7 Descriptive Statistics: Univariate Data

> *Tester*: This can't be right. This man has 104% body fat! [turns to Homer] Hey, no eating in the tank!
>
> From: *The Simpsons*

7.1 Introduction

The word "statistics" probably suggests to you a collection of data. Indeed, the subject of statistics does deal with data--- how to gather it, organize it, and most importantly analyze it. In this and the following two chapters we will concentrate on the second of these goals, organizing or presenting data. After an excursion into probability theory we will develop some of the rudiments of statistical analysis. Although we will say very little regarding proper methods for gathering data, perhaps the following story will point out how even a very large data set may produce spurious conclusions.

In 1936 a well-known magazine, *The Literary Digest*, sent presidential election surveys to some 10 million potential voters. On the basis of more than 2.3 million returns it predicted that the Republican candidate, Alfred Landon, would win decisively. In fact, President Franklin Roosevelt was reelected in a landslide, receiving almost 61% of the total popular vote of 45 million.

How did the *Digest's* survey go wrong? It was not its first experience in presidential polling; it had accurately predicated the winner in the previous four presidential elections. However, its voter pool was drawn from lists of telephone and automobile owners, who in 1936 tended to represent the most economically well off segments of the population. Since the election was virtually a referendum on the economic policies of Roosevelt's New Deal, the *Digest's* selection criteria had unwittingly introduced a bias that was fatal to the accuracy of its predictions. This example might seem extreme, but it typifies the difficulties involved in selecting a sample from a much larger population, for the purpose of predicting some characteristic of this population.

While it is easy with hindsight to criticize the *Digest* poll for its selection criteria, it exhibits a more basic fault that reveals a lack of understanding of modern statistical methods. A few percentage points typically decide elections. To make a reliable prediction in such cases requires, as we shall learn, polling only several thousand *randomly* chosen voters, whether the voting population is one million or 40 million. The prodigious amount of work needed to process the millions of extra opinions adds very little to the practical value of the survey, and may not even insure against the effects of badly biased sampling.

A poll is fairly easy to understand. Everyone is tallied as either "for" or "against", or perhaps "undecided". Most data that people have to think about are more difficult to understand than such a simple tabulation. An important part of statistical analysis, therefore, concerns presenting data in a way that makes it more comprehensible. Many of these techniques involve visual representations or numerical summaries, to which we now turn our attention.

7.2 The Histogram

A data set that contains one number for each observation is called *univariate*. For example, if an exam were given to 30 students, then the grades received by each of them would constitute a univariate data set. A histogram is a graphical method for representing such data. The English political economist William Playfair created it for this purpose in the 18th century. To illustrate the construction it is useful to consider an example.

Example 7.1: Prepare a histogram for the times between successive eruptions of the geyser, Old Faithful.

Solution:

Old Faithful is a geyser in Yellowstone National Park. It erupts quite frequently, though perhaps not as regularly as its name might indicate. The table below contains a sample of 108 data values from a larger file (*oldFaith.xls*) giving the time in minutes from one eruption to the next. The entries are not records of consecutive eruptions.

Time in Minutes Between Eruptions of Old Faithful, Yellowstone National Park								
Source: D. Howell at http://www.uvm.edu/~dhowell								
77	88	75	96	51	92	54	93	78
85	71	52	50	89	81	45	88	77
57	58	***65***	60	80	78	79	85	79
87	84	58	73	74	83	60	71	93
57	62	51	72	78	85	87	73	82
77	54	61	62	***69***	91	93	92	***65***
71	89	***68***	88	87	50	49	56	***69***
75	71	74	81	59	56	62	77	86
81	60	82	57	87	93	84	80	93
80	91	92	60	87	84	78	60	75
57	72	80	54	47	53	87	54	83
50	49	***68***	78	57	87	72	51	75

Table 7.1

It's difficult to grasp this data just by looking at the numbers. A histogram gives a pictorial representation. To construct it, we first find the largest and smallest values. In this case the smallest is 45 minutes and the largest is 96 minutes. We then divide the slightly larger interval from 40 to 100 into a number of equal sub-intervals (called *bins*) and find how many data values lie in each of these subintervals. Let's choose bins of length 6 so we obtain the ten intervals (40, 46], (46, 52], (52, 58], (58, 64], (64, 70], (70, 76], (76, 82], (82, 88], (88, 94], and (94,100]. The open parenthesis, "(", indicates that a data value of 52, say, is to be counted in the second bin, but not in the third. For example, there are six entries that fall in the interval (64, 70]. They appear in bold italics in Table 7.1. The resulting frequencies are given in the following table. The third column shows the fraction of the 108 data values that fall in each bin. This is called the *relative frequency* and is often expressed as a percent, instead of a decimal. For example, in 18.5% of eruptions, the next eruption occurred between 76 and 82 minutes later.

Bins	*Freq.*	*Rel. Freq.*
(40, 46]	1	.009
(46, 52]	10	.093
(52, 58]	14	.130
(58, 64]	10	.093
(64, 70]	6	.056
(70, 76]	15	.139
(76, 82]	20	.185
(82, 88]	19	.176
(88, 94]	12	.111
(94, 100]	1	.009

Table 7.2

Finally, the last table is converted to a (relative frequency) histogram, which is just a bar chart whose y axis plots the relative frequency over each of the bin intervals. In the version below we label the midpoint of each of the bin intervals.

