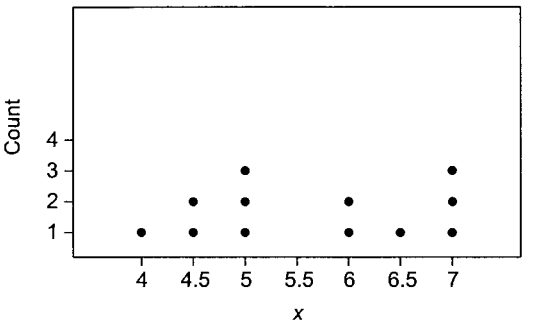

FIGURE 2.4 Example of a dot plot.

2.2.1 Center

First consider *center*. By eyeballing the histogram (Figure 2.2) of flight delay times, we might put the center at around 0. Two statistics commonly used to describe the *center* of a variable include the *mean* and *median*.

If $x_1, x_2, \ldots, x_n$ are n data values, then the mean is

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

The median is the middle value in a sorted arrangement of the values; that is, half the values are less than or equal to the median and half are greater. If $y_1 \le y_2 \le \cdots \le y_n$ denotes a sorted list of values and n is odd, the median is the middle value $y_{(n+1)/2}$. If n is even, then the median is the average of the two middle values, $(1/2)(y_{n/2} + y_{(n/2+1)})$.

A compromise between the mean and the median is a *trimmed mean*. The mean is the average of all observations, while the median is the average of the middle one or two observations. For a 25% trimmed mean, for example, you sort the data, omit 25% of the observations on each side, and take the mean of the remaining middle 50% of the observations. The 25% trimmed mean is also known as *midmean*.

Example 2.1 The mean of the 12 values 1, 3, 3, 4, 4, 7, 8, 10, 14, 21, 24, 26 is 10.42, the median is the average of the sixth and seventh values, $(7 + 8)/2 = 7.5$, and the midmean is the average of fourth through ninth values, 7.83.

The mean of the 15 values 1, 3, 3, 4, 4, 7, 8, 10, 14, 21, 24, 28, 30, 30, 34 is 14.73, the median is the 8th value, 10, and the midmean is the average of the 4th through 12th values, 13.33. □

Example 2.2 The mean length of a departure delay for United Airlines was 15.9831 min. The median length of a departure delay was −1.00 min; that is, half of the flights left more than 1 min earlier than their scheduled departure time. □

Remark Software may differ in how it calculates trimmed means. In R, `mean(x, trim = 0.25)` rounds $0.25n$ down; thus, for $n = 15$, three observations are omitted. ||

2.2.2 Spread

To describe spread, three common choices are the range, the interquartile range, and the standard deviation.

The *range* is the difference between the largest and smallest values.

The *interquartile range* (IQR) is the difference between the third and the first quartiles. It gives a better measure of the center of the data than does the range and is not sensitive to outliers.

The *sample standard deviation*, or *standard deviation*, is

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{2.1}$$

To motivate the standard deviation, we begin with a less common measure of spread, the *mean absolution deviation* (MAD), $(1/n)\sum_{i=1}^{n}|x_i - \bar{x}|$. This is the average distance from the mean and is a natural measure of spread. In contrast, the standard deviation is roughly the average squared distance from the mean, followed by a square root; the combination is roughly equal to the MAD, though usually a bit larger. The standard deviation tends to have better statistical properties.

There are a couple of versions of standard deviation. The *population standard deviation* is the square root of the *population variance*, which is the average of the squared distances from the mean, $(1/n)\sum_{i=1}^{n}(x_i - \bar{x})^2$. The *sample variance* is similar but with a divisor of $n-1$,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{2.2}$$

and the sample standard deviation is its square root. The population versions are appropriate when the data are the whole population. When the data are a sample from a larger population, we use the sample versions; in this case, the population versions tend to be too small—they are *biased*; we will return to this point in Section 6.3.1. For n large, there is little practical difference between using $n-1$ or n.

Example 2.3 The standard deviation of the departure delay times for United Airlines flights is 45.119 min. Since the observations represent a population (we compiled *all* United Airlines flights for the months of May and June), we use the definition with the $1/n$ term rather than the $1/(n-1)$ term. Using Equation 2.1 gives 45.139. □

2.2.3 Shape

To describe the shape of a data set, we may use skewness and kurtosis (see page 28.) However, more common and intuitive is to use the *five-number summary*: the minimum, first quartile, median, third quartile, and maximum value.

Example 2.4 Consider the 15 numbers 9, 10, 11, 11, 12, 14, 16, 17, 19, 21, 25, 31, 32, 41, 61.

