Chapter 3

Graphing and Summarizing Data

Chapter Objectives

- Produce and interpret dot plots and histograms.
- Compute and interpret measures of center (mean, median, trimmed mean, geometric mean).
- Compute and interpret measures of variation (variance and standard deviation, interquartile range, median absolute deviation).
- Compute and interpret measures of skewness (coefficient of skewness and quartile skew coefficient).
- For a given data set, decide which measures of center, variation, and skewness are appropriate.
- Produce and interpret box plots.

Key Takeaways

- Histograms and dot plots reveal the shape, center, and spread of the distribution of a variable's values in a random sample, and these features can be inferred in the population.
- Statistics are used to summarize various features of a data set, including its center, variation, and skewness, and to infer these features in the population.
- There is usually more than one way to summarize a given feature of a data set, and the choice will depend on whether the data have outliers and whether their distribution is skewed.
- Boxplots are useful for comparing samples from two or more populations side by side in the same graph.

3.1 Introduction

The amount of information in large data sets can be overwhelming, so when analyzing a data set we usually prefer to just focus on its key features such as any overall patterns of variation or relationships between variables. To *explore* and then *communicate* these aspects of the data, we use graphical displays and numerical summaries.

One goal of graphing or summarizing data is to describe the *distribution* of the variable in the data set. The ***distribution*** of a variable refers to the range of values that the variable takes and the frequencies with which it takes those values. One way to represent a distribution is in a ***frequency distribution table*** showing how many times each value appears in the data set, as illustrated in the next example.

Example 3.1: Distribution of a Variable

A researcher at the Connecticut Agricultural Experiment Station took a random sample of $n = 150$ leaves from McIntosh apple trees on July 18, 1951, and counted the number of European red mites

on each leaf [4], [6]. The data are below.

0	2	4	0	0	0	0	1	0	0	0	2	0	0	0	0	0	1	1	0	0	1	0	1
0	0	0	0	0	1	3	1	1	1	1	2	1	2	0	2	0	0	1	0	1	0	5	5
0	1	1	0	0	0	6	1	4	2	3	1	0	0	1	0	4	0	0	2	0	3	1	0
0	2	2	3	0	4	1	2	0	0	0	0	1	0	4	0	1	0	0	2	2	0	0	3
1	0	0	2	5	1	2	4	1	1	2	1	0	1	0	3	0	0	4	0	1	0	1	2
0	4	1	1	0	1	2	3	0	1	0	1	0	0	0	6	0	0	7	3	1	0	0	3
1	0	4	3	0	1																		

The distribution of the variable (number of mites on a leaf) is represented by the following *frequency distribution table*.

Value of the variable	0	1	2	3	4	5	6	7
Number of leaves (frequency)	70	38	17	10	9	3	2	1

The table shows that most leaves (108 out of 150) have either no mites at all or just one, but a few have as many as six or seven. The distribution could also be represented by this *relative frequency distribution table*:

Value of the variable	0	1	2	3	4	5	6	7
Proportion of leaves (relative frequency)	0.47	0.25	0.11	0.07	0.06	0.02	0.01	0.01

Notice that the frequencies in the first table sum to 150 (the total sample size), and the relative frequencies (proportions) in the second sum to one.

A variable's distribution can also be represented a few different ways in a graph, as will be seen in the next section.

3.2 Graphing Data

Two useful graphs for displaying distributions are *dot plots* and *histograms*.

3.2.1 Dot Plots

Dot plots are especially useful for displaying small to moderate sized data sets containing only a handful of distinct values, some of which may appear multiple times.

Creating a Dot Plot: To construct a dot plot,

1. Draw a horizontal axis that spans the range of values in the data set.
2. For each observation in the data set, place a dot just above the horizontal axis at that value. If two or more observations have the same value, stack the dots. Rounding the data before plotting them can result in more observations having the same value, which sometimes enhances the appearance of the plot.

Example 3.2: Dot Plots

A dot plot of the data on counts of mites on apple tree leaves from Example 3.1 is below.

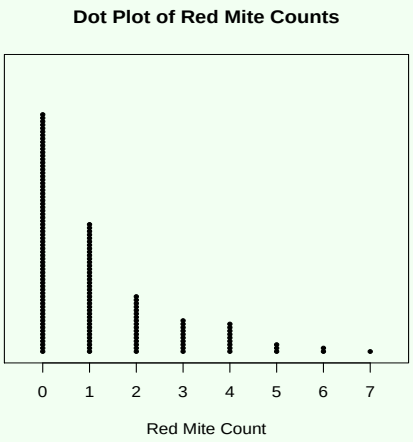

Figure 3.1: Dot plot of counts of European red mites on apple tree leaves.

In the plot, there are 70 dots over the value 0, 38 over the value 1, etc., reflecting the frequencies given in the first table of Example 3.1.

Dot plots are useful for comparing the distributions of a variable in samples drawn from two or more populations, as in the next example.

Example 3.3: Dot Plots

A report by the U.S. Environmental Protection Agency to the U.S. Congress gave mercury concentrations (in μg/g wet fish weight) for three categories of fish: freshwater fish, marine finfish, and marine shellfish [1]. The reported values are shown in the tables below.

Mercury in Freshwater Fish

Fish	Mercury (μg/g)
Bass	0.157
Bloater	0.093
Bluegill	0.033
Smallmouth Buffalo	0.096
Carp, Common	0.093
Catfish (channel, largemouth, rock, etc.)	0.088
Crappie (black, white)	0.114
Fresh-water Drum	0.117
Northern Squawfish	0.330
Northern Pike	0.127
Perch (white and yellow)	0.110
Sauger	0.230
Sucker (bridgelip, carpsucker, klamath, etc.)	0.114
Trout (brown, lake, rainbow)	0.149
Walleye	0.100

Mercury in Marine Finfish

Fish	Mercury (μg/g)	Fish	Mercury (μg/g)
Anchovy	0.047	Pompano	0.104
Barracuda, Pacific	0.177	Porgy	0.522
Cod	0.121	Ray	0.176
Croaker, Atlantic	0.125	Salmon	0.035
Eel, American	0.213	Sardines	0.100
Flounder	0.092	Sea Bass	0.135
Haddock	0.089	Shark	1.327
Hake	0.145	Skate	0.176
Halibut	0.250	Smelt, Rainbow	0.100
Herring	0.013	Snapper	0.250
Kingfish	0.100	Sturgeon	0.235
Mackerel	0.081	Swordfish	0.950
Mullet	0.009	Tuna	0.206
Ocean Perch	0.116	Whiting (silver hake)	0.041
Pollack	0.150		

Mercury in Marine Shellfish

Fish	Mercury (μg/g)
Abalone	0.016
Clam	0.023
Crab	0.117
Lobster	0.232
Oysters	0.023
Scallop	0.042
Shrimp	0.047

Dot plots of these three data sets are shown below using a common horizontal scale to facilitate comparisons.

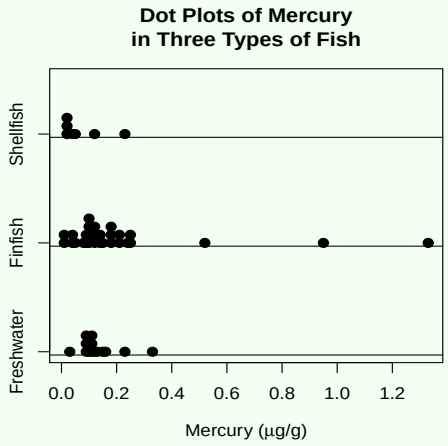

Figure 3.2: Dot plots of mercury concentrations (μg/g) in three types of fish: marine shellfish (top), marine finfish (middle), and freshwater fish (bottom).

We see from the dot plots that with the exception of the three extreme observations in the finfish data set, mercury concentrations in freshwater fish and marine finfish tend to be similar, but

concentrations in shellfish tend to be lower than in the other two fish types.

3.2.2 Histograms

For large data sets, the dots in a dot plot would need to be made so small to fit into the plot that they would be tiny specks. For such data sets *histograms* are preferred because the vertical scale on the graph can easily be adjusted to accommodate any number of observations. Histograms are a favorite among statisticians for graphing the distribution of a variable.

Creating a Histogram: To construct a histogram,

1. Choose the number of ***class intervals*** (or ***bins***). Usually 5-15 works well for small data sets. More than 15 is better for larger data sets.
2. Determine the class interval width:

$$\text{Class interval width} \approx \frac{\text{Largest observation} - \text{Smallest observation}}{\text{Number of class intervals}}$$

   Adjust the class interval width (round) and position the interval endpoints at convenient values (such as multiples of five). The leftmost interval should extend below the smallest observation and the rightmost above the largest. Each observation in the data set should fall into one of the intervals.
3. Determine the frequency (number of observations) or relative frequency (proportion of observations) for each class interval. Observations falling on the borderline between two class intervals should be placed in the upper interval.
4. Mark the class interval endpoints on a horizontal axis and place a bar over each interval, with bar height equal to the frequency (or relative frequency) for that interval.

When making a histogram, it's useful to first construct a frequency distribution table showing the class intervals and their frequencies, as in the next example.

Example 3.4: Histograms

Mercury contamination in fish is a serious concern. Citizens of a group of islands in the Indian Ocean called the Republic of Seychelles are among those who consume the most fish in the world (80-100 kg per person per year, which translates to more than half a pound per person per day), much of it predatory species. The following data are observations of the mercury content (in ppm) in the hair of 40 fishermen in the Seychelles [11].

13.3	32.4	18.1	58.2	64.0	68.2	35.4	33.9	23.9	18.3
22.1	39.1	31.4	18.5	21.0	5.5	7.9	5.2	28.7	26.3
13.9	25.9	9.8	26.9	16.8	37.7	19.6	21.8	31.6	30.1
42.4	16.5	21.2	33.0	9.8	10.6	29.6	40.7	12.9	13.8

We'll make a histogram of these data using seven class intervals. Since the data range from 5.2 to 68.2 ppm, the class interval width is

$$\text{Class interval width} = \frac{68.2 - 5.2}{7} = 9,$$

which we round up to 10 for convenience. The table below shows the class intervals and their frequencies (and relative frequencies).

Class Interval	Frequency	Relative Frequency
$[0, 10)$	5	0.125
$[10, 20)$	11	0.275
$[20, 30)$	10	0.250
$[30, 40)$	9	0.225
$[40, 50)$	2	0.050
$[50, 60)$	1	0.025
$[60, 70)$	2	0.050
	40	1.000

Note that the frequencies sum to the total sample size ($n = 40$) and the relative frequencies sum to one. the histogram, shown below, is obtained by marking the class interval endpoints on a horizontal axis and placing over each interval a bar whose height is equal to the frequency (or relative frequency) of the interval.

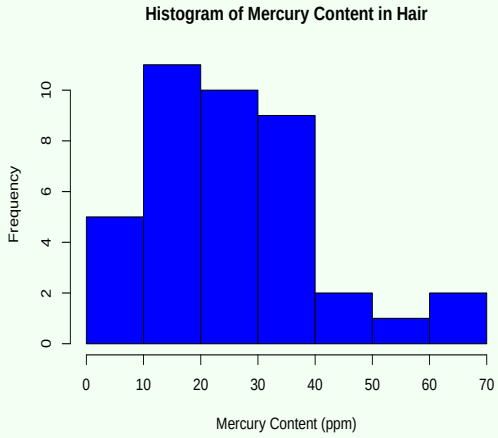

Figure 3.3: Histogram of mercury (ppm) in the hair of Seychelles fishermen.