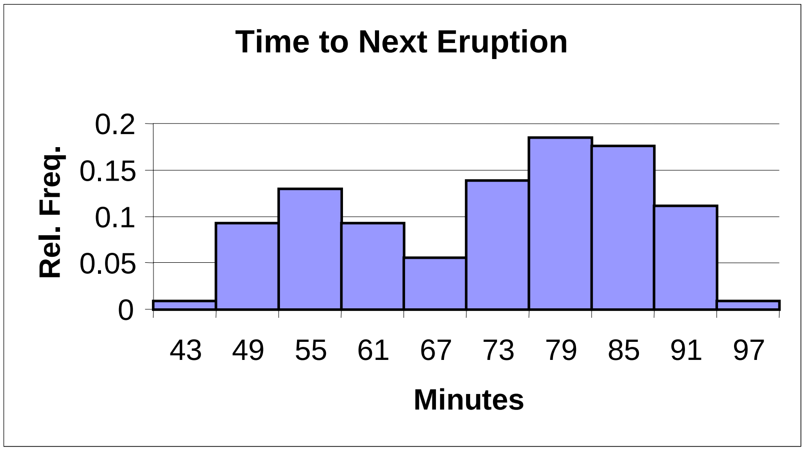


Figure 7.1

The graph illustrates a striking feature of the data. There exist two peaks in the time between eruptions, one at approximately 55 minutes and the second at approximately 80 minutes. Such a data set is called *bimodal,* a mode being a value where the histogram has a local maximum, to use some terminology from calculus.■

The histogram is visually appealing and, as we will see later, a useful tool for understanding some concepts associated with probability theory. ***However, the shape of a histogram may be quite sensitive to the choices made in the construction***. For example, using subintervals of length 10 instead of 6 gives the following histogram for the data set considered above.

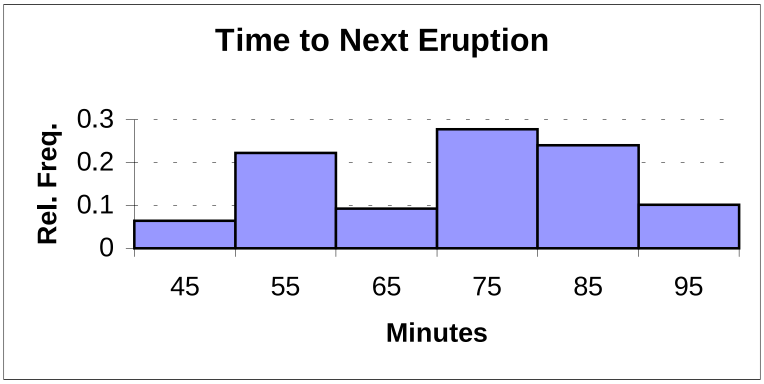


Figure 7.2

The bimodal distribution is somewhat less striking here. Too few intervals prevent patterns from appearing. Too many intervals will fragment the data and will also not exhibit any clustering patterns. More subtle deficiencies arise when we try to look at the histogram and estimate the values of the median, average, and other simple numerical measures associated with the data (we'll define these below).

Finally, it is very difficult to compare the distributions of two data sets by comparing their histograms. It is visually difficult to line up the pictures, and different interval sizes make comparisons hard to see. In section 7.5 we will construct a visual display that is more compact than a histogram, provides significant numerical information regarding the data, and also allows one to readily make visual comparisons between data sets.

Let's summarize the procedure for constructing a histogram:

- Find the largest and smallest data values in your set.
- Divide the interval from the smallest to the largest value (or a slightly larger one) into a number of equal subintervals (bins). The number of bins should be between 5 and 20, and you should try to arrange so that the endpoints that define the bins are “round” numbers.
- Count the number of data values that lie in each bin. It is probably easiest, and least likely to produce errors, if you first sort the data from smallest to largest. Otherwise, you can simply go through the data, keeping tally marks (𝍸) to tabulate the data that falls in each bin.

- Construct a bar graph whose height over each bin interval is the frequency or relative frequency you obtained in the previous step.
- Label your axes with both numerical values and text identifying the variable associated with each axis.

7.3 Numerical Measures: Central Tendency

Numerical measures provide the most common method for succinctly summarizing a data set. We will see later how they may also be used to give a compact visual representation of the data. We first try to answer the question, “What is a typical value in a set of data?” Two commonly used measures are the *median* and the *mean.*

The median m of a data set is a number with the property that there are as many numbers in the set which are less than m, as there are greater than m. For example, if the data set consists of the numbers 3, 8, 4, 2, and 9 then we first order them as 2, 3, 4, 8, and 9. The median is 4, since there are two numbers less than 4, and two numbers greater than 4. If the data were 3, 8, 4, 2 we would again order the numbers as 2, 3, 4, and 8. This time there are many numbers that meet the definition of the median. Any number which is greater than 3 and less than 4 will do. To get an unambiguous answer in this case we conventionally define the median to be the average of the two values closest to the middle, obtaining $m = 3.5$. As this example shows, the median need not be a value from the original data. We summarize this in a formal definition, after introducing some notation.

We will denote the elements of some unspecified data set by $x_1, x_2, \ldots, x_n$ where n is the number of elements in the set. For example, we might have, as above, $x_1 = 3$, $x_2 = 8$, $x_3 = 4$, $x_4 = 2$. Notice that this notation does not necessarily mean that the numbers $x_1, x_2, \ldots, x_n$ are in size order. When we want to refer to the elements of the data set in size order we will use the notation $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. In the previous example, $x_{(1)} = 2$, $x_{(2)} = 3$, $x_{(3)} = 4$, $x_{(4)} = 8$. With this notation, we can state

Definition 7.1: If $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ is an <u>ordered</u> data set of n elements, the median m is defined as $x_{(\frac{n+1}{2})}$ (i.e. the middle value) if n is odd, and as $\frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})$ (i.e. the average of the two middle values) if n is even.■

It is important to emphasize that in applying this definition the data must first be placed in increasing order. You then locate the middle value, if the number of data elements is odd, or take the average of the two values closest to the middle, if the number of data points is even. The median is an example of a measure of central tendency. By itself, however, it gives no indication of the spread of the data. We will consider measuring the spread of data in the next section.

Example 7.2: Find the median age of the signers of the Declaration of Independence.

Solution:

The table below lists the ages at the signing of the Declaration of Independence of 55 of the 56 signers (the age of the 56th is not known).