The median is 17. Now, find the median of the numbers less than or equal to 17. This will be the first quartile $Q_1 = 11.5$. The median of the numbers greater than or equal to 17 is the third quartile $Q_3 = 28$. Thus, the five-number summary is 9, 11.5, 17, 28, 31. □

Remark Different software packages use different algorithms for computing quartiles, so do not be alarmed if your results do not match exactly. ||

2.3 BOXPLOTS

A boxplot is a type of graph that can be used to visualize the five-number summary of a set of numeric values.

Example 2.5 Consider the following set of 21 values (Table 2.4).

TABLE 2.4 A Set of 21 Data Values

5	6	6	8	9	11	11
14	17	17	19	20	21	21
22	23	24	32	40	43	49

The five-number summary for these data is 5, 11, 19, 23, 48 and the interquartile range is 23−11 = 12. The corresponding boxplot is shown in Figure 2.5. □

To create a boxplot

- Draw a box with the bottom placed at the first quartile and the top placed at the third quartile. Draw a line through the box at the median.
- Compute the number $Q_3 + 1.5 \times$ IQR, called the *upper fence*, and then place a cap at the largest observation that is less than or equal to this number.
- Similarly, compute the *lower fence*, $Q_1 - 1.5 \times$ IQR, and place a cap at the smallest observation that is greater than or equal to this number.
- Extend *whiskers* from the edge of the box to the caps.
- The observations that fall outside the caps will be considered *outliers* and separate points are drawn to indicate these values.

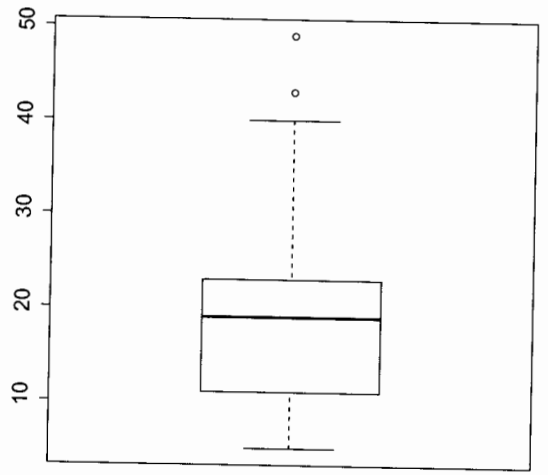

FIGURE 2.5 Boxplot for Table 2.4.

In the above example, the upper fence is $23 + 1.5 \times 12 = 41$. The largest observation that falls below this fence is 40, so a cap is drawn at 40. The lower fence is $11 - 1.5 \times 12 = -7$. The smallest observation that falls above this fence is 5, so a cap is drawn at 5. The outliers are 43 and 49.

Example 2.6 For the length of United Airlines flight delays, the five-number summary is -17.00, -5.00, -1.00, 12.50, 377.00. Thus, the interquartile range is $12.50 - (-5.00) = 17.50$ and half of the 1123 values are contained in an interval of length 17.50. □

Boxplots are especially useful in comparing the distribution of a numeric variable across levels of a factor variable.

Example 2.7 We can compare the lengths of the flight delays for United Airlines across the days of the week for which the departure was scheduled.

For instance, we can see that the most variability in delays seems to occur on Thursdays and Fridays (Figure 2.6). □

2.4 QUANTILES AND NORMAL QUANTILE PLOTS

For the random sample of 1009 babies born in North Carolina in 2004, the distribution of their weights is unimodal and roughly symmetric (Figure 2.7). We introduce another graph that allows us to compare this distribution with the normal distribution.

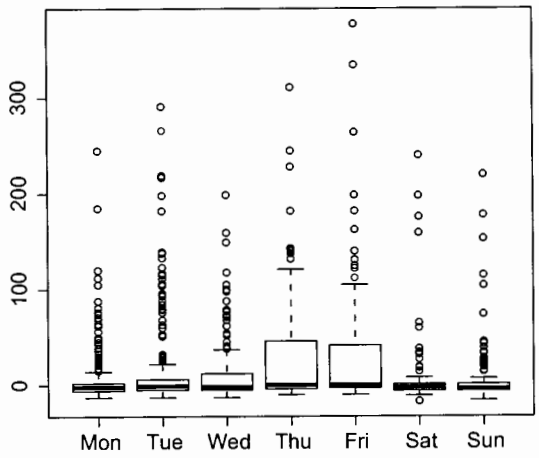

FIGURE 2.6 Distribution of lengths of the flight delays for United Airlines across the days of the week.