Comment: A histogram that uses too few class intervals might conceal important information in the data. One that uses too many intervals might reveal information that's too detailed. Usually between about 5 and 30 bars is appropriate.

Example 3.5: Histograms

The histograms below are both made from the same hair mercury data that were used to make the one in Fig. 3.3. The histogram on the left has too few bars and the one on the right has too many.

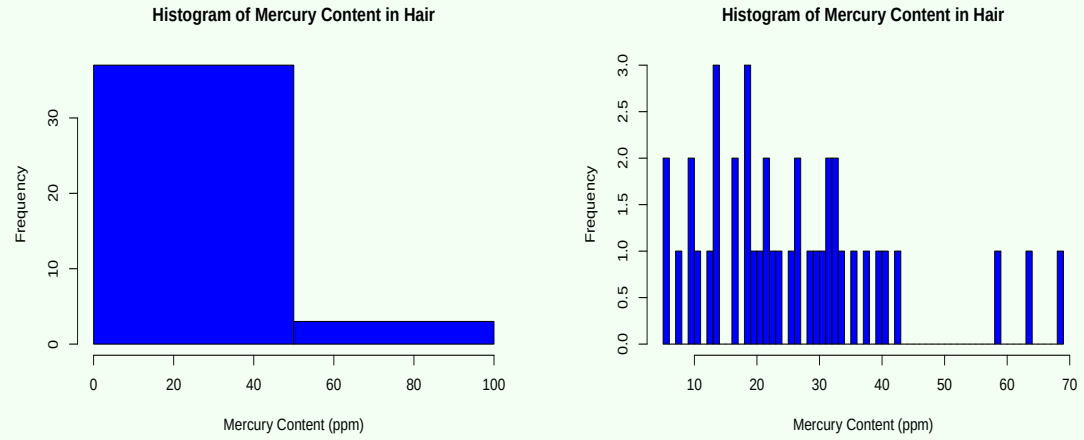

Figure 3.4: Histograms of mercury (ppm) in the hair of Seychelles fishermen with too few bars (left) and too many bars (right).

3.2.3 Interpreting Dot Plots and Histograms

A histogram or dot plot, by showing the distribution of values of a variable in a random sample, provides a glimpse of the variable's distribution in the population because we expect the two distributions to resemble each other when the sample is representative of the population. We're usually interested in just a few key features of the population, so it's useful to know what they are and how to examine them in a histogram or dot plot of the sample.

Among the things to look for in histograms and dot plots are:

1. **Shape**
   - ***Symmetric*** (left and right halves are "mirror images" of each other, usually in the form of a ***bell-shape***)
   - ***Right skewed*** (long "tail" extending to the right)
   - ***Left skewed*** (long "tail" extending to the left)
2. **Number of *modes*** (peaks)
   - ***Unimodal*** (having a single peak)
   - ***Bimodal*** (having two peaks)
   - ***Multimodal*** (having multiple peaks)
3. **Center** (the value of a "typical" observation)
4. **Spread** (the amount of variation in the observations)
5. **Outliers** (extreme observations lying outside the overall pattern, often corresponding to an individual from a population different from the one sampled)
6. **Other interesting features** (such as clumps of observations separated by large gaps, unusual or unexpected patterns, etc.)

The figure below illustrates some of these histogram shapes, and the ensuing examples contextualize some of the other features we're interested in when examining histograms and dot plots.

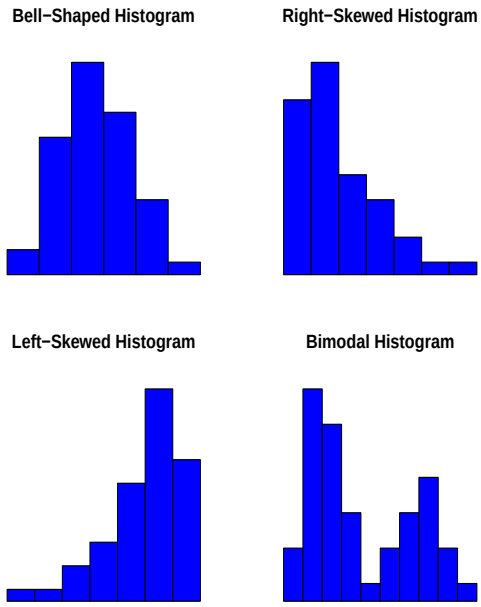

Figure 3.5: Histograms showing different shapes: symmetric and bell-shaped (top left), right skewed (top right), left skewed (bottom left), and bimodal (bottom right).

Example 3.6: Interpreting Dot Plots and Histograms

The histogram of hair mercury in Fig. 3.3 shows a slightly right skewed distribution that's essentially unimodal, with a mode (peak) at about 15 ppm. It's centered on 25, suggesting that a typical mercury concentration is about 25 ppm, but it's also quite spread out, indicating that there's considerable variation in the concentrations – they range from near zero to around 70 ppm.

Example 3.7: Interpreting Dot Plots and Histograms

The dotplot of red mite counts in Fig. 3.1 of Example 3.2 is severely right skewed with a mode (peak) at zero. A typical count is around one or two, and the counts vary over the range from zero to seven.

When unusual and unexpected features show up in a graph of data, effort should be made to identify their causes. For example, an outlier in laboratory data might be a contaminated specimen or the result of lab equipment malfunction. Separate mounds in a histogram of ecological data might correspond to different habitat types, different environmental conditions, or distinct animal populations or plant species. Unexpected variation in soil quality data might reflect disparity across sampling sites in human influences or natural ones (or both).

Example 3.8: Interpreting Dot Plots and Histograms

The two rightmost outliers in the middle dot plot of Fig. 3.2 correspond to shark and swordfish. Both have dangerously high mercury levels – the U.S. Food and Drug Administration's limit for human consumption is 1 ppm, or equivalently, 1 μg/g. One way fish accumulate mercury is by eating other fish, and the cited EPA report placed these two fish in a separate category of finfish

(along with barracuda) because, to quote,

> These are predatory, highly migratory species that spend much of their lives at the high end of marine the food web. These fish are large and accumulate higher concentrations of mercury than do lower trophic level, smaller fish.

3.3 Summarizing Data

3.3.1 Introduction and Notation

Recall that a ***statistic*** is any numerical quantity calculated from a set of random sample data. One use of statistics is to summarize the data set's essential information using just a few numerical values, called ***summary statistics***. At the very least, the summary should indicate the *center* of the distribution of values in the data set, representing a typical value, and the *spread* of the distribution, representing the amount of variation. It's sometimes also useful to include a statistic that indicates the *shape* of the distribution, and in particular its degree of *skewness*.

Throughout this manuscript, we'll denote the observations in a numerical data set of size $\boldsymbol{n}$ by

$$\boldsymbol{X_1, X_2, \ldots, X_n}.$$

The subscripts on the X's merely distinguish one observation from the next, and usually correspond to the distinct individuals, items, or specimens that make up the sample. Thus X_1 is the value of the variable for the first individual in the sample, X_2 the value for the second individual, and so on.

3.3.2 Measures of Center

We'll look at four statistics all of which measure the center of a data set:

1. The sample mean
2. The sample median
3. The trimmed mean
4. The geometric mean

The choice will depend largely on the shape of the distribution and whether there are outliers. We'll see how to decide which one to use for a given data set after first looking at how they're computed and interpreted.

The Sample Mean

The ***sample mean***, denoted $\bar{\boldsymbol{X}}$, is the most commonly used measure of center and is defined as follows.

Sample Mean: For data $X_1, X_2, \ldots, X_n$, the sample mean is

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

Thus $\bar{X}$ is just the arithmetic average of the observations in the data set. Its computation is illustrated in Example 3.9.

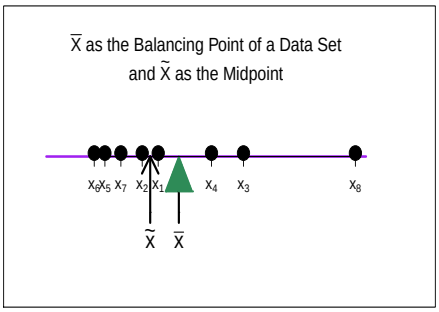

Figure 3.6: The sample mean $\bar{X}$ shown as the point where the data would balance if they were the positions of weights on a (weightless) horizontal axis, and the sample median $\tilde{X}$ shown as a value separating the smallest 50% of the data from the largest 50%.

The sample mean can be interpreted as the "balancing point" or "center of mass" of the data set in the sense that if equal weights were placed at positions $X_1, X_2, \ldots, X_n$ along a (weightless) horizontal axis, they'd balance on a fulcrum at the point $\bar{X}$. See Fig. 3.6.

If data are a random sample from a population, then because samples tend to be representative of populations, we can use $\bar{X}$ as an *estimate* of the ***population mean***, which is denoted by $\boldsymbol{\mu}$. We'll see how to gauge how far off the mark the estimate might be in Chapter 6.

Properties of $\bar{X}$: For any set of observations $X_1, X_2, \ldots, X_n$, the sample mean satisfies the following properties.

1. Changing the measurement scale of the data results in the same change of scale for the mean. More specifically, if we make a ***linear transformation*** of the X_i's, that is, if for some constants a and b we compute
$$Y_i \;=\; aX_i + b$$
for each $i = 1, 2, \ldots, n$, then the mean of $Y_1, Y_2, \ldots, Y_n$ is
$$\bar{Y} \;=\; a\bar{X} + b.$$

2. The ***deviations*** $X_i - \bar{X}$ of the observations away from the mean sum to zero. That is, for any set of data,
$$\sum_{i=1}^{n}(X_i - \bar{X}) \;=\; 0.$$

The first of the above two facts implies, for example, that if $X_1, X_2, \ldots, X_n$'s are temperatures in degrees Celsius with mean, say, $\bar{X} = 24$, and we convert each temperature to Fahrenheit via the transformation $Y_i = (9/5)X_i + 32$, then the mean of the Fahrenheit temperatures $Y_1, Y_2, \ldots, Y_n$ will be $\bar{Y} = (9/5)(24) + 32 = 75.2$. The second fact holds because the above average and below average data values cancel each other out when the deviations are summed. Both facts can be verified using the properties of summations given in the appendix.

The Sample Median

There's no guarantee that half of the observations will be less than the mean and the other half greater. For example, in Fig. 3.6, five observations are below $\bar{X}$ and only three above it. The *sample median*, on the other hand, is a statistic which *does* have this property. For example, in Fig. 3.6, four observations are below the median and four above it.

The ***sample median***, also called the ***50th percentile*** of the sample, is denoted by $\tilde{\boldsymbol{X}}$ and defined as follows.