Ages of Signers of Declaration of Independence									
26	26	29	30	31	32	33	33	34	34
35	35	35	37	37	38	38	39	39	39
40	41	41	42	42	42	43	44	45	45
45	45	46	46	46	46	47	48	49	50
50	50	50	51	52	53	53	55	60	60
61	63	63	69	70					

Table 7.3

The ages are arranged in sorted order. Since there are an odd number of values, Definition 7.1 implies that the median is the $\frac{55+1}{2} = 28^{th}$ value or 44. Roughly speaking, half of the signers were younger than 44 years old and half were older.■

The median is very useful in describing arbitrary data. However, many data sets exhibit symmetry, and more specifically a “bell-shaped” histogram, which can be described more usefully by a different measure of central tendency, the *mean*.

Definition 7.2: The mean of a data set $x_1, x_2, \ldots, x_n$ (not necessarily in size order) is the number $\overline{x}$ given by

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.■$$

The mean is sometimes also called the *average* of the numbers $x_1, x_2, \ldots, x_n$.

Example 7.3: Prepare a histogram and find the mean and median for the data on university spending in Table 7.4 below.

Solution:

US News & World Report annually publishes a rating of U.S. colleges and universities. One factor the magazine uses in its rating is the amount that the institution spends per student. The following table reports these numbers for the 50 top rated schools from the 1991 report. The numbers are sorted in decreasing order and not in order of the school's overall ranking. (In recent years the magazine has stopped publishing these numbers.)

Spending per Student at 50 top rated US Universities (thousands of $)									
Source: *US News & World Report*, 1991									
106.7	74.8	72.6	63.6	57.9	57.7	51.3	50.8	50.7	50
44.5	44	42.4	40.8	40.3	40.2	37.6	36.7	36.6	35.4
31.8	29.6	29.5	28.4	28.2	28	28	27.7	26.3	26.3
25.9	25.5	25	24.8	24	24	24	23.9	22.5	22.2
22.1	21.4	21	19.4	18.9	17.3	16.9	16.5	15.6	12.8

Table 7.4

Below is a histogram for this data.

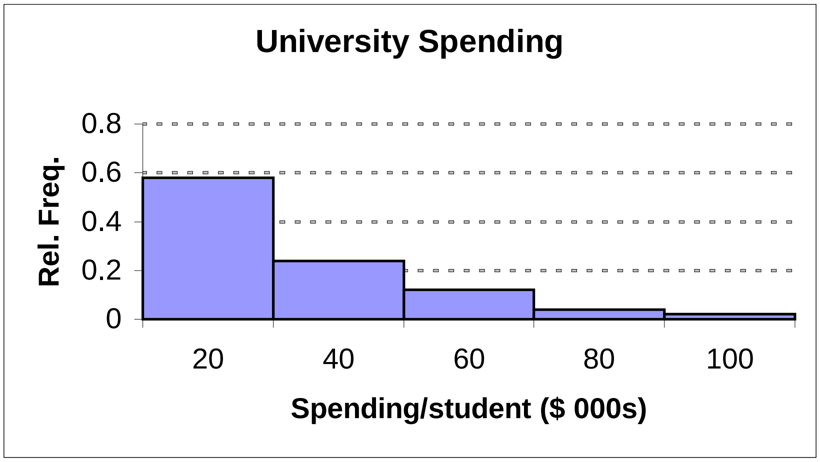


Figure 7.3

Histograms such as this, where the data tails off further in one direction than another are called *skewed*. The university data illustrated in Figure 7.3 is skewed to the right since exceptionally large values occur. Right skewed data arises frequently in economic statistics, where very large data values may occur beyond what is "usually" expected. The mean and median reflect this skewness. In fact, here we have the median $m = 28.1$ (check!) and the mean $\bar{x} = 34.8$.

The divergence of the mean and the median is quite typical for data that is skewed. The mean is affected by all values. In particular, a few extremely large values can greatly influence the size of the mean, whereas they will have little effect on the median. For the data in Table 7.4, if the largest value were changed to 200, there would be no change in the median, but the mean would become 36.7. Using the mean to summarize skewed data can be very misleading, as it may produce a value very far from the middle of the distribution.■

7.4 Numerical Measures: Spread

Knowing the median or the mean tells us something about the center of the distribution, but gives no indication as to the spread of the data. We introduce several ways to measure spread - the *range*, the *quartiles* and the *standard deviation*.

The range is the simplest numerical measure of spread. Its definition is simple.

Definition 7.3: The *range* of a data set is the difference between the biggest and smallest values.■

Note that the range is a single number, not a pair. Thus for the data in Table 7.3 the range is $70 - 26 = 44$ years.

The quartiles are a measure of spread associated with the median. As the name indicates there are four quartiles. Roughly speaking, the first quartile q_1 is a value for which 25% of the data is less than q_1 and 75% is greater than q_1. The second quartile is just the median. The third quartile q_3 has the property that 75% of the data is less than q_3 and 25% is greater. The fourth quartile is simply the largest data value.

While these describe the idea of a quartile, the actual definition, as for the median, has some technical complications. More annoyingly, there is no universally agreed upon method for defining these quantities. Different statistical software packages will often produce somewhat different values for the quartiles. One definition that is simple to apply, although certainly not in universal use, is the following.

Definition 7.4: The *first quartile* q_1 is the median of all values smaller than the median of the data set. The *third quartile* q_3 is the median of all data values that exceed the median of the data set.■

Example 7.4: Find the quartiles for the ages of the signers of the Declaration of Independence. (See Table 7.3.)