Definition 2.1 Let X denote a random variable. For $0 \leq p \leq 1$, the pth *quantile* of X is the value q_p such that $P(X \leq q_p) = p$. That is, q_p is the value at which the amount of area under the density curve (of X) to the left of q_p is p, or $p \times 100\%$ of the area under the curve is to the left of q_p.

Some books may use the term *percentile* rather than quantile. For instance, the 0.3 quantile is the same as the 30th percentile. ||

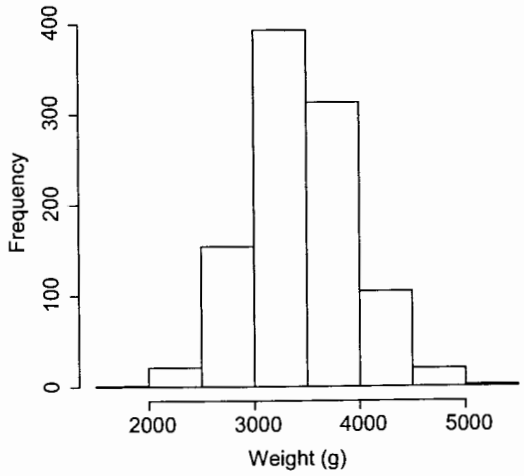

FIGURE 2.7 Distribution of birth weights for North Carolina babies.

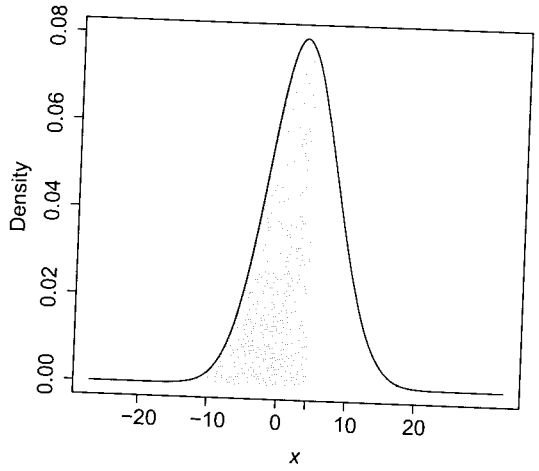

FIGURE 2.8 Density for $N(3, 5^2)$ with $P(X \le 4.27) = 0.6$.

Example 2.8 Let Z denote the standard normal distribution. Let $p = 0.5$. Then, the 0.5 quantile of Z is 0 since $P(Z \le 0) = 0.5$. That is, 0 is the 50th percentile of the standard normal distribution.

Let $p = 0.25$. Then, $q_{0.25} = -0.6744$ since $P(Z \le -0.6744) = 0.25$. That is, -0.6744 is the 25th percentile of the standard normal distribution. □

Example 2.9 Let X be a normal random variable, $N(3, 5^2)$. Find the $p = 0.6$ quantile.

We want q_p such that $P(X \le q_p) = 0.6$. The desired value is $q_p = 4.3$ (see Figure 2.8). □

R Note:

Use the `qnorm` command to find normal quantiles:

```
> qnorm(.25)          # standard normal
[1] -0.6744898
> qnorm(.6, 3, 5)     # N(3,5^2)
[1] 4.266736
```

We can also formulate quantiles in terms of the cumulative distribution function F of a random variable X since

$$F(q_p) = P(X \le q_p) = p \text{ implies } q_p = F^{-1}(p).$$

Example 2.10 Let X be an exponential random variable with $\lambda = 3$. The cdf of X is given by $F(x) = 1 - e^{-3x}$. Since $F^{-1}(y) = (-1/3)\ln(1 - y)$, the pth quantile is given by $q_p = (-1/3)\ln(1 - p)$.

Alternatively, since we know the pdf of X is $f(x) = e^{-3x}$, $x \ge 0$, we could also solve for q_p in

$$p = P(X \le q_p) = \int_0^{q_p} e^{-3t}\, dt.$$ □

Suppose the (sorted) data are $x_1 \le x_2 \le \cdots \le x_n$ and we wish to see if these data come from the normal distribution, $N(0, 1)$. The *normal quantile plot* for comparing distributions is a plot of the x's against $(q_1, x_1), (q_2, x_2), \ldots, (q_n, x_n)$, where q_k is the $k/(n + 1)$ quantile of the standard normal distribution. If these points fall (roughly) on a straight line, then we conclude that the data follow an approximate normal distribution. This is one type of *quantile–quantile plot*, or *qq plot* for short, in which quantiles of a data set are plotted against quantiles of a distribution or of another data set.