Sample Median: For data $X_1, X_2, \ldots, X_n$, after putting them *in order* from smallest to largest, the sample median is

$$\tilde{X} = \begin{cases} \text{The } (\frac{n+1}{2})\text{th ordered value if } n \text{ is odd.} \\ \text{The average of the } (\frac{n}{2})\text{th and the } (\frac{n+2}{2})\text{th ordered values if } n \text{ is even.} \end{cases}$$

For example, if n is five, the median is the third among the ordered data values, but if n is six, it's the average of the third and fourth values. The median separates the smallest 50% of the data from the largest 50%. (If n is odd, the median is one of the data values and is considered to be "half" in the smallest 50% of the data and "half" in the largest 50%.) Like the mean, the median of a linear transformation of the data set is equal to that same linear transformation of the original median. Computation of $\tilde{X}$ is illustrated in Example 3.9.

The Trimmed Mean

A ***trimmed mean*** of a sample is denoted $\bar{\boldsymbol{X}}_{\text{tr}}$ and defined as follows.

Trimmed Mean: For data $X_1, X_2, \ldots, X_n$, after putting them *in order* from smallest to largest, the trimmed mean is

$\bar{X}_{\text{tr}}$ = The mean of the remaining observations after discarding a selected number of observations from *both ends* of the ordered data.

The ***trimming percentage*** of a trimmed mean is the percentage of observations deleted from *each* end of the ordered list (same percentage deleted from each end). Thus a trimmed mean with trimming percentage 10% is obtained by eliminating the smallest 10% of the data *and* the largest 10%, and then calculating the mean of the remaining observations. Good choices for a trimming percentage are values between 5% and 25%. Computation of $\bar{X}_{\text{tr}}$ is illustrated in Example 3.9.

Comment: If the trimming percentage is 0% (that is, you don't trim *any* observations), then the trimmed mean is the same as the usual sample mean. At the other extreme, if the trimming percentage is close enough to 50% that you trim all but the single middle value (when n is odd) or two middle values (when n is even), then the trimmed mean is the same as the sample median. In this sense, a trimmed mean can be thought of as a "compromise" between the sample mean and sample median, and the value of $\bar{X}_{\text{tr}}$ will usually lie between the values of $\bar{X}$ and $\tilde{X}$.

The Geometric Mean

Before defining the *geometric mean*, we'll need to review *logarithmic scales* and *logarithmic transformations*. We begin with the *base-10* logarithmic scale and transformation, and then turn to the *base-e*, or *natural* logarithmic scale and transformation.

Recall that a ***base-10 logarithmic*** (or ***base-10 log***) ***scale*** is a measurement scale on which each increase of one unit corresponds to a *tenfold* increase in the quantity being measured. Two familiar base-10 log scales are the Richter scale (for measuring earthquake magnitudes) and the pH scale (for measuring acidity or alkalinity of a substance). An earthquake whose magnitude is 5.0 on the Richter scale is ten times as strong as one whose magnitude is 4.0, and a substance whose pH is 9.0 is ten times more alkaline than one whose pH is 8.0.

The ***base-10 logarithmic*** (or ***base-10 log***) ***transformation*** of a non-negative variable X is a conversion

$$Y = \log_{10}(X)$$

of X to a variable Y that satisfies

$$X = 10^Y. \tag{3.1}$$

Notice from (3.1) that each one-unit increase in Y corresponds to a tenfold increase in X, so Y is measured on a base-10 log scale.

The ***natural*** (or ***base-e***) ***logarithmic scale*** is similar to the base-10 log scale, but the ***exponential constant***

$$e = 2.718282\ldots$$

takes the place of 10. Thus a one-unit increase on the natural log scale corresponds to an increase by a multiplicative factor e in the quantity being measured.

The ***natural*** (or ***base-e***) ***logarithmic transformation*** of a non-negative variable X is the conversion

$$Y = \log(X) \tag{3.2}$$

of X to a variable Y that satisfies

$$X = e^Y. \tag{3.3}$$

We see from (3.3) that for each one-unit increase in Y, X increases by the multiplicative factor e, so Y is measured on a natural log scale.

The right side of (3.3), $\boldsymbol{e^Y}$, is sometimes called the ***antilog*** of Y. Together, expressions (3.2) and (3.3) show us how to convert back and forth between two measurement scales: the original measurement scale of a variable X and the natural log scale of Y. For example, if X is measured in inches, then (3.2) converts X to log inches, and if Y is a value measured in log inches, then taking the antilog as in (3.3) converts Y back to inches.

Properties of the Log: The natural log $Y = \log(X)$ of a variable X has the following properties:

1. Each one-unit increase in Y corresponds to an increase in X by a multiplicative factor e, so Y is measured on a natural log scale.
2. $\log(1) = 0$ (since $X = 1$ in (3.3) implies $Y = 0$).
3. $\log(0) = -\infty$ (since $X = 0$ in (3.3) implies $Y = -\infty$).
4. If $0 < X < 1$ then $-\infty < \log(X) < 0$.
5. $\log(X)$ is *not defined* for $X < 0$ (since there is no value for Y that satisfies (3.3) when $X < 0$).
6. $\log(e) = 1$ (since $X = e$ in (3.3) implies $Y = 1$).
7. The natural log is an *order preserving* transformation, meaning that if $X_1 < X_2$ then $\log(X_1) < \log(X_2)$.

The Log of a Product: The natural log has the following property:

8. The natural log of a product equals the sum of the natural logs. Thus if $X_1 > 0$ and $X_2 > 0$,

$$\log(X_1 X_2) \;=\; \log(X_1) + \log(X_2).$$

To see, notice that if $Y = \log(X_1 X_2)$, then by (3.3), Y is the value satisfying $X_1 X_2 = e^Y$. But if we define Y_1 and Y_2 as $Y_1 = \log(X_1)$ and $Y_2 = \log(X_2)$, then by (3.3), $X_1 = e^{Y_1}$ and $X_2 = e^{Y_2}$, so $X_1 X_2 = e^{Y_1} e^{Y_2} = e^{Y_1 + Y_2}$. It follows that the value Y satisfying $X_1 X_2 = e^Y$ is $Y = Y_1 + Y_2$, which is to say $\log(X_1 X_2) = \log(X_1) + \log(X_2)$.

The Log of an Exponentiated Variable: The natural log has the following property:

9. For *any* constant a,

$$\log(X^a) \;=\; a \log(X).$$

For intuition, when a is a nonnegative integer, $X^a = XX \cdots X$ (a X's multiplied together), and by Property 8 (which extends to more than two X's), $\log(XX \cdots X) = \log(X) + \log(X) + \ldots + \log(X) = a \log(X)$.

The ***geometric mean*** of a sample is denoted **GM** and defined as follows.

Geometric Mean: For non-negative data $X_1, X_2, \ldots, X_n$, the geometric mean is

$$\text{GM} \;=\; e^{\frac{1}{n}\sum_{i=1}^n Y_i} \;=\; e^{\bar{Y}}, \tag{3.4}$$

where $Y_i = \log(X_i)$ is the natural log of X_i for each $i = 1, 2, \ldots, n$, and $\bar{Y}$ is the sample mean of $Y_1, Y_2, \ldots, Y_n$.

Thus the geometric mean is the antilog of the mean of the logs of the data. Computation of GM is illustrated in Example 3.9.

It can be shown, using the properties of log transformations listed above, that an equivalent expression for the geometric mean is as follows.

Geometric Mean (Alternative Formula): For non-negative data $X_1, X_2, \ldots, X_n$, the geometric mean can be calculated as

$$\text{GM} \;=\; (X_1 X_2 \cdots X_n)^{\frac{1}{n}} \;=\; \sqrt[n]{\prod_{i=1}^n X_i}\,. \tag{3.5}$$

The notation $\prod_{i=1}^n X_i$ is shorthand for the product $X_1 X_2 \cdots X_n$, and the notation $\sqrt[n]{}$ means the nth root, which is the same as raising to the power $1/n$.

The geometric mean is used as a measure of center for data sets that have right skewed distributions. For such data sets, GM usually approximately equals the sample median. We'll see why in Subsection 4.5.2 of Chapter 4.

The geometric mean is used also for averaging multiplicative factors. To illustrate, consider two successive dilutions of polluted water by mixing in clean water. If the original water contains 1,000 ppm of the pollutant, and we first dilute it to 10% of its original concentration and then to 20% of that, the final concentration is $1{,}000 \times 0.1 \times 0.2 = 20$ ppm. The geometric mean gives a more meaningful average of these two dilution factors than the usual mean.

To see, suppose we averaged them using the *usual* mean. We'd get $(0.1 + 0.2)/2 = 0.15$. But diluting the polluted water successively to this concentration gives $1{,}000 \times 0.15 \times 0.15 = 22.5$ ppm, which is different

from the 20 ppm obtained using the original dilution factors. However, if we average them using the *geometric* mean, we get (from (3.5)) GM $= (0.1 \times 0.2)^{1/2} = \sqrt{0.1 \times 0.2} = 0.141$. Diluting the polluted water successively to this concentration gives $1{,}000 \times 0.141 \times 0.141 = 20$ ppm, the same as using the original dilution factors.

Calculating the Measures of Center

We now have four statistics for measuring the center of a data set. In the next example we'll compute all four and compare their values for one set of data. After the example, we'll address the question "Which one should be used?"

Example 3.9: Calculating Measures of Center

The following data, ordered from smallest to largest, are carbon dioxide (CO_2) emissions (in millions of metric tons) from fossil fuel combustion (commercial, industrial, residential, transportation, and electric utilities) for each of the $n = 48$ continental United States in 2004, as reported by the U.S. Environmental Protection Agency.

CO_2 Emissions from Fossil Fuel Combustion

State	CO_2	State	CO_2	State	CO_2
Vermont	7	Utah	64	New Jersey	130
Rhode Island	11	Mississippi	65	Missouri	138
South Dakota	14	Kansas	77	Alabama	140
Idaho	16	Iowa	80	Kentucky	150
Delaware	17	Maryland	81	North Carolina	150
New Hampshire	22	Massachusetts	83	Georgia	174
Maine	23	Washington	85	Michigan	187
Montana	35	South Carolina	89	Louisiana	198
Oregon	43	Colorado	92	New York	216
Nebraska	43	Arizona	96	Indiana	233
Connecticut	45	Oklahoma	99	Illinois	236
North Dakota	46	Minnesota	100	Florida	256
Nevada	47	Wisconsin	107	Ohio	262
New Mexico	58	West Virginia	112	Pennsylvania	275
Arkansas	63	Tennessee	125	California	394
Wyoming	64	Virginia	127	Texas	688

A histogram of the data, below, shows a right skewed distribution.

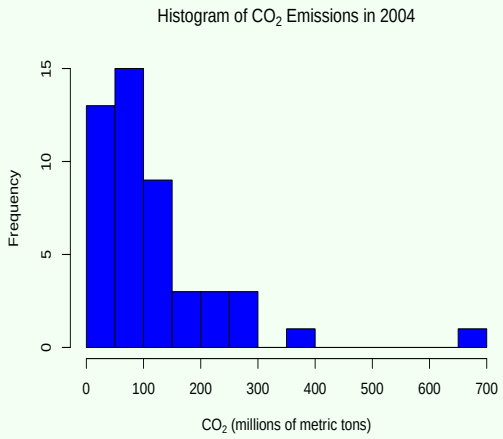

Figure 3.7: Histogram of carbon dioxide emissions for the 48 continental United States in 2004.