Solution:

Referring to the data in Table 7.3 we found the median to be 44 so the first quartile is the median of all values smaller than 44. There are 27 such values. The median of these values is the 14th value or $q_1 = 37$. Similarly the third quartile is the median of the 27 values exceeding 44. This gives $q_3 = 50$.■

Roughly 25% of the data are less than q_1 and 75% of the data are less than q_3. Therefore, half the data lies between these two values. The difference between these quartiles measures the spread of the middle 50% of the data. This leads to

Definition 7.5: The *interquartile range* (IQR) is the difference between q_3 and q_1.■

When the mean is used as an indicator of centrality, we get some sense of the spread of the data from the *standard deviation.*

Definition 7.6: The standard deviation s of a data set $x_1, x_2, \ldots, x_n$ is the number given by

$$s = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}}\ .■$$

The definition is rather complicated, and in fact is a bit messy to compute. It is worth going through a simple example to understand exactly what the formula says.

Example 7.5: Find the standard deviation of the data values 2, 4, 4, 3, 7, and 8.

Solution:

The mean $\overline{x}$ of the numbers is 4.67, to two decimal places. To compute the standard deviation, for each number in the data set we find its difference from the mean and then square the result. The numerator inside the square root in Definition 7.6 is given by

$$(2-4.67)^2 + (4-4.67)^2 + (4-4.67)^2 + (3-4.67)^2 + (7-4.67)^2 + (8-4.67)^2 = 27.333 \qquad (7.1)$$

Finally, we divide this by $n-1=5$ and take the square root, obtaining $s = \sqrt{27.333/5} = 2.34$.■

One might well ask why we employ such a complicated definition. Our objective is to measure, in some sense, how much the data values on average differ from the mean. You might ask why we don't simply use the quantity

$$\frac{|x_1 - \overline{x}| + |x_2 - \overline{x}| + \cdots + |x_n - \overline{x}|}{n},$$

which is the average of the absolute deviation of each data value from the mean value. This is used at times, but in fact it is not as useful as the standard deviation, because it is difficult to manipulate algebraically. Rather, the squared values $(x_i - \overline{x})^2$ are used to measure deviation from the mean; we then average these (for technical reasons dividing by $n-1$ instead of n). Finally, we take the square root so that the "units" in which we measure s will be the same as those associated with the data set values.

On route to computing the standard deviation we have also computed a related measure known as the *variance*. We introduce this here, although its principal role is in probability theory.

Definition 7.7: The variance of a data set is the square of the standard deviation, and is denoted by s^2, where s is the standard deviation. In other words,

$$\text{variance} = s^2 = \frac{\left(x_1 - \overline{x}\right)^2 + \left(x_2 - \overline{x}\right)^2 + \cdots + \left(x_n - \overline{x}\right)^2}{n-1}. \blacksquare$$

For the numbers 2, 4, 4, 3, 7, and 8 considered above, the calculation in (7.1) gives the numerator of the variance and so $s^2 = 27.333/5 \approx 5.47$. Note that if the data have units, the variance is measured in square units.

How well do the mean and standard deviation describe a data set? We interpret this question as asking for some estimate for the fraction of the data that lies within one, two, or three standard deviations of the mean. For data having an approximately bell-shaped histogram the following rule holds. We will discuss its mathematical origins later in the course.

Rule 7.1 (The Bell Curve Rule): If a data set has an approximately bell-shaped histogram with mean $\overline{x}$ and standard deviation s then

a) The interval from $\overline{x} - s$ to $\overline{x} + s$ contains approximately 68% of all the data values.

b) The interval from $\overline{x} - 2s$ to $\overline{x} + 2s$ contains approximately 95% of all the data values.

c) The interval from $\overline{x} - 3s$ to $\overline{x} + 3s$ contains approximately 99.7% of all the data values.■

A data value x in the interval from $\overline{x} - s$ to $\overline{x} + s$ satisfies $\left|x - \overline{x}\right| \le s$ and is said to lie within one standard deviation of the mean. Thus item a) above asserts that 68% of the data values lie within one standard deviation of the mean, for data that has an approximately bell-shaped histogram. Similarly, 95% of such data lie within two standard deviations of the mean. Let's see how closely this rule is attained in one of our examples.

Example 7.6: Check how closely the Bell Curve rule is followed by the data in Table 7.3.

Solution:

Consider the signers' ages from Table 7.3**.** We want to see how well the Bell Curve Rule (Rule 7.1) applies. Let's first construct a histogram to see whether the data has the appropriate shape.

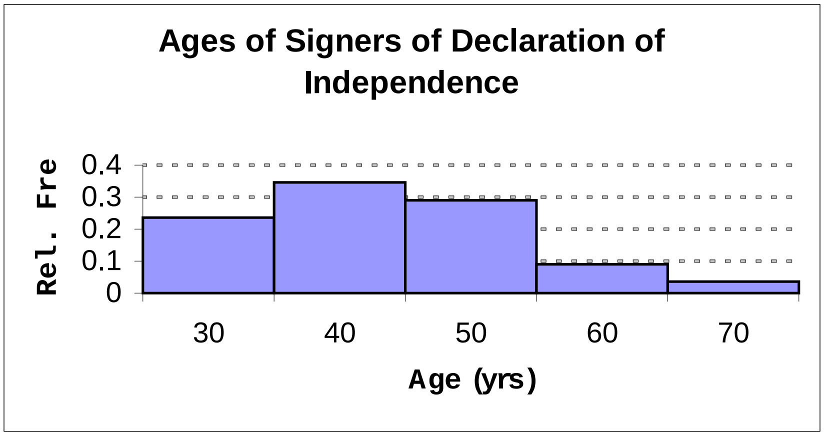


Figure 7.4

The histogram is slightly skewed to the right, but probably not enough to invalidate the Bell Curve Rule (Rule 7.1); let's check it. The value of $\overline{x} = 44.2$ and $s = 10.4$ so the interval one standard deviation from the mean extends from $\overline{x} - s = 33.8$ to $\overline{x} + s = 54.6$. There are 39 values in this interval (check this from Table 7.3), which is 39/55 = 71% of all data values, reasonably close to the predicted 68%. The interval two standard deviations on either side of the mean extends from 23.4 to 65.0 and includes 53 of the 55 values, or 53/55 = 96% of all values. Again, this agrees quite well with the rule.■

If the Bell Curve Rule (Rule 7.1) applies, we have a convenient way of classifying extreme data values. A data value more than two standard deviations from the mean would be regarded as unusual, since at most 5% of the values have that property. Similarly, a value more than three standard deviations from the mean might be considered exceptional.