Example 2.11 Here, there are $n = 10$ points. We will look at the $i/(n + 1) = i/(11\text{th})$ quantiles, $i = 1, \ldots, 10$, of the standard normal.

x	17.7	22.6	26.1	28.3	30.0	31.2	31.5	33.5	34.7	36.0
p_i	1/11	2/11	3/11	4/11	5/11	6/11	7/11	8/11	9/11	10/11
q_p	−1.34	−0.91	−0.60	−0.35	−0.11	0.11	0.35	0.60	0.91	1.34

For instance, the q_p entry corresponding to $p_5 = 5/11 = 0.455$ (the 45.5th percentile) is

$$q_{0.455} = -0.11 \text{ because } P(Z \le -0.11) = 0.455.$$

To create a normal quantile plot, we graph the pairs (q_p, x). A straight line is often drawn through the points corresponding to the first and third quartiles of each variable (see Figure 2.9).

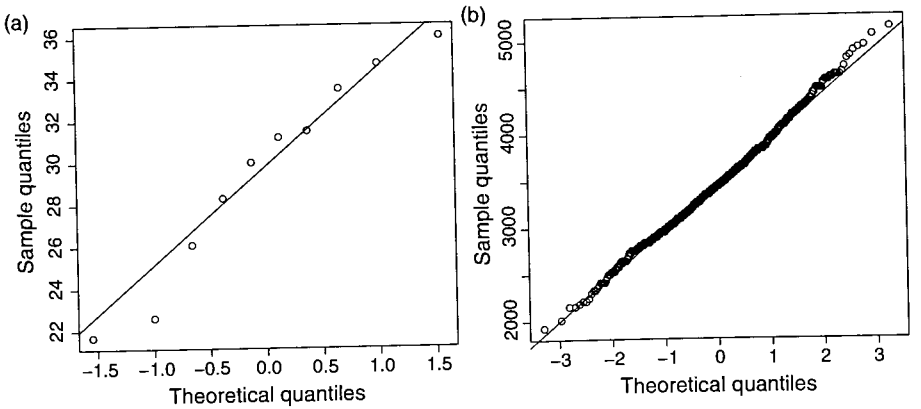

FIGURE 2.9 (a) Example of normal quantile plot for data in Example 2.11. (b) Normal quantile plot for weights of NC babies.

R Note:

The commands `qqnorm` and `qqline` can be used to create normal quantile plots:

```
x <- c(21.7, 22.6, 26.1, 28.3, 30, 31.2, 31.5, 33.5, 34.7, 36)
qqnorm(x)       # plot points
qqline(x)       # add straight line

qqnorm(NCBirths$Weight)
qqline(NCBirths$Weight)
```

The `qqnorm` command plots the quantiles of the standard normal on the x axis. The `qqline` command adds a straight line through the first and third quartiles of the data.

□

Recall that the distribution of the flight delay times for United Airlines is strongly right skewed (Figure 2.2). The normal quantile plots for these data and for the left-skewed distribution of average January temperatures in Washington state (Figure 2.3) are shown in Figure 2.10.

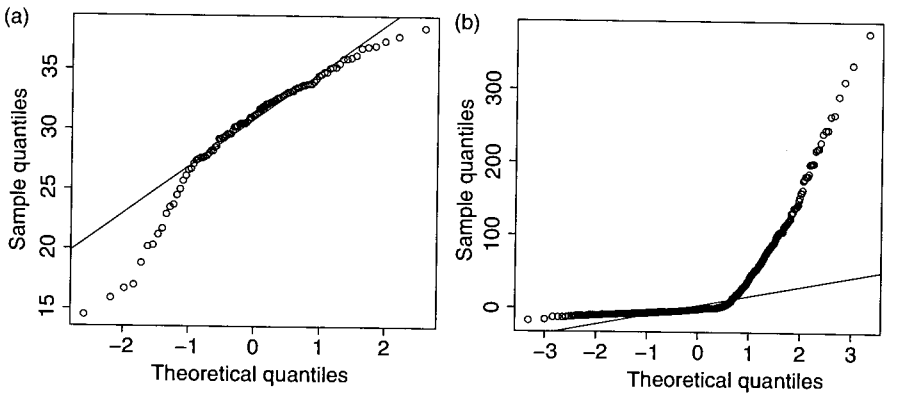

FIGURE 2.10 (a) Normal quantile plot for average January temperatures in Washington state. (b) Normal quantile plot for flight delay times for United Airlines.