The sample mean is

$$\begin{aligned} \bar{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ &= \frac{1}{48}(7 + 11 + 14 + 16 + 17 + \ldots + 394 + 688) \\ &= 122.2. \end{aligned}$$

Since $n = 48$ is even, the sample median is the average of the 24th and 25th observations ($n/2 = 24$ and $(n+2)/2 = 25$) in the ordered list:

$$\begin{aligned} \tilde{X} &= \frac{89 + 92}{2} \\ &= 90.5. \end{aligned}$$

Trimming six observations off of each end of the ordered data amounts to a trimming percentage of 12.5% (since $6/48 = 0.125$). The mean of the remaining 36 observations is the 12.5% trimmed mean:

$$\begin{aligned} \bar{X}_{tr} &= \frac{1}{36}(23 + 35 + 43 + 43 + 45 + \ldots + 216 + 233) \\ &= 101.8. \end{aligned}$$

Finally, we'll use expression (3.4) to calculate the geometric mean. Taking the natural log of each observation (using statistical software) gives the following log-transformed values:

Log CO_2 Emissions from Fossil Fuel Combustion

State	Log CO_2	State	Log CO_2	State	Log CO_2
Vermont	1.9	Utah	4.2	New Jersey	4.9
Rhode Island	2.4	Mississippi	4.2	Missouri	4.9
South Dakota	2.6	Kansas	4.3	Alabama	4.9
Idaho	2.7	Iowa	4.4	Kentucky	5.0
Delaware	2.8	Maryland	4.4	North Carolina	5.0
New Hampshire	3.1	Massachusetts	4.4	Georgia	5.2
Maine	3.1	Washington	4.4	Michigan	5.2
Montana	3.6	South Carolina	4.5	Louisiana	5.3
Oregon	3.7	Colorado	4.5	New York	5.4
Nebraska	3.8	Arizona	4.6	Indiana	5.5
Connecticut	3.8	Oklahoma	4.6	Illinois	5.5
North Dakota	3.8	Minnesota	4.6	Florida	5.5
Nevada	3.9	Wisconsin	4.7	Ohio	5.6
New Mexico	4.1	West Virginia	4.7	Pennsylvania	5.6
Arkansas	4.2	Tennessee	4.8	California	6.0
Wyoming	4.2	Virginia	4.8	Texas	6.5

The sample mean of the log-transformed values is

$$\begin{aligned} \bar{Y} &= \frac{1}{48}(1.9 + 2.4 + 2.6 + 2.7 + 2.8 + \ldots + 6.0 + 6.5) \\ &= 4.41, \end{aligned}$$

and its antilog is the geometric mean:

$$\text{GM} = e^{4.41} = 82.3.$$

The four measures of center are shown below for comparison:

Statistic	Value
Sample mean $\bar{X}$	122.2
Sample median $\tilde{X}$	90.5
Trimmed mean $\bar{X}_{tr}$	101.8
Geometric mean GM	82.3

The mean is the larger than the median, which is typical for a right skewed distribution. The trimmed mean lies between the mean and the median, which is typical regardless of the shape of the distribution. The median and geometric mean are relatively close to each other, which is again typical for a right skewed distribution.

Comparison of the Measures of Center

The table below lets us compare the four measures of center to decide which one to use for a given set of data.

Statistic	Interpretation	Properties	Uses
Sample mean $\bar{X}$	Arithmetic average, "balancing point" of the data	Not resistant to outliers	Commonly used, except with skewed distributions and data with outliers
Sample median $\tilde{X}$	Middle value of the data, 50th percentile	Resistant to outliers	Commonly used, especially with skewed distributions and data with outliers
Trimmed mean $\bar{X}_{tr}$	Average of middle portion of data	Resistant to outliers	Rarely used, but occasionally with data with outliers
Geometric mean GM	Antilog of average of logs of data	Not resistant to outliers	Rarely used, but occasionally with skewed distributions or data that are multiplicative factors

A statistic is said to be ***resistant*** to outliers if its value isn't influenced by their presence in the data set. The sample mean can be *strongly influenced by outliers* – a single large outlier can inflate its value dramatically – so it's *not* resistant. The median and trimmed mean, however, are both resistant. The geometric mean isn't entirely resistant, but its value usually isn't severely affected by outliers when they're present. The following example illustrates.

Example 3.10: Resistance to Outliers

The outlier on the right side of the histogram in Example 3.9 is Texas, whose CO_2 emissions were 688 million tons. The four measures of center from that example are shown again below along with their recomputed values after reducing Texas' emissions to 394 million tons, the same level as the next highest state (California):

Statistic	Original Value	Recomputed Value	Change
Sample mean $\bar{X}$	122.2	116.1	$\downarrow$ 6.1
Sample median $\tilde{X}$	90.5	90.5	None
Trimmed mean $\bar{X}_{tr}$	101.8	101.8	None
Geometric mean GM	82.3	81.5	$\downarrow$ 0.8

Notice that the mean decreased upon diminishing the large outlier, but neither the median nor the trimmed mean changed. The geometric mean decreased, but only slightly.

The mean will typically be larger than the median when the data follow a right skewed distribution, but approximately equal to the median when they follow a symmetric one. To see why, recall that the mean is the "balancing point" of a data set and is inflated by large outliers, whereas the median is the "middle value" and is resistant to outliers. In a histogram, the bars would balance (approximately) along the horizontal axis at the mean if they were weights, and their total *area* would be split into equal halves at the median (the "equal areas point"). As seen in Fig. 3.8, the "balancing point" (mean) is also the equal-areas split point (median) for the symmetric distribution. But the "balancing point" is pulled rightward of the equal-areas point for the right skewed distribution by the observations in the right tail.

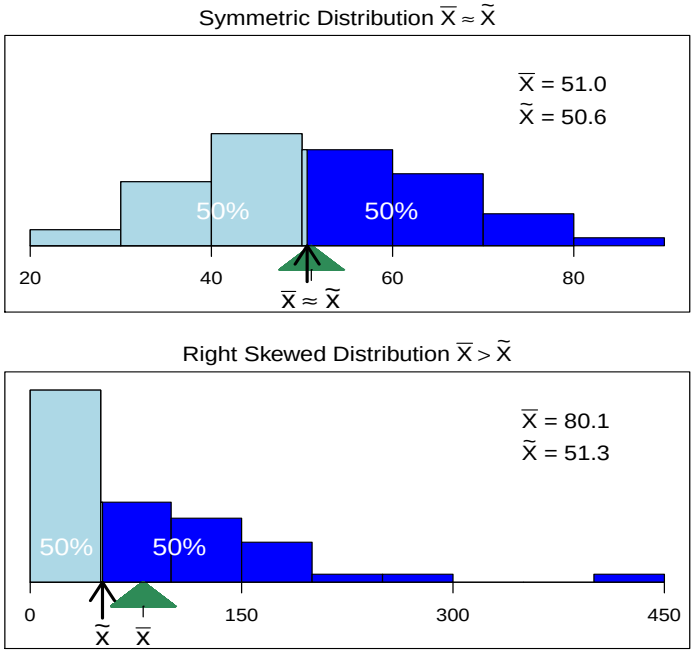

Figure 3.8: Histograms illustrating the relative positions of the sample mean $\bar{X}$ (green triangle) and median $\tilde{X}$ (light and dark blue areas split point). For a symmetric distribution (top), the mean and median are approximately equal. For a right skewed distribution (bottom), the mean is greater than the median.

Choosing a Measure of Center: The following guidelines can be used when choosing which measure of center to use for a given set of data.

- For data with a symmetric distribution and no outliers, the mean and median are *both representative* of a typical value. The mean is preferred because, as we'll see (in later chapters), it's more amenable to making inferences about the population.
- For data with a right skewed distribution, the median is *representative* of a typical value and is preferred over the mean, which is too large.
- For data with outliers, the median is *representative* of a typical value because it's resistant, so it's preferred over mean, which is influenced by the outliers.

3.3.3 Measures of Variation

We'll also be interested in summarizing the amount of variation, or dispersion, in a set of data. For example, if the data are temperatures, we'll want to summarize how much they fluctuate. If they're rainfall measurements, we may want to summarize how much they typically differ from their average. We'll look at four measures of variation in data:

1. The sample variance
2. The sample standard deviation
3. The interquartile range
4. The median absolute deviation

We'll also compare their properties, which will help us decide which one to use for a given set of data.

The Sample Variance and Sample Standard Deviation

The *sample variance* and *sample standard deviation* are based on the deviations $X_i - \bar{X}$ of the observations $X_1, X_2, \ldots X_n$ away from their mean $\bar{x}$.

The ***sample variance***, denoted $\boldsymbol{s^2}$, is defined as

Sample Variance: For data $X_1, X_2, \ldots, X_n$, the sample variance is

$$s^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2. \tag{3.6}$$

The variance is computed by "averaging" the squared deviations (using $n-1$ instead of n). Its value is interpreted as the size of a typical *squared* deviation away from the mean. It's measured in the *squared* units of the original data (for example inches squared if the data are rainfalls in inches), so it's not very useful. Instead, we take its square root, which gives the ***sample standard deviation***, denoted by $\boldsymbol{s}$:

Sample Standard Deviation: For data $X_1, X_2, \ldots, X_n$, the sample standard deviation is

$$s \;=\; \sqrt{s^2} \;=\; \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

The standard deviation is interpreted as the size of a typical deviation away from the mean, and is measured in the same units as the original data (inches if the data are in inches). It's the most commonly used measure of variation. Example 3.11 illustrates its computation.

If the data are a random sample from a population, we can use s^2 as an estimate of the ***population variance***, which is denoted $\boldsymbol{\sigma^2}$.

Some Properties of $\boldsymbol{s}$: For any set of observations $X_1, X_2, \ldots, X_n$, the sample standard deviation s satisfies the following properties.

1. $s = 0$ only when the observations are all the same value, that is, when there's no variation in their values. Otherwise $s > 0$, and the more the data values vary, the larger s will be.
2. If we change the measurement scale of the X_i's by making a linear transformation, that is,

   $$Y_i \;=\; aX_i + b \qquad \text{for } i = 1, 2, \ldots, n,$$

   then the standard deviation s_y of $Y_1, Y_2, \ldots, Y_n$ is

   $$s_y \;=\; |\,a\,|\,s,$$

   where $|\ |$ denotes the absolute value.

Comment: Why do we divide by $n-1$ instead of by n when we compute s^2 and s? It turns out that if we divided by n, the resulting statistic would tend to underestimate the true value σ^2. The reason for this is that the deviations $X_i - \bar{X}$ used to compute s^2 tend to be smaller than the deviations $X - \mu$ used to compute σ^2 because the value of $\bar{X}$ varies so as to always be "close" to (somewhere in the middle of) the sample values $X_1, X_2, \ldots, X_n$, but the value of μ is fixed. When we divide by $n-1$ instead of n, the resulting statistic s^2 no longer has the tendency to underestimate the true value.