The Bell Curve Rule (Rule 7.1), when it applies, provides a convenient way of making a crude estimate of the standard deviation from the range. Namely, since almost all of the data lies in the in the interval $\overline{x} - 2s$ to $\overline{x} + 2s$, the range of the data should be approximately the length of this interval. Thus range $\approx 4s$, which gives

Rule 7.2 (Range Rule): If the data have an approximately bell-shaped histogram then the standard deviation s is approximately equal to $\frac{\text{range}}{4}$.■

For example, for the signers' data the range is 44 years. Dividing by 4 gives as estimate for the standard deviation $s \approx 11$. As noted above (Example 7.6) the exact value is 10.4.

7.5 Box Plots

Using the median and quartiles we can construct a useful visualization tool called a *box plot.* A box plot for the university data of Table 7.4 is shown below, with explanations of the various components.

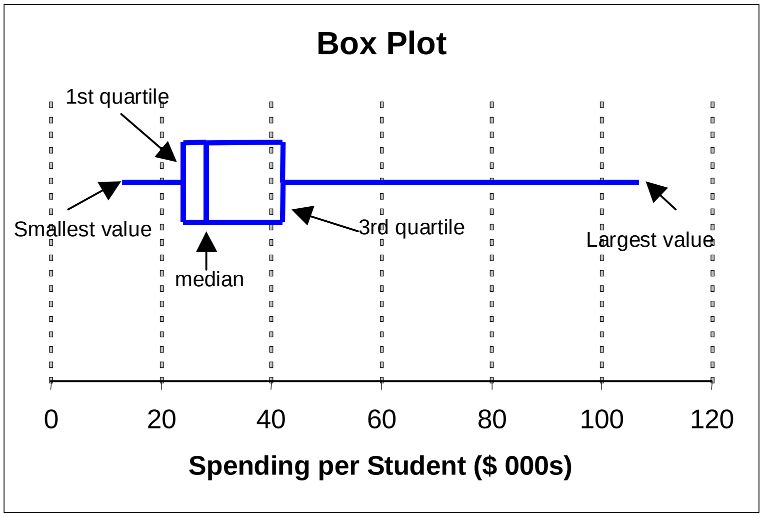


Figure 7.5

In a box plot, the central portion of the data from q_1 to q_3 is marked using a rectangular box extending parallel to a scale axis, usually drawn horizontally. The vertical width of the box has no quantitative significance. A vertical line is drawn through the box at the position corresponding to the median value for the data. Drawn outward from either side of the box are two lines (sometimes called "whiskers") extending to the smallest and largest values.

The box plot is particularly good at exhibiting the symmetry or lack thereof in a data set. The fact that in the figure above the median line is much closer to the first quartile than it is to the third shows that the data is skewed to the right. In fact, by definition, the interval from the first quartile to the median and the interval between the median and the third quartile each contain approximately 25% of the data. Yet the latter set of values is spread out over more than twice as large a range. The reader should compare the box plot with the histogram of the same data in Figure 7.3. Both give a similar picture of the distribution, with the box plot having the advantage of providing, through the horizontal scale, the numerical values of the median, quartiles, and maximum and minimum values.

A main advantage of a box plot is that it allows one to compare very quickly two or more data sets by simply stacking the box plots on top of each other using the same reference scale. The reader is invited to try this in some of the exercises.

7.6 Estimation

We noted above that the standard deviation, which is certainly messy to compute by hand, can be quickly estimated for certain data sets using the Range Rule (Rule 7.2). We now consider methods to estimate the median, mean and standard deviation if the original data are no longer available, but a data summary table is. To estimate the median we expand the frequency table by adjoining another column that gives the cumulative relative frequency. This column simply adds the percentages or relative frequency of all data up to and including the current bin.

Example 7.7: Compute the cumulative frequencies associated with the frequency distribution for the Old Faithful eruptions (Table 7.2). Use these to estimate the median and quartiles.

Solution:

We add a third column, the cumulative (relative) frequency to the data on Old Faithful, Table 7.2. The entry in the last row of this column should be one, but round-off may result in a slight discrepancy, as in Table 7.5.

Bins	*Freq.*	*Rel. Freq.*	*Cumul. Freq.*
(40, 46]	1	.009	.009
(46, 52]	10	.093	.102
(52, 58]	14	.130	***.232***
(58, 64]	10	.093	.325
(64, 70]	6	.056	.381
(70,76]	15	.139	.520
(76, 82]	20	.185	.705
(82, 88]	19	.176	.881
(88, 94]	12	.111	.992
(94,100]	1	.009	1.001

Table 7.5

The value 0.232, highlighted in row four, indicates that approximately 23% of the data values lie below 58, the upper bin boundary for that row. Therefore, we see that q_1 (the 25th percentile) must fall in the interval (58, 64], since at the beginning of this interval we have accounted for only 23% of the values, while 32.5% are accounted for at the end. Since 25% is only slightly larger than 23%, we might reasonably estimate the value of q_1 to be 59 or 60 minutes. (There is a very precise procedure for doing the latter estimation called linear interpolation, but we will not go into the details.) MS *Excel* gives $q_1 = 60$ for the data in Table 7.1. Similarly the median is in the interval (70, 76] and we would estimate its value as approximately 74 or 75. The precise value is 75. Using these ideas the reader should make an estimate for q_3.■

We can also estimate the mean and standard deviation from such a table. To see the idea, suppose we had the data 3, 4, 6, 2, 3, 6, 3, 2 and we wanted to find the mean. Well of course we could just compute

$$(3+4+6+2+3+6+3+2)/8 .$$

However, in computing the numerator we observe that some values are repeated and so we can simplify the calculation by grouping those together. This gives

$$\left((3\times 3)+(2\times 2)+(2\times 6)+(1\times 4)\right)/8 = 29/8 = 3.625\,.$$

Let's apply this idea to the Old Faithful data summarized in Table 7.5.