Remark Even for samples drawn from a normal distribution, the points on a normal quantile plot do not lie *exactly* on a straight line. See Exercise 14. ||

2.5 EMPIRICAL CUMULATIVE DISTRIBUTION FUNCTIONS

The *empirical cumulative distribution function* (ecdf) is an estimate of the underlying cumulative distribution function (page 363) for a sample. The empirical cdf, denoted

by $\hat{F}$, is a step function

$$\hat{F}(x) = \frac{1}{n}\big(\text{number of values} \le x\big),$$

where n is the sample size.

For instance, consider the set of values 3, 6, 15, 15, 17, 19, 24. Then, $\hat{F}(18) = 5/7$ since there are five data values less than or equal to 18.

More generally,

$$\hat{F}(x) = \begin{cases} 0, & x < 3, \\ 1/7, & 3 \le x < 6, \\ 2/7, & 6 \le x < 15, \\ 4/7, & 15 \le x < 17, \\ 5/7, & 17 \le x < 19, \\ 6/7, & 19 \le x < 24, \\ 1, & x \ge 18. \end{cases}$$

Figure 2.11 displays the empirical cdf for this example as well as the ecdf for a random sample of size 25 from the standard normal distribution. The graph of the cdf for the standard normal $\Phi(t)$ is added for comparison.

The empirical cumulative distribution function is useful for comparing two distributions. Figure 2.12 shows the ecdf's of beer consumption for males and females

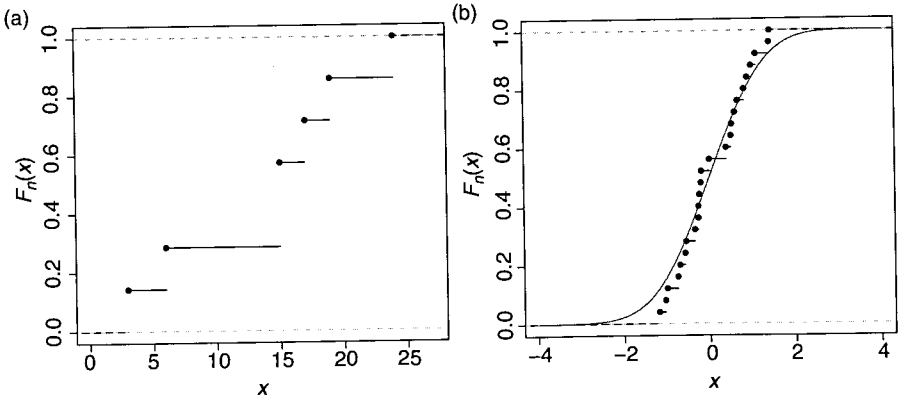

FIGURE 2.11 (a) Empirical cumulative distribution function for the data 3, 6, 15, 15, 17, 19, 24. (b) Ecdf for a random sample from $N(0, 1)$ with the cdf for the standard normal.

from the Beer and Hot Wings Case Study in Section 1.8. With the vertical line at 25 ounces, we can see that about 30% of the males and nearly 70% of the females have consumed 25 or fewer ounces of beer.

R Note:

The command `plot.ecdf` plots the empirical cumulative distribution function.

```
x <- c(3, 6, 15, 15, 17, 19, 24)
plot.ecdf(x)
x <- rnorm(25)                  # random sample of size 25 from N(0,1)
plot.ecdf(x, xlim = c(-4, 4))                # adjust x range
curve(pnorm(x), col = "blue", add = TRUE) # impose normal cdf
abline(v = 25, col = "red")                  # add vertical line
```

For the Beer and Hot Wings Case Study, we first create vectors that hold the data for the men and women separately.

```
beerM <- subset(Beerwings, select = Beer, subset = Gender == "M",
          drop = T)
beerF <- subset(beerwings, select = Beer, subset = Gender == "F",
          drop = T)
```

The `subset` command creates a new vector from the data set `Beerwings` by selecting the column `Beer` and extracting those rows corresponding to the males (`subset=Gender=="M"`) or females (`subset=Gender=="F"`). The `drop=T` argument ensures that we have a vector object (as opposed to a data frame).

```
plot.ecdf(beerM, xlab = "ounces")
plot.ecdf(beerF, col = "blue", pch = 2, add = TRUE)
abline(v = 25, lty = 2)
legend(c(5, .8), legend = c("Males", "Females"),
   col = c("black", "blue"), pch = c(19, 2))
```

In the last `plot.ecdf` command above, the `pch=2` changes the plotting character while the `add=TRUE` adds this plot to the existing plot.

2.6 SCATTER PLOTS

In the Beer and Hot Wings Case Study in Section 1.8, one question that the student asked was whether there was a relationship between the number of hot wings eaten and the amount of beer consumed. A way to visualize the relationship between two numeric variables is with a scatter plot (see Figure 2.13).