The Interquartile Range

The ***sample interquartile range***, denoted $\boldsymbol{IQR}$, is defined as

Interquartile Range: For data $X_1, X_2, \ldots, X_n$, the interquartile range is

$$IQR = Q_3 - Q_1,$$

where Q_1 is the ***25th sample percentile*** (or ***first quartile***), and Q_3 is the ***75th sample percentile*** (or ***third quartile***), defined by

$$\begin{aligned} Q_1 &= \text{The median of the observations that are} \\ &\quad \text{less than or equal to the overall median } \tilde{X} \\ Q_3 &= \text{The median of the observations that are} \\ &\quad \text{greater than or equal to the overall median } \tilde{X} \end{aligned}$$

Together, the two quartiles and the median split the data into fourths, with Q_1 separating the smallest fourth from the rest, $\tilde{X}$ splitting the data into halves, and Q_3 separating the largest fourth from the rest. Thus the middle 50% of the data set lies between Q_1 and Q_3, and so the interquartile range measures the spread in the middle 50% of the data. Example 3.11 illustrates the computation of the IQR.

The Median Absolute Deviation

The ***median absolute deviation***, denoted ***MAD***, is defined as

Median Absolute Deviation: For data $X_1, X_2, \ldots, X_n$, the median absolute deviation is

$$MAD = \text{The median of } |X_1 - \tilde{X}|, |X_2 - \tilde{X}|, \ldots, |X_i - \tilde{X}|,$$

where $\tilde{X}$ is the median of the data set and $|\ |$ denotes the absolute value.

The median of the absolute deviation is the median of the absolute values of the deviations away from the sample median. It's interpreted as the size of a typical deviation away from the median. Example 3.11 illustrates its computation.

A Comparison of the Measures of Variation

In the next example, we'll compute each of the four measures of variation for a data set. After the example, we'll discuss how to decide which one to use for a given set of data.

Example 3.11: Comparison of Measures of Variation

The amount of solar radiation received at a greenhouse plays an important role in determining the rate of photosynthesis. The following are $n = 7$ observations on incoming solar radiation (in $MJ/m^2/d$) in one particular greenhouse [7].

8.4 8.8 9.0 10.2 10.7 11.2 11.9

The mean of the data is $\bar{X} = 10.03$. To compute the sample variance and standard deviation, we first calculate the squared deviations away from the mean, shown in the last column below.

X_i	Deviation $X_i - \bar{X}$	Squared Deviation $(X_i - \bar{X})^2$
8.4	$8.4 - 10.03 = -1.63$	2.66
8.8	$8.8 - 10.03 = -1.23$	1.51
9.0	$9.0 - 10.03 = -1.03$	1.06
10.2	$10.2 - 10.03 = 0.17$	0.03
10.7	$10.7 - 10.03 = 0.67$	0.45
11.2	$11.2 - 10.03 = 1.17$	1.37
11.9	$11.9 - 10.03 = 1.87$	3.50
		$\sum(X_i - \bar{X})^2 = 10.58$

The sum of the squared deviations, shown at the bottom of the last column, is 10.58. Thus, using (3.6), the sample variance is

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\ &= \frac{1}{6}(10.58) \\ &= 1.76, \end{aligned}$$

and the sample standard deviation is

$$\begin{aligned} s &= \sqrt{s^2} \\ &= \sqrt{1.76} \\ &= 1.33. \end{aligned}$$

Together the mean and standard deviation tell us that in this data set, a typical sunlight measurement is about 10.03 $MJ/m^2/d$, plus or minus 1.33 $MJ/m^2/d$.

Because n is odd, the median is just the middle value in the sorted list,

$$\tilde{X} = 10.2.$$

The first quartile Q_1 is the median of the observations that are less than or equal to 10.2. Thus $Q_1 = (8.8 + 9.0)/2 = 8.9$. Likewise, the third quartile Q_3 is the median of the observations that are greater than or equal to 10.2, so $Q_3 = (10.7 + 11.2)/2 = 10.95$. Therefore the interquartile range is

$$\begin{aligned} IQR &= Q_3 - Q_1 \\ &= 10.95 - 8.9 \\ &= 2.05. \end{aligned}$$

To compute the median absolute deviation, we first need the absolute values of the deviations away from the median, which are shown in the last column of the table below:

Observation X_i	Deviation $X_i - \tilde{X}$	Absolute Deviation $\lvert X_i - \tilde{X} \rvert$
8.4	$8.4 - 10.2 = -1.8$	1.8
8.8	$8.8 - 10.2 = -1.4$	1.4
9.0	$9.0 - 10.2 = -1.2$	1.2
10.2	$10.2 - 10.2 = 0.0$	0.0
10.7	$10.7 - 10.2 = 0.5$	0.5
11.2	$11.2 - 10.2 = 1.0$	1.0
11.9	$11.9 - 10.2 = 1.7$	1.7

The median absolute deviation is the median the absolute deviations,

$$MAD = 1.2.$$

Properties of the Measures of Variation

Because the variance and standard deviation are based on the squared deviations $(X_i - \bar{X})^2$, which can be very large if X_i is an outlier, they're *not resistant* to outliers. More precisely, outliers at either extreme of the data set will inflate the values of the variance and standard deviation. The interquartile range and median absolute deviation are both *resistant* to outliers, though. The next example illustrates.

Example 3.12: Resistance to Outliers

Suppose in Example 3.11 that the largest radiation reading was 17.4 (an outlier) instead of 11.9. The recomputed variance, standard deviation, interquartile range, and median absolute deviation are shown below, along with their original values (from Example 3.11) for comparison:

Statistic	Original value	Recomputed value after inserting the outlier
Sample variance	$s^2 = 1.76$	$s^2 = 9.51$
Sample standard deviation	$s = 1.33$	$s = 3.08$
Interquartile range	$IQR = 2.05$	$IQR = 2.05$
Median absolute deviation	$MAD = 1.20$	$MAD = 1.20$

We see that s and s^2 are inflated by the outlier, but the IQR and MAD are unchanged.

Choosing a Measure of Variation: The following guidelines can be used when choosing between the four statistics that summarize the variation in a data set:

- Use the standard deviation when the mean is used to summarize the center, that is, if there are no outliers and the distribution is roughly symmetric.
- Use the interquartile range or median absolute deviation if the median is used to summarize the center, that is, if there are outliers or the distribution is skewed (in either direction).
- The sample variance is rarely used because it's measured in the squared units of the data.

3.3.4 Measures of Skewness

Because environmental data often exhibit right skewed distributions, it's useful to be able to summarize the extent of the skewness with a statistic. We'll look at two measures of skewness:

1. The coefficient of skewness
2. The quartile skew coefficient

We'll also look their properties, which will guide us when deciding which one to use for a given set of data.

The Coefficient of Skewness

The ***coefficient of skewness***, denoted $\boldsymbol{g}$, is defined (for $n \geq 3$) as

Coefficient of Skewness: For data $X_1, X_2, \ldots, X_n$, the coefficient of skewness is

$$g = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \frac{(X_i - \bar{X})^3}{s^3}. \tag{3.7}$$

The coefficient of skewness is computed by calculating the deviations $X_i - \bar{X}$ away from the mean, cubing each one, dividing each cubed deviation by the cube of the standard deviation, summing the results, and multiplying by $n/((n-1)(n-2))$.

If we let

$$z_i = \frac{X_i - \bar{X}}{s},$$

we can write coefficient of skewness as

Coefficient of Skewness (Alternative Formula Using *Z*-Scores): For data $X_1, X_2, \ldots, X_n$, the coefficient of skewness is

$$g = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} z_i^3.$$

The quantities $z_1, z_2, \ldots, z_n$ are sometimes called ***z*-scores**. They indicate how many standard deviations their corresponding X_i values are above or below the mean. Example 3.13 illustrates the computation of the coefficient of skewness.

Some Properties of *g*: The following properties are useful for interpreting the value of g:

- For data whose distribution is symmetric, $g \approx 0$.
- For data whose distribution is right skewed, $g > 0$. The more right skewed the distribution is, the larger g will be.
- For data whose distribution is left skewed, $g < 0$. The more left skewed the distribution is, the larger in the negative direction g will be.

To get a feel for how the value of g reflects different degrees of skewness, the figure below shows several histograms with varying degrees of skewness along with the corresponding value of g.

Example 3.13: The Coefficient of Skewness

We'll compute the coefficient of skewness for the $n = 7$ mercury concentrations in marine shellfish from Example 3.3, shown again below ordered from smallest to largest.

0.016 0.023 0.023 0.042 0.047 0.117 0.232

A dot plot (below) reminds us that the distribution is right skewed, so we can expect to get a positive value for g.

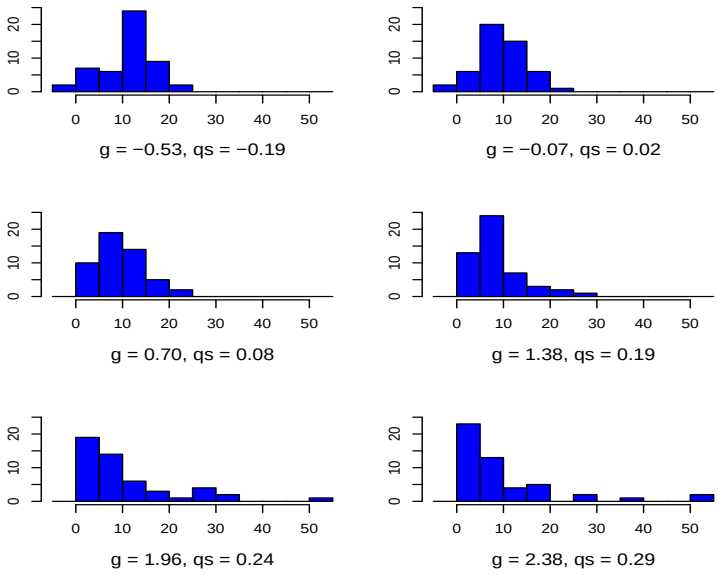

Figure 3.9: Distributions of data with varying degrees of skewness and the corresponding values of the coefficient of skewness g and quartile skew qs.

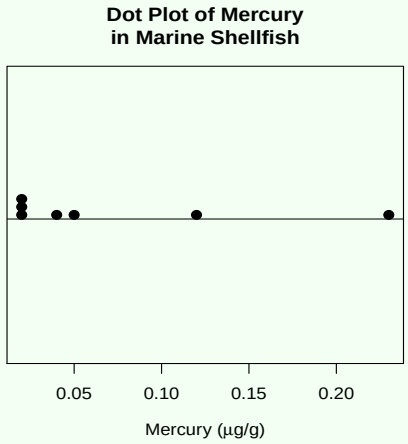

Figure 3.10: Dot plots of mercury concentrations (μg/g) in marine shellfish.

The sample mean and standard deviation are $\bar{X} = 0.071$ and $s = 0.079$. The cubed deviations away from the mean, divided by the cubed standard deviation, are given below.