Example 7.8: Using Table 7.5 estimate the mean and standard deviation for time between eruptions of Old Faithful.

Solution:

To estimate the mean time between eruptions using Table 7.5, first find the total number of data values, n. This is found by adding the entries in the frequency column (column 2). We find $n = 108$. The numerator of $\overline{x}$ is obtained by adding all the data values. One of these values falls into the first data bin (40, 46]. We estimate the value of this data point using the midpoint 43 of the bin interval. There are 10 data values in the second bin, probably spread out through the interval (46, 52]. We will use as a common estimate for each of the ten values the midpoint 49 of this bin. Hence, the total contribution of these values to the sum should be approximately 10× 49. We can make a similar argument for the remaining intervals. This gives the estimate

$$\overline{x} \approx \frac{(1)(43)+(10)(49)+\ \ldots\ +(19)(85)+(12)(91)+(1)(97)}{108} = \frac{7794}{108} \approx 72.2\,.$$

The exact value using the original data is $\overline{x} = 72.6$, so the estimate is quite good.

A similar methodology can be used to estimate the standard deviation. We imagine the data values are grouped into the bins and then in the formula for s replace each data value by the value at the middle of its respective bin. For this data the calculation would give (using our estimate of $\overline{x} \approx 72.2$)

$$s \approx \sqrt{\frac{(1)(43-72.2)^2+(10)(49-72.2)^2+\ \ldots\ +(1)(97-72.2)^2}{107}} = \sqrt{\frac{20743.6}{107}} \approx 13.92\,.$$

The exact value is 13.94 minutes.■

7.7 Tech Notes

The *Excel* interface has been enhanced with an additional menu, *Stats*. This menu provides many of the common functions needed in univariate statistics, as well as some of the graphing tools discussed in this chapter. This section illustrates the use of these tools.

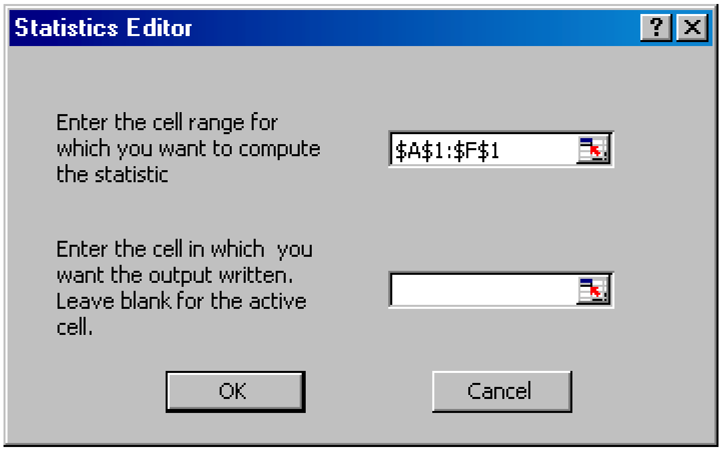
Example 7.9: Using the *Stats* menu**.**

Solution:

The table below displays some data on a worksheet for which we want to compute the mean.

	A	B	C	D	E	F
1	3	7	-5	15	12	8

Click on the *Stats* menu and select Mean. The dialog box below opens. This box requests two pieces of information. First, you need to enter the reference for the cells in your data. This can be done either by typing the reference (in this case simply a1:f1) or selecting the data with the mouse, which was done for this illustration. The latter method writes the references using the absolute reference system, as discussed in Appendix A. The second piece of information requested is the cell reference of the location where you want the answer written. You can select a cell or leave the box blank, in which case the active (highlighted) cell will be chosen (which here was $A$3).



Statistics Editor

Enter the cell range for which you want to compute the statistic

$A$1:$F$1

Enter the cell in which you want the output written. Leave blank for the active cell.

Clicking OK places the answer together with the word “Mean” in the worksheet as shown below. The active cell automatically moves to the cell below the word “Mean.”

	A	B	C	D	E	F
1	3	7	-5	15	12	8
2						
3	Mean	6.666667				

The numerical value displayed in B3 is computed using the *Excel* command, =average(a1:f1).■

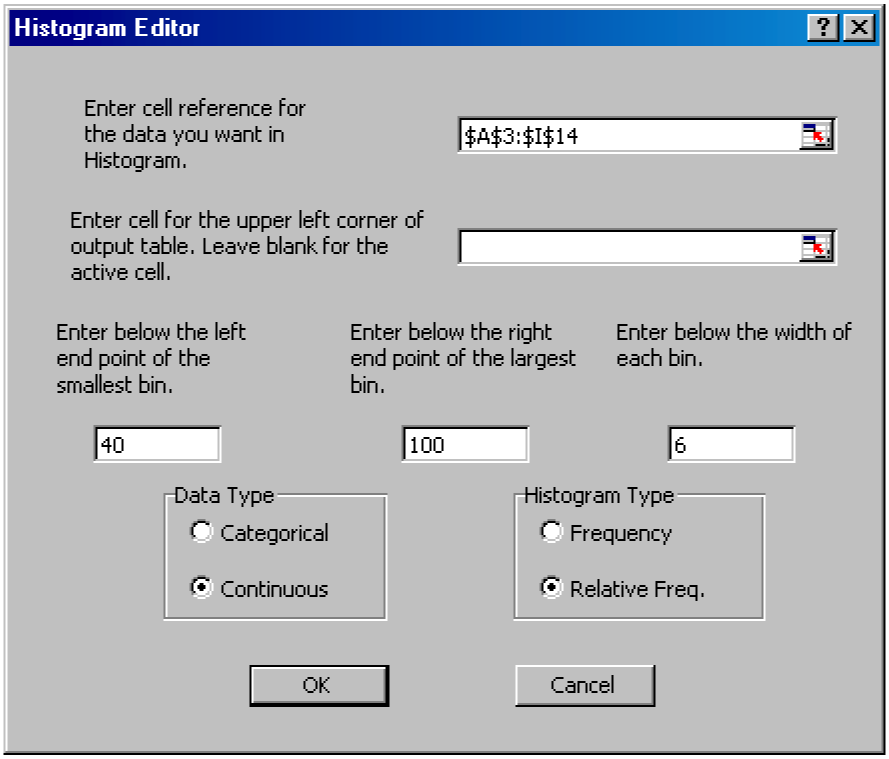
One of the reasons for using the *Stats* menu for such computations is to emphasize the importance of proper labeling of all computations on a spreadsheet. Although examining a worksheet on the computer and paying attention to the formula bar will usually reveal the logic behind the computations, without proper labeling it is virtually impossible to figure out what the computations mean from printed output.