Each point in the scatter plot represents a single observation—that is, a single person who took part in the study. From the graph, we note that there is a positive,

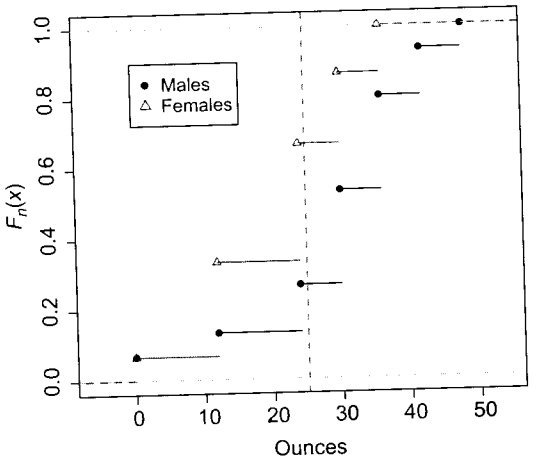

FIGURE 2.12 Ecdf's for male and female beer consumption. The vertical line is at 25 ounces.

roughly linear, association between hot wings and beer: as the number of hot wings eaten increases, the amount of beer consumed also increases.

Remark In statistics, the convention is to put the variable of primary interest on the y-axis, and the variable that may help predict or explain that variable as x, and to "plot y against x." ||

Further examples are shown in Figure 2.14. In general, when describing the relationship between two numeric variables, we will look for *direction*, *form*, and *strength*.

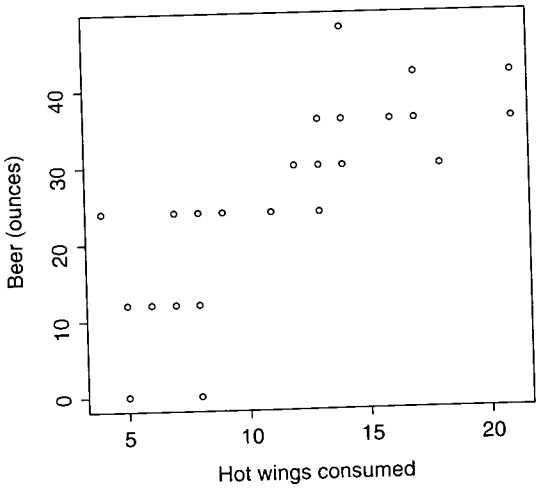

FIGURE 2.13 A scatter plot of `Beer` against `Hotwings`.

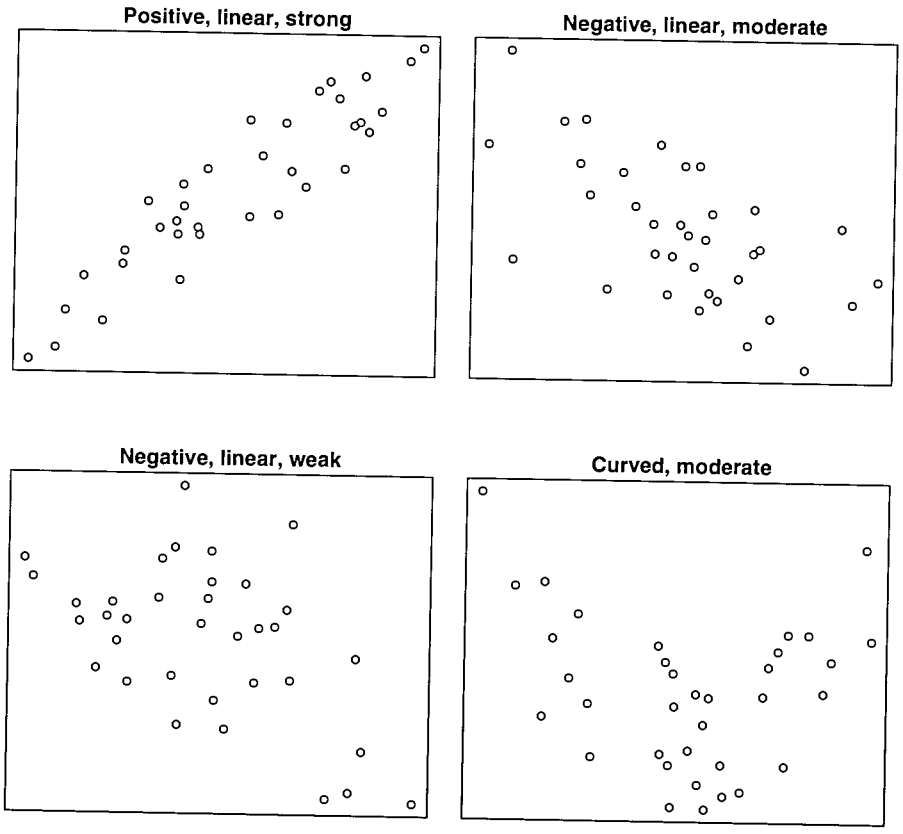

FIGURE 2.14 Examples of scatter plots.