X_i	$z_i^3 = \frac{(X_i-\bar{X})^3}{s^3}$
0.016	$\frac{(0.016-0.071)^3}{0.079^3} = -0.027$
0.023	$\frac{(0.023-0.071)^3}{0.079^3} = -0.018$
0.023	$\frac{(0.023-0.071)^3}{0.079^3} = -0.018$
0.042	$\frac{(0.042-0.071)^3}{0.079^3} = -0.004$
0.047	$\frac{(0.047-0.071)^3}{0.079^3} = -0.002$
0.117	$\frac{(0.117-0.071)^3}{0.079^3} = 0.016$
0.232	$\frac{(0.232-0.071)^3}{0.079^3} = 0.669$
	$\sum \frac{(X_i-\bar{X})^3}{s^3} = 0.616$

Their sum, shown at the bottom of the table above, is 0.616. Thus, using (3.7), we get the coefficient of skewness:

$$g = \frac{7}{(6)(5)}(0.616) = 0.144.$$

As expected, g is positive because the distribution of the data is right skewed.

Comment: How is it that g measures skewness? First note that the multiplier $n/((n-1)(n-2))$ and the divisor s^3 in (3.7) are always positive, so the sign of g is determined solely by that of the sum $\sum(X_i - \bar{X})^3$. A cubed deviation $(X_i - \bar{X})^3$ will be positive if the observation X_i is greater than the mean $\bar{X}$, and negative if it's less than the mean. The farther X_i is away from the mean in either direction, the larger the magnitude of its cubed deviation will be. If the distribution is right skewed, the X_i's in the long, right tail of the distribution will produce large positive cubed deviations, forcing $\sum(X_i - \bar{X})^3$, and therefore g, to be positive. If the distribution is left skewed, the X_i's in the left tail will produce large negative cubed deviations, forcing g to be negative. The more pronounced the skewness is in either direction, the farther g will be away from zero in that direction. On the other hand, if the distribution is symmetric, any positive cubed deviations will tend to be canceled out by negative ones of equal magnitude, leading to a g value near zero.

Comment: The multiplier $n/((n-1)(n-2))$ in the formula for g plays a role similar to the multiplier $1/(n-1)$ in the the sample variance, that is to "average" the (cubed) deviations. The exact form $n/((n-1)(n-2))$ is used because it makes g a more accurate estimator of the skewness in the population, correcting for any tendency to underestimate the true value.

The Quartile Skew Coefficient

The ***quartile skew coefficient***, denoted qs, is defined as

Quartile Skew Coefficient: For data $X_1, X_2, \ldots, X_n$, the quartile skew coefficient is

$$qs = \frac{(Q_3 - \tilde{X}) - (\tilde{X} - Q_1)}{IQR}, \tag{3.8}$$

where $\tilde{X}$ is the median and IQR is the interquartile range.

In words, the quartile skew coefficient is the difference between the spreads of the upper and lower quartiles away from the median, relative to the size of the IQR. Example 3.14 will illustrate its computation.

Some Properties of *qs*: The following properties are useful for interpreting the value of qs:

- For data whose distribution is symmetric, $qs \approx 0$.
- For data whose distribution is right skewed, $qs > 0$. The more right skewed the distribution is, the larger qs will be.
- For data whose distribution is left skewed, $qs < 0$. The more left skewed the distribution is, the larger qs will be in the negative direction.

Fig. 3.9 shows the value of qs for several histograms with varying degrees of skewness.

Example 3.14: The Quartile Skew Coefficient

We'll compute the quartile skew coefficient for the $n = 7$ mercury concentrations in marine shellfish from Examples 3.3 and 3.13, shown again below.

$$0.016 \quad 0.023 \quad 0.023 \quad 0.042 \quad 0.047 \quad 0.117 \quad 0.232$$

The median is $\tilde{X} = 0.042$ and the quartiles are $Q_1 = 0.023$ and $Q_3 = 0.082$. Thus the interquartile range is

$$IQR \;=\; 0.082 - 0.023 \;=\; 0.059,$$

so the quartile skew coefficient is

$$\begin{aligned} qs &= \frac{(Q_3 - \tilde{X}) - (\tilde{X} - Q_1)}{IQR} \\ &= \frac{(0.082 - 0.042) - (0.042 - 0.023)}{0.059} \\ &= 0.356. \end{aligned}$$

As expected, qs is positive because the distribution of the data is right skewed.

Comment: To see how the quartile skew qs measures skewness, first note that the IQR will always be positive, and so the sign of quartile skew is determined solely by the sign of the numerator in (3.8). If the distribution is right skewed, upper half of the data will be more spread out than the lower half, meaning that Q_3 will be farther above the median $\tilde{X}$ than Q_1 will be below it, so the numerator in (3.8) will be positive, and therefore the quartile skew will be positive too. By a similar argument, if the distribution is left skewed the numerator in (3.8), and thus qs, will be negative. If the distribution is symmetric, Q_1 and Q_3 will be approximately equidistant from the median, so the numerator of (3.8), and therefore the quartile skew, will be close to zero.

Some Properties of the Measures of Skewness

Because the coefficient of skewness is based on the cubed deviations $(X_i - \bar{X})^3$, which can be very large (in magnitude) if X_i is an outlier, it's *not resistant* to outliers. Outliers in the positive direction will inflate the value of the coefficient of skewness, and outliers in the negative direction will inflate it's value in the negative direction. The quartile skew coefficient, on the other hand, is *resistant* to outliers.

Choosing a Measure of Skewness: The following guidelines can be used when choosing a statistic to summarize the skewness of a set of data:

- Use the coefficient of skewness when the mean and standard deviation are used to summarize the center and variation, and in particular, when there are no outliers in the data.

- Use the quartile skew when the median and interquartile range or median absolute deviation are used to summarize the center and variation, and in particular, when there are outliers in the data.

3.4 Boxplots and the Five Number Summary

We can simultaneously summarize the center, variation, and skewness of a set of data using the ***five number summary***, defined as the smallest observation, first quartile, median, third quartile, and largest observation:

Five Number Summary: For data $X_1, X_2, \ldots, X_n$, the five number summary is

$$\text{Minimum}, \quad Q_1, \quad \tilde{X}, \quad Q_3, \quad \text{Maximum}.$$

A ***boxplot*** is a graphical display of the five number summary.

Creating a Boxplot: To construct a boxplot,

1. Draw a box next to a vertical axis, with bottom of the box bottom level with the first quartile, and the top of the box level with the third quartile.
2. Draw a horizontal line through the box level with the median.
3. Draw lines (called ***whiskers***) extending from the bottom of the box down to the minimum observation and from the top of the box up to the maximum.

Example 3.15: Boxplots and the Five Number Summary

On November 28, 2011 a spill of toxic materials from the Suncor Energy oil refinery north of Denver, Colorado was discovered seeping into Sand Creek, which flows into the South Platte River, the Denver area's main water source [3]. The Colorado Department of Public Health and Environment monitored the spill by measuring benzene concentrations (ppb) on $n = 13$ days at two locations, the confluence of Sand Creek with the South Platte River, and a location on the South Platte downstream from its confluence with Sand Creek. The data are below.

Benzene in South Platte River

Date	Benzene at Confluence with Sand Creek	Benzene Downstream from the Confluence
Dec. 27	640	190
Dec. 28	240	300
Dec. 29	140	130
Dec. 30	190	130
Dec. 31	170	160
Jan. 2	300	240
Jan. 3	730	250
Jan. 4	630	240
Jan. 5	650	240
Jan. 6	190	590
Jan. 7	310	260
Jan. 8	400	260
Jan. 9	720	240

The data from the first location, in sorted order, are

140 170 190 190 240 300 310 400 630 640 650 720 730

The five number summary is

Min	Q_1	$\tilde{X}$	Q_3	Max
140	190	310	640	730

and the boxplot (with five number summary labels) is shown below.

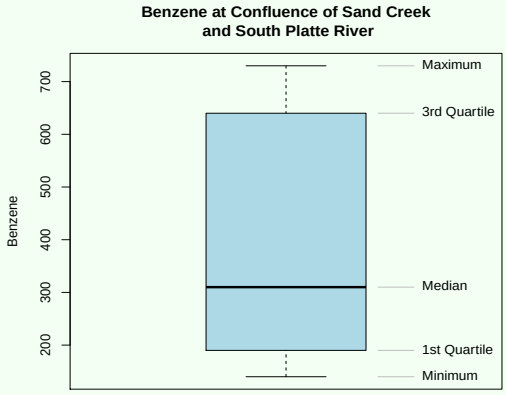

Figure 3.11: Boxplot of benzene concentrations at the confluence of the South Platte River with Sand Creek in Denver, Colorado.

When there are outliers in the data, we show them as isolated points in the boxplot. For this purpose, an outlier is defined to be any observation that lies farther than one and a half interquartile ranges away from the nearest quartile, that is, outside the two "***fences***":

Lower and Upper Fences: For data $X_1, X_2, \ldots, X_n$, the lower and upper fences for deciding whether an observation is an outlier are

$$\text{Lower fence} \;=\; Q_1 - 1.5(IQR) \qquad \text{Upper fence} \;=\; Q_3 + 1.5(IQR)$$

When outliers are shown as isolated points in a boxplot, the whiskers extend only as far as the largest and smallest observations that *aren't* outliers. The next example illustrates.

Example 3.16: Boxplots Showing Outliers

Refer to the data on benzene concentrations in the South Platte River in Example 3.15. For the data collected downstream of the confluence with Sand Creek, the five number summary is

Min	Q_1	$\tilde{X}$	Q_3	Max
130	190	240	260	590

and so the interquartile range is $IQR = 260 - 190 = 70$. We'll deem as outliers any observations greater than

$$Q_3 + 1.5(IQR) \;=\; 260 + 1.5(70) \;=\; 365$$

or less than

$$Q_1 - 1.5(IQR) \;=\; 190 - 1.5(70) \;=\; 85.$$

The largest benzene concentration, 590 ppb, is therefore an outlier, but it's the only one. A boxplot showing this outlier as an isolated point is shown below. Notice that the upper whisker extends only as far as 300 ppb, the largest observation that's *not* an outlier.

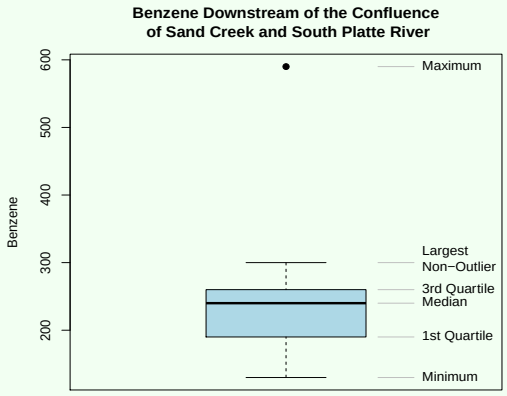

Figure 3.12: Boxplot of benzene concentrations downstream of the confluence of the South Platte River with Sand Creek in Denver, Colorado.

Comment: Although the "1.5 IQRs" criterion is a useful way of identifying unusual observations, it often singles out observations that *shouldn't* be considered unusual. In particular, observations in the long tail of a skewed distribution often appear as outliers in boxplots even though such isolated observations are to be expected in skewed distributions.

The true usefulness of boxplots is for comparing samples from two or more populations, side by side in the same graph, as in the next example.