The *Stats* menu also provides for the ability to generate histograms. A dialog box (also called a user form) prompts you for the necessary information. Here the output is more extensive, consisting of a table and a graph; there are also several choices to be made in the dialog box in addition to the ones we have already alluded to in our general discussion of histograms. These are discussed in Example 7.10.

Example 7.10: Using the Histogram command**.**

Solution:

The figure below shows the settings for the Histogram Editor that were used to generate Figure 7.1 in this chapter.



Histogram Editor

Enter cell reference for the data you want in Histogram. $A$3:$I$14

Enter cell for the upper left corner of output table. Leave blank for the active cell.

Enter below the left end point of the smallest bin. 40

Enter below the right end point of the largest bin. 100

Enter below the width of each bin. 6

Data Type
○ Categorical
● Continuous

Histogram Type
○ Frequency
● Relative Freq.

Figure 7.6

- **Data Reference** - The data in Table 7.1 was located in cells A3:I14 and this reference is entered in the top box of the editor.

- **Output Table Reference** - The editor generates a table (see Table 7.6 below) and you need to specify the cell reference for the top left corner. You will be warned if the generated table will overwrite any data already on the sheet. You can avoid the display of the latter message by making sure there is no data in the region below and to the right of the chosen anchor point. In Figure 7.6 the output table reference has been left blank, meaning the active cell is selected as the anchor.

- **Bin Endpoints and Bin Width** - We have already discussed the guidelines for selecting these. ***Because of the way the results are tabulated it is usually a good idea if the left endpoint is strictly smaller than the smallest data value***.

- **Data Type** - This selection affects how the data (horizontal) axis is labeled. A data set is classified as *categorical* if its values fall into a limited number of discrete categories. By contrast, *continuous* data in principle takes on all possible values in some interval. The examples in this chapter have been of the latter sort. When data is considered categorical, we tabulate the outcomes in each category and place a label with the corresponding data value below each bar on the histogram. For continuous data, we tabulate the values in intervals and use the middle of each interval as a label for the histogram bar.

- **Histogram Type** - A frequency histogram plots the actual numerical frequencies on the vertical axis, whereas using the relative frequency option plots the relative frequency (between 0 and 1) on this axis. Since either of these graphs can be obtained from the other by a change of vertical scale, they will look identical when generated by the computer. ***The advantage of a relative frequency graph is that it allows meaningful comparison between two univariate data sets that take on the same range of values***. For example, we might want to compare the grades on the same exam for students from two different classes (each with different numbers of students). A frequency histogram that simply counts the number of students scoring in a particular interval would make such a comparison meaningless.

The table that is produced by the histogram routine differs slightly in appearance from the tables we have used in the text. Table 7.6 shows a portion of the table produced by the Histogram Editor given in Figure 7.6.

Bins	*Freq*	*Rel. Freq*	*Cumul. Freq*	*Alt x Values*
40	0	0	0	37
46	1	0.009259	0.009259	43
52	10	0.092593	0.101852	49
58	14	0.12963	0.231481	55
64	10	0.092593	0.324074	61

Table 7.6

The last column is simply a list of midpoint values to be used in labeling the data axis in the continuous case. The first row counts the number of data values that are ≤ the first bin value of 40. This is of course zero, since 40 was selected as smaller than the least data value 45. The second row tabulates all data values that are in the interval (40, 46]; the third row all data values in the interval (46, 52], etc.■

In addition to histograms, the Box Plot command on the *Stats* menu provides for the generation of multiple box plots. The reader should have no trouble following the simple directions provided in the dialog boxes for this command.

Although *Excel's* internal commands for finding medians and quartiles do not require that the data be sorted, it is sometimes convenient to do so for other purposes. For example, having the data sorted makes it easier to perform the tabulations needed to test the Bell Curve Rule (Rule 7.1). *Excel* has a built-in sort routine that is found on the Data menu. Assuming your data is given in columns, you first select all the data you wish to sort, including any associated labels. The dialog box for the sort command asks you for the column you wish to sort and whether you wish the data arranged in ascending or descending order. The data is then rearranged according to your specifications. Remember, if you make a mistake you can always undo it!

7.8 Summary

Descriptive statistics provides methods for extracting structure from data. In this chapter we have considered some techniques that are useful when analyzing ***univariate data***. Graphical techniques organize the data visually. These representations include ***histograms*** and ***box plots***, both of which try to inform the viewer regarding the distribution of data values. The histogram does this through counting the number of data values (or the frequency of such values) in equally spaced bins from the largest to smallest data value. The box plot visually encodes some numerical statistics that are useful markers for the center and spread of the data, namely the ***median***, ***quartiles***, and ***largest*** and ***smallest*** values. In addition to the latter numerical quantities, the ***mean*** and ***standard deviation*** are convenient summary statistics that also play an important role in statistical prediction. In particular, when the histogram of a data set is approximately bell shaped, the ***Bell Curve Rule*** shows how the mean and standard deviation can be used to assess which data values might be considered unusually large or small.