In Chapter 9, we will investigate the relationship between two numeric variables in more detail.

R Note:

For scatter plots, use the `plot` command:

```
plot(Beerwings$Hotwings, Beerwings$Beer, xlab = "Hot wings eaten",
  ylab = "Beer consumed")
```

2.7 SKEWNESS AND KURTOSIS

Asymmetry and peakedness are often measured using *skewness* and *kurtosis*, which are defined using third and fourth central moments (Section A.7).

Definition 2.2 Let X be a random variable with mean μ and standard deviation σ. The *skewness* of X is

$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} \tag{2.3}$$

and the *kurtosis* of X is

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3. \tag{2.4}$$

||

A variable with positive skewness typically has a longer or heavier tail on the right than on the left; for negative skewness, the opposite holds. A variable with positive kurtosis typically has a higher central peak and a longer or heavier tail on at least one side than a normal distribution, while a variable with negative skewness is flatter in the middle and has shorter tails. Figure 2.15 shows some examples.

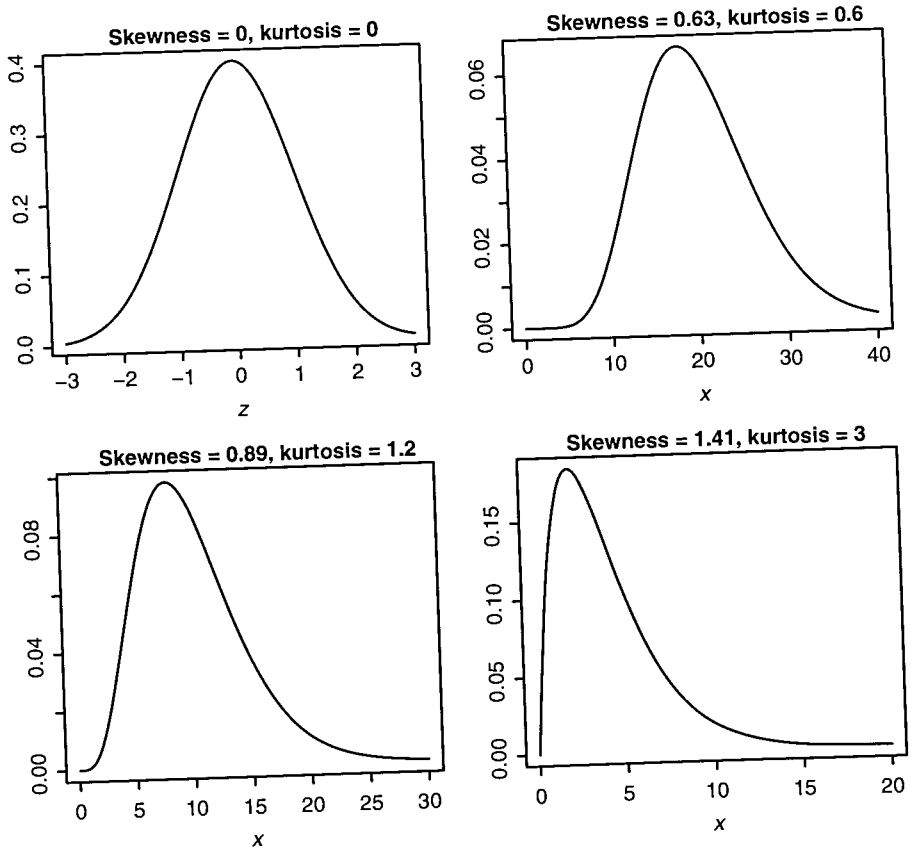

FIGURE 2.15 Examples of skewness and kurtosis for four distributions, including the standard normal (top left).