Example 3.17: Side-By-Side Boxplots

For the benzene concentrations from Examples 3.15 and 3.16, we can make comparisons between the two locations using boxplots shown side by side in the same graph, as below.

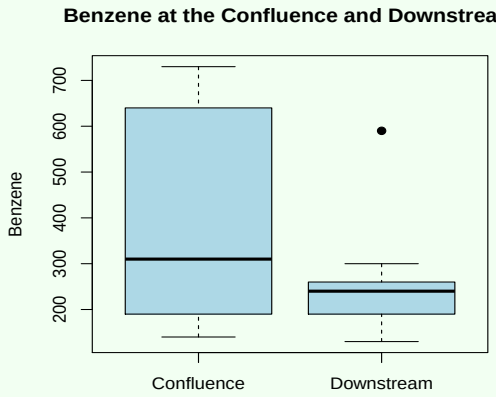

Figure 3.13: Side-by-side boxplots of benzene concentrations at two locations along the South Platte River in Denver, Colorado.

It's clear from the boxplots that the benzene concentrations tend to be lower at the location farther downstream from the chemical spill. It's also clear that there's more day-to-day variation in the concentrations at the location nearer to the spill.

Comment: Boxplots of data whose distribution is skewed will appear asymmetrical, with a shorter whisker at one end and a longer whisker, often with isolated points, extending in the direction of the long tail of the skewed distribution. As an example, the boxplots below were made from the same six data sets whose histograms are shown in Fig. 3.9. The values of the coefficients of skewness, g, and quartile skew coefficients, qs are also given.

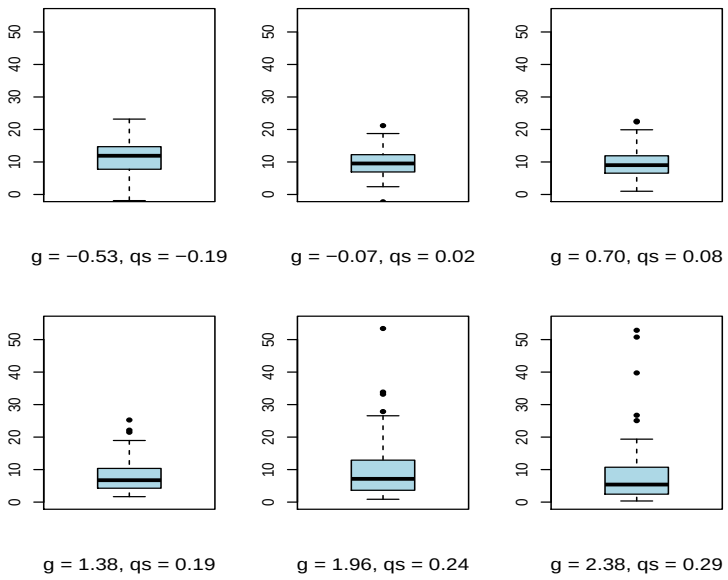

Figure 3.14: Boxplots of data whose distributions have varying degrees of skewness and the corresponding values of the coefficient of skewness g and quartile skew qs.

3.5 Problems

3.1 The following data are water temperature measurements (°C) made in 2001 at nine sites on Lake Tahoe and Donner Lake, near the border of California and Nevada [10].

Lake Temperatures

Site Name	Temperature
Tahoe City	17.5
TRG Buoy	17.5
Sugar Pine Point	17.8
Emerald Bay	18.3
Ski Run Marina	18.8
Camp Richardson	17.8
Tahoe Keys Marina East	19.3
Donner Lake Boat Ramp	20.3
Donner Lake Buoy	20.0

a) Calculate the mean of this set of data.

b) Calculate the median.

c) Calculate the geometric mean.

3.2 The table below shows data on the numbers of injuries and deaths by lightning strikes in the U.S. for each of the years 1959 - 2005, as compiled by the National Climatic Data Center from reports by the National Weather Service [8].

Lightning Casualties

Year	Injuries	Deaths	Year	Injuries	Deaths
1959	134	75	1983	203	49
1960	91	48	1984	156	33
1961	166	61	1985	149	34
1962	100	48	1986	120	32
1963	129	150	1987	190	35
1964	95	49	1988	146	30
1965	113	57	1989	175	23
1966	91	39	1990	152	39
1967	61	27	1991	207	36
1968	132	51	1992	161	25
1969	79	46	1993	147	20
1970	140	50	1994	288	32
1971	118	62	1995	233	43
1972	90	51	1996	309	52
1973	123	50	1997	306	42
1974	84	58	1998	285	44
1975	115	38	1999	243	46
1976	91	34	2000	364	51
1977	138	59	2001	370	47
1978	114	44	2002	256	51
1979	79	24	2003	236	44
1980	186	39	2004	278	33
1981	184	40	2005	309	38
1982	63	33			

a) Make a histogram of the deaths due to lightning strikes. What unusual feature do you notice in the histogram?

b) Compute the mean and the median of the death counts. Which statistic has the larger value? Explain why.

c) Referring to the outlier in the data, the cited report states:

> On December 8, 1963 the crash of a jetliner killing 81 people near Elkin, Maryland, was attributed to lightning by the Civil Aeronautics Board investigators.

Subtracting 81 from the tabulated value (150) for 1963 gives a more reasonable death count for that year. Recompute the mean and median with 69 replacing the outlier 150 for the year 1963. How do the values of the mean and median compare to those computed in part *b*?

3.3 One possible way of reducing atmospheric carbon dioxide (CO_2) emissions from power plants is to store them in the earth's subsurface below a sealing caprock. Depleted oil reservoirs are potential geologic storage sites for CO_2.

One criterion for determining the suitability of a depleted oil reservoir as a CO_2 storage site is its distance from the power plant. The table below shows distances (km) of 15 depleted oil reservoirs in the Galveston, Texas area from a power plant in Texas City, Texas [12].

CO_2 Storage Sites

Field Name/Reservoir Name	Distance from CO_2 Source
Cedar Point/No name	32.9
Chocolate Bayou/Frio Upper	21.6
Chocolate Bayou/Alibel	21.6
Fig Ridge/Seabreeze	71.5
Franks/8900 Sand	8.9
Gillock/Big gas	12.6
Gillock/East Segment	12.6
South Gillock/No name	7.9
E. U. Frio Hastings/No name	26.1
W. Frio Hastings/No name	26.1
Moores Orchard/No name	91.0
Oyster Bayou/No name	60.4
Red Fish reef/Combined	25.2
North Thompson/No name	59.8
Webster/No name	31.4

a) Make a histogram of the data.

b) Is the shape of the distribution left skewed, right skewed, or approximately symmetric?

c) Based on your answer to part *b*, which measure of center's value would you expect to be larger, the sample mean's or the sample median's?

d) Calculate the sample mean and the sample median.

3.4 A study was carried out to investigate the use of a new method for measuring concentrations of organic compounds in water [13], [5]. Fifteen laboratories participated in the study by using the method to measure chlorobenzene (μg/L) in four reference water specimens having known chlorobenzene concentrations 0.88, 1.10, 4.41, and 5.29 μg/L. The table below shows the resulting data.

Chlorobenzene Measurements

Laboratory	Conc= 0.88	Conc= 1.10	Conc= 4.41	Conc= 5.29
1	1.08	1.24	4.45	5.71
2	2.35	0.96	4.53	5.24
3	1.30	1.30	4.90	6.80
4	1.20	1.40	3.90	4.80
5	2.20	0.93	4.90	4.00
6	1.21	1.10	4.50	5.37
7	1.20	1.20	4.40	4.90
8	1.10	1.00	4.30	5.80
9	0.80	1.00	5.30	5.50
10	1.30	1.70	4.70	6.60
11	1.10	1.20	4.10	5.30
12	1.00	1.30	4.90	5.40
13	1.20	1.10	4.80	5.60
14	0.55	0.79	3.33	3.65
15	1.00	1.30	4.70	5.80

a) Make a dot plot of the chlorobenzene measurements made on the 1.10 μg/L reference specimens.

b) Is the shape of the distribution left skewed, right skewed, or approximately symmetric?

c) Calculate the sample mean and compare it to the true concentration value 1.10.

3.5 Repeat Problem 3.4 using the chlorobenzene measurements made on the 0.88 μg/L reference specimens.

3.6 Repeat Problem 3.4 using the chlorobenzene measurements made on the 4.41 μg/L reference specimens.

3.7 Repeat Problem 3.4 using the chlorobenzene measurements made on the 5.29 μg/L reference specimens.

3.8 In a study of heavy metal contamination in soil due to industrialization near the Manali area in Chennai, Southern India, metals were measured in soil at 32 sites in the region [9]. The table below shows their concentrations (mg/kg).

Metals in Soil

Site	As	Co	Cr	Pb	Sr
M-1	1.02	32.9	207.4	3.66	172.3
M-2	0.69	30.8	150.2	1.80	63.9
M-3	1.36	32.1	156.1	1.50	122.5
M-4	0.89	15.9	150.7	10.30	256.3
M-5	0.58	17.1	158.3	5.35	214.3
M-6	0.74	23.1	191.0	6.17	112.5
M-7	0.25	9.2	230.0	8.84	213.6
M-8	0.99	11.1	150.0	9.79	86.5
M-9	0.36	13.4	240.0	20.70	69.3
M-10	1.12	11.5	197.0	24.80	87.8
M-11	1.06	58.5	149.8	10.19	152.3
M-12	0.63	10.5	218.0	22.76	123.6
M-13	0.88	7.96	151.0	10.97	55.9
M-14	0.96	3.4	215.0	8.24	63.5
M-15	1.15	6.7	306.0	12.70	105.6
M-16	2.03	12.3	172.0	94.00	114.9
M-17	0.36	15.4	157.0	11.30	158.3
M-18	0.55	12.9	385.0	48.80	223.3
M-19	1.85	14.5	395.0	80.10	89.6
M-20	1.11	12.4	255.0	42.50	99.6
M-21	1.30	15.3	201.0	101.40	236.3
M-22	0.53	12.6	183.0	95.00	149.5
M-23	0.96	15.2	204.0	101.00	163.3
M-24	0.85	9.8	158.0	77.20	102.3
M-25	0.45	14.8	247.0	140.20	152.3
M-26	1.36	20.4	246.0	78.30	88.9
M-27	2.30	11.3	309.0	83.30	59.6
M-28	0.21	16.7	418.0	50.00	113.6
M-29	0.54	19.2	328.0	67.50	145.3
M-30	0.78	11.1	251.0	65.80	185.9
M-31	0.98	22.3	154.0	19.40	148.3
M-32	0.22	20.8	161.0	24.30	230.3

a) Make a histogram of the arsenic (As) concentrations, and describe the shape of the distribution.

b) Which of the two measures of center, the sample mean or median, would you expect to be larger for this set of data?

c) Compute the sample mean and median.

d) Which measure of center, the sample mean or median, is more representative of a typical value for this data set?

3.9 Repeat Problem 3.8 using the cobalt (Co) concentrations.

3.10 Repeat Problem 3.8 using the chromium (Cr) concentrations.

3.11 Repeat Problem 3.8 using the lead (Pb) concentrations.

3.12 Repeat Problem 3.8 using the strontium (Sr) concentrations.

3.13 The data below are cadmium concentrations (mg/kg) from the Sacramento Army Depot Superfund Site (in EPA Region 9), as reported in [14].