Example 2.12 Let Z be the standard normal variable with $\mu = 0$ and $\sigma = 1$. Then the skewness of Z is

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^3 e^{-z^2/2}\, dz = 0$$

and the kurtosis is

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^4 e^{-z^2/2}\, dz - 3 = 0. \qquad \square$$

Example 2.13 Let X be an exponential random variable with parameter $\lambda = 1$. Then $\mu = 1 = \sigma$, the skewness of X is

$$\int_0^{\infty} (x-1)^3 e^{-x}\, dx = 2,$$

and the kurtosis is

$$\int_0^{\infty} (x-1)^4 e^{-x}\, dx - 3 = 6. \qquad \square$$

Example 2.14 Let X be the standard uniform random variable, $f(x) = 1$ for $0 < x < 1$. Then $\mu = 0.5$, $\sigma^2 = 1/12$, the skewness is zero, and the kurtosis is

$$\frac{\int_0^1 (x-0.5)^4\, dx}{(\int_0^1 (x-0.5)^2\, dx)^2} - 3 = -1.2. \qquad \square$$

2.8 EXERCISES

1. Compute the mean $\bar{x}$ and median m of the six numbers 3, 5, 8, 15, 20, 21. Apply the logarithm to the data and then compute the mean $\tilde{x}$ and median $\tilde{m}$ of the transformed data. Is $\ln(\bar{x}) = \tilde{x}$? Is $\ln(m) = \tilde{m}$?
2. Compute the mean $\bar{x}$ and median m of the eight numbers 1, 2, 4, 5, 6, 8, 11, 15. Let $f(x) = \sqrt{x}$. Apply this function to the data and then compute the mean $\tilde{x}$ and the median $\tilde{m}$ of the transformed data. Is $f(\bar{x}) = \tilde{x}$? Is $f(m) = \tilde{m}$?
3. Let $\bar{x}$ and m denote the mean and median, respectively, of $x_1 < x_2 < \cdots < x_n$. Let f be a real-valued function.
   (a) Is $f(\bar{x})$ the mean of $f(x_1), f(x_2), \ldots, f(x_n)$?
   (b) Is $f(m)$ the median of $f(x_1), f(x_2), \ldots, f(x_n)$?
   (c) Are there any conditions that would ensure that $f(\bar{x})$ is the median of the transformed data?
   (d) Are there any conditions that would ensure that $f(m)$ is the median of the transformed data?
4. Import data from the Flight Delays Case Study in Section 1.1 data into R.
   (a) Create a table and a bar chart of the departure times (`DepartTime`).

   (b) Create a contingency table of the variables `Day` and `Delayed30`. For each day, what is the proportion of flights delayed at least 30 min?
   (c) Create side-by-side boxplots of the lengths of the flights, grouped by whether or not the flight was delayed at least 30 min.
   (d) Do you think that there is a relationship between the length of a flight and whether or not the departure is delayed by at least 30 min?
5. Import data from the General Social Survey Case Study in Section 1.6 into R.
   (a) Create a table and a bar chart of the responses to the question about the death penalty.
   (b) Use the `table` command and the `summary` command in R on the gun ownership variable. What additional information does the `summary` command give that the `table` command does not?
   (c) Create a contingency table comparing responses to the death penalty to the question about gun ownership.
   (d) What proportion of gun owners favor the death penalty? Does it appear to be different from the proportion among those who do not own guns?
6. Import data from the Black Spruce Case Study in Section 1.9 into R.
   (a) Compute the numeric summaries for the height changes (`Ht.Change`) of the seedlings.
   (b) Create a histogram and normal quantile plot for the height changes of the seedlings. Is the distribution approximately normal?
   (c) Create a boxplot to compare the distribution of the change in diameters of the seedlings (`Di.change`) grouped by whether or not they were in fertilized plots.
   (d) Use the `tapply` command to find the numeric summaries of the diameter changes for the two levels of fertilization.
   (e) Create a scatter plot of the height changes against the diameter changes and describe the relationship.
7. Let $x_1 < x_2 < \cdots < x_n$ and $y_1 < y_2 < \cdots < y_n$ be two sets of data with means $\bar{x}$, $\bar{y}$ and medians m_x, m_y, respectively. Let $w_i = x_i + y_i$ for $i = 1, 2, \ldots, n$.
   (a) Prove or give a counterexample: $\bar{x} + \bar{y}$ is the mean of $w_1, w_2, \ldots, w_n$.
   (b) Prove or give a counterexample: $m_x + m_y$ is the median of $w_1, w_2, \ldots, w_n$.
8. Find the median m and the first and third quartiles for the random variable X having
   (a) the exponential distribution with pdf $f(x) = \lambda e^{-\lambda x}$.
   (b) the Pareto distribution with parameter $\alpha > 0$ with pdf $f(x) = \alpha/x^{\alpha+1}$ for $x \geq 1$.
9. Let the random variable X have a Cauchy distribution with pdf $f(x) = 1/(\pi(1 + (x-\theta)^2))$ for $-\infty < x < \infty$. Show that θ is the median of the distribution.