Cadmium Concentrations

26.2	27.6	445.0	30.8	486.3	513.8	112.8	159.3	1300.0	6.7
33.7	35.0	11.0	22.1	830.9	125.1	40.8	345.5	384.8	183.0
2300.0	1500.0	260.3	32.1	166.2	31.7	12.4	614.5	639.5	116.2
119.4	111.6	10.3	1.7	3.3	10.5	11.7	10.3	122.3	283.0
265.1	125.5	131.1	47.9	119.3					

a) Make a histogram of the data and describe its shape.

b) Calculate the sample mean and median of the data. Which is larger?

c) Take the log of each cadmium concentration and make a histogram of the log concentrations. How does the shape of this histogram differ from the one in part *a*?

d) Calculate the geometric mean of the (original) cadmium concentrations.

3.14 The data below are counts of greenbugs (aphids) on each of 20 oat plants recorded in April, 1981 and reported in [15].

1	18	2	0	40	3	8	5	12	11
8	7	6	3	15	27	8	4	3	2

a) Make a histogram of the data and describe its shape.

b) Compute the sample mean and median of the data. Which is larger?

c) Compute the coefficient of skewness.

3.15 Refer to the data on water temperatures in Lake Tahoe and Donner Lake given in Problem 3.1.

a) Compute the sample standard deviation.

b) Compute the interquartile range.

c) Compute the median absolute deviation.

d) Compute the coefficient of skewness.

e) Compute the quartile skew.

3.16 Increased usage of groundwater for irrigation in Jordan in recent decades has raised concerns about potential shortages where groundwater mechanics aren't well understood.

To explore the origin and movement of groundwater in the upper Yarmouk Basin, northern Jordan, hydrochemical measurements were made on water samples from 40 wells throughout the region and examined for spatial patterns [2]. The data below are the chlorine (Cl) measurements (mg/L).

Chlorine in Groundwater

Well	Cl	Well	Cl	Well	Cl	Well	Cl	Well	Cl
AD1001	113.0	AD1024	138.0	AD1064	141.0	AD1306	127.9	AD3023	87.6
AD1003	111.4	AD1037	111.0	AD1160	111.0	AD1307	87.7	AD3024	103.6
AD1007	220.0	AD1046	152.0	AD1168	97.0	AD1319	116.0	AD3025	91.1
AD1015	116.0	AD1050	98.0	AD1170	127.0	AD1320	134.0	AD3040	106.7
AD1016	103.0	AD1054	137.0	AD1230	102.0	AD1323	96.0	AD3044	88.7
AD1020	108.0	AD1055	123.0	AD1258	112.4	AD3004	134.8	AD3047	89.7
AD1021	117.0	AD1056	124.0	AD1260	109.5	AD3021	90.0	AD3057	128.0
AD1023	545.0	AD1063	130.8	AD1281	94.0	AD3022	97.9	AD3058	106.0

a) Determine if there are any outliers in the data using the "1.5 IQRs" rule.

b) Make a boxplot of the data showing outliers, if any, as isolated points.

c) Choose and then compute an appropriate measure of center for this data set.

d) Choose and then compute an appropriate measure of variation for this data set.

e) Choose and then compute an appropriate measure of skewness for this data set.

3.17 The study cited in Problem 3.16 also reported the magnesium (Mg) measurements (mg/L) for the 40 wells in the Yarmouk Basin in Jordan. The Mg measurements are shown in the table below.

Magnesium in Groundwater

Well	Mg	Well	Mg	Well	Mg	Well	Mg	Well	Mg
AD1001	23.3	AD1024	21.8	AD1064	25.5	AD1306	23.2	AD3023	21.4
AD1003	23.2	AD1037	20.2	AD1160	26.4	AD1307	25.5	AD3024	21.6
AD1007	26.3	AD1046	25.8	AD1168	26.0	AD1319	25.5	AD3025	20.0
AD1015	21.5	AD1050	21.7	AD1170	20.1	AD1320	25.8	AD3040	25.4
AD1016	21.6	AD1054	23.3	AD1230	21.5	AD1323	25.8	AD3044	23.2
AD1020	21.4	AD1055	25.6	AD1258	23.2	AD3004	25.6	AD3047	25.5
AD1021	23.4	AD1056	23.3	AD1260	23.4	AD3021	26.1	AD3057	23.0
AD1023	38.5	AD1063	25.7	AD1281	23.0	AD3022	20.1	AD3058	19.2

a) Determine if there are any outliers in the data using the "1.5 IQRs" rule.

b) Make a boxplot of the data showing outliers, if any, as isolated points.

c) Choose and then compute an appropriate measure of center for this data set.

d) Choose and then compute an appropriate measure of variation for this data set.

e) Choose and then compute an appropriate measure of skewness for this data set.

3.18 In Example 2.9 of Chapter 2, two-stage sampling was used in a study of the zinc (Zn) and calcium (Ca) concentrations in soil on a research field in Slovenia.

In the first stage, the field was partitioned into subplots, and a simple random sample of $m = 5$ subplots was selected. In the second stage, from each of the selected subplots, a simple random sample of $\tilde{n} = 3$ soil specimens was selected and the Zn and Ca (both mg/kg) concentrations measured in each specimen. The table below shows the data.

Zinc and Calcium in Soil

Subplot	Specimen from the Subplot	Zn	Ca
	1	51	0.78
1	2	41	0.57
	3	52	0.44
	1	54	0.57
2	2	55	0.51
	3	47	0.53
	1	47	0.41
3	2	48	0.55
	3	48	0.54
	1	48	0.55
4	2	45	0.43
	3	48	0.57
	1	39	0.49
5	2	39	0.58
	3	43	0.47

In this problem we'll summarize the data one subplot at a time.

a) To assess spatial variability in Zn concentrations, compute and compare the five subplot sample means.

b) Based on the five sample means computed in part *a*, which subplot appears to have the highest Zn concentrations?

c) It's often the case that measurements made in close proximity spatially tend to be similar in value, whereas ones farther apart tend to be dissimilar. This suggests that within each of the five subplots, the Zn measurements should be fairly homogeneous, and therefore have small standard deviation. Calculate the sample variance and sample standard deviation for each of the five subplots.

d) Based on the five sample standard deviations calculated in part *c*, for which subplot do the Zn concentrations vary the most?

e) Make side by side boxplots of the Zn concentrations for the five subplots (one boxplot for each subplot).

3.19 Repeat Problem 3.18 using the calcium (Ca) concentrations.

3.20 Problem 2.6 in Chapter 2 described a study of contaminants in eggs of herons and egrets in the Mai Po Marshes Nature Reserve in Hong Kong. One question of interest was whether contaminant concentrations in eggs would differ for the two species, for example due to different diets and metabolic characteristics.

An egg was taken from each of nine randomly selected Little Egret nests in the Mai Po egretry and nine randomly selected Black-Crowned Night Heron nests in the A Chau egretry. Several contaminants were measured in each of the sampled eggs. The data on chlordanes (CHLs, ng/g) and total organochlorines (OCs, ng/g) are shown below.

Contaminants in Eggs

	Egg No.	CHLs	OCs
Little Egret Eggs from Mai Po Village	1	390	3710
	2	240	2060
	3	320	2100
	4	470	3760
	5	310	3530
	6	450	2870
	7	100	940
	8	81	1140
	9	180	1850
Black-Crowned Night Heron Eggs from A Chau	10	44	180
	11	19	1270
	12	66	1010
	13	75	1150
	14	30	550
	15	13	470
	16	6	870
	17	24	480
	18	6	360

In this problem we'll examine the CHL concentrations.

a) Make side by side boxplots of the CHL concentrations for the two bird species.

b) Based on the boxplots in part *a*, which type of eggs, egrets' or herons', tend to have higher CHL concentrations?

c) Based on the boxplots in part *a*, which type of bird eggs exhibit more variable concentrations of CHLs?

3.21 Refer to the study described in Problem 3.20 and the data given there. In this problem we'll examine the OC concentrations.

a) Make side by side boxplots of the OC concentrations for the two bird species.

b) Based on the boxplots in part *a*, which type of eggs, egrets' or herons', tend to have higher OC concentrations?

c) Based on the boxplots in part *a*, which type of bird eggs exhibit more variable concentrations of OCs?

Bibliography

[1] Mercury study report to congress, volume IV: An assessment of exposure to mercury in the United States. Technical Report EPA-452/R-97-006, United States Environmental Protection Agency, 1997.

[2] Nizar Abu-Jaber and Alaa Kharabsheh. Ground water origin and movement in the Upper Yarmouk Basin, Northern Jordan. *Environmental Geology*, 54(7):1355–1365, 2008.

[3] Bruce Finley. Toxics from Suncor Refinery spill still seeping into water; Colorado vows to "accelerate" response. *The Denver Post*, Jan. 21 2012.

[4] R. A. Fisher and C. I. Bliss. Fitting the negative binomial distribution to biological data. *Biometrics*, 9(2):176 – 200, June 1953.

[5] American Society for Testing Standards Designation D 5790 (2006). *Standard Test Method for Measurement of Organic Compounds in Water by Capillary Column Gas Chromatography/Mass Spectrometry*, volume 11. ASTM, West Conshohocken, PA, 2006.

[6] P. Garman. Original data on European red mite on apple leaves. Technical report, Connecticut, 1951.

[7] A. M. Hasson. Radiation components over bare and planted soils in a greenhouse. *Solar Energy*, 44:1–6, 1990.

[8] Stuart Hinson, Rhonda Herndon, and William Angel (Ed.). 2005 annual summaries. Technical report, National Oceanic and Atmospheric Administration, National Environmental Satellite Data Information Service, and National Climatic Data Center. Published by the National Climatic Data Center, Asheville, NC.

[9] A. K. Krishna and P. K. Govil. Assessment of heavy metal contamination in soils around Manali Industrial Area, Chennai, Southern India. *Environmental Geology*, 54(7):1465–1472, 2008.

[10] Micheal S. Lico. Gasoline-related organics in Lake Tahoe before and after prohibition of carbureted two-stroke engines. *Lake and Reservoir Management*, 20(2):164–1747, 2004.

[11] A. Matthews. Mercury content of commercially important fish of the Seychelles, and hair mercury levels of a selected part of the population. *Environmental Research*, 30(2):305–312, 1983.

[12] Vanessa Nunez-Lopez et al. Quick-look assessments to identify optimal CO_2 EOR storage sites. *Environmental Geology*, 54(8):1695–1706, 2008.

[13] C.H. Proctor. A simple definition of detection limit. *Journal of Agricultural, Biological, and Environmental Statistics*, 13(1):99–120, March 2008.

[14] Anita Singh, Ashok K. Singh, and George Flatman. Estimation of background levels of contaminants. *Mathematical Geology*, 26(3):361 – 388, 1994.

[15] Linda J. Willson, J. Leroy Folks, and J. H. Young. Multistage estimation compared with fixed-sample-size estimation of the negative binomial parameter k. *Biometrics*, 40(1):109 – 117, Mar 1984.