7.9 Exercises

1. The following data gives the heights in inches of the male singers of the New York Choral Society (1979), according to their singing parts. The parts are arranged according to decreasing pitch of the singing range.

1st tenor:										
64	65	66	66	66	67	68	68	69	70	70
71	71	72	72	73	74	76				

2nd tenor:										
66	68	68	69	69	69	69	69	69	70	71
71	71	71	71	71	73	76				

1st Bass:										
66	66	68	68	68	68	68	68	69	69	69
70	70	70	70	70	70	70	70	71	71	71
71	71	71	72	72	72	72	72	72	73	73
74	75	75	75							

2nd Bass:										
66	67	67	68	69	70	70	70	70	71	72
72	72	72	72	72	72	74	74	74	74	75
75	75									

Table 7.7

Prepare a histogram for the heights of all tenors (combining 1st and 2nd) and another for all the basses. Discuss whether there is any evidence that the people who sing the lower pitched parts tend to be taller than those who sing higher pitches.

2. The Table 7.8 below gives 100 measurements of the speed of light in air that were made in 1879 by the American physicist A. A. Michelson. The data has been arranged in increasing size order and not in the order in which it was obtained. The units are millions of meters per second. Prepare a histogram of the deviations of each measurement from 299. Does the histogram display a bell-shaped appearance?

Michelson Speed of Light Data									
299.62	299.65	299.72	299.72	299.72	299.74	299.74	299.74	299.75	299.76
299.76	299.76	299.76	299.76	299.77	299.78	299.78	299.79	299.79	299.79
299.8	299.8	299.8	299.8	299.8	299.81	299.81	299.81	299.81	299.81
299.81	299.81	299.81	299.81	299.81	299.82	299.82	299.83	299.83	299.84
299.84	299.84	299.84	299.84	299.84	299.84	299.84	299.85	299.85	299.85
299.85	299.85	299.85	299.85	299.85	299.86	299.86	299.86	299.87	299.87
299.87	299.87	299.88	299.88	299.88	299.88	299.88	299.88	299.88	299.88
299.88	299.88	299.89	299.89	299.89	299.9	299.9	299.91	299.91	299.92
299.93	299.93	299.94	299.94	299.94	299.95	299.95	299.95	299.96	299.96
299.96	299.96	299.97	299.98	299.98	299.98	300	300	300	300.07

Table 7.8

3. The data below is from a study at the National Institute of Standards and Technology. The study was used to prepare a certification value for the transmittance of an optical filter. The units of the data are not given in the source (see *transmitt.xls*). Prepare a histogram for the deviations of the data from 2.

Transmittance Data									
2.0013	2.0013	2.0013	2.0014	2.0014	2.0014	2.0014	2.0014	2.0015	2.0015
2.0015	2.0015	2.0015	2.0015	2.0015	2.0015	2.0015	2.0016	2.0016	2.0016
2.0016	2.0017	2.0017	2.0017	2.0018	2.0018	2.0018	2.0018	2.0018	2.0019
2.0019	2.0019	2.0019	2.002	2.002	2.002	2.0021	2.0021	2.0022	2.0023
2.0024	2.0024	2.0025	2.0025	2.0026	2.0026	2.0026	2.0026	2.0027	2.0027

4. Find the mean, standard deviation, median and quartiles for the normal daily maximum temperatures each month at New York and New Orleans as given in the following table.

	Normal Maximum daily Temperatures (Fahrenheit degrees)					
	Jan.	Feb	March	April	May	June
New York	37.6	40.3	50	61.2	71.7	80.1
New Orleans	60.8	64.1	71.6	78.5	84.4	89.2
New York	85.2	83.7	76.2	65.3	54	42.5
New Orleans	90.6	90.2	86.6	79.4	71.1	64.3
	July	August	September	October	November	December

5. Consider the data 3, -2, 7, 10, 5, and 3. (The value 3 occurs twice.)

   a) Find the median, mean, standard deviation and variance for this data. Do the calculation by hand, using a calculator if necessary only to perform arithmetic.

   b) Suppose 15 is added to each of the data values. What are the new values for the median, mean, standard deviation, and variance? (Hint: You can do this without repeating all the details of the calculation in a).)

   c) Suppose each data value is multiplied by -5. What are the new values of the median, mean, standard deviation, and variance?

   d) State general rules based on the observations in parts b) and c).

6. The deviations from 299 for the data in Table 7.8 have a mean of .852 and a standard deviation of .079. Using the results of exercise 5, what are the mean, standard deviation, and variance for the actual data in Table 7.8?

7. Test the conclusions of the Bell Curve Rule (Rule 7.1) for the data you analyzed in exercise 2 (See exercise 6 for the values of $\overline{x}$ and s).

8. For the university expense data (Table 7.4) the mean value is $34.8 (thousand) and the standard deviation is $17.9 (thousand). Examine the percent of the data that lies within one and two standard deviations of the mean. How well do the results agree with the Bell Curve Rule (Rule 7.1)? Would you expect them to?

9. Based on the data in Table 7.1, the average time between eruptions of Old Faithful is 72.6 minutes. If someone just missed viewing an eruption and you were to advise that person to return in 73 minutes, do you think it likely he or she would turn out to be pleased with your advice? Could you think of how you might improve the advice? (We will consider this question again in Chapter 9.)

10. a) Using the data in exercise 1, find the median and quartiles for the heights of tenors and basses (grouping all tenors together and all basses). Use these to construct box plots over the same scale for the singers' heights for the two vocal parts. Do the box plots suggest any

relation between the vocal range and singer height? What might account for the relationship you observe?

b) Repeat part a) using the file *singers_99.xls* giving the heights of singers in the same chorus in 1999. Does the relationship observed for the male singers apply to the female singers?

11. Prepare box plots for the normal monthly high temperatures in New York and New Orleans (exercise 4). What do the plots show regarding the annual variation in high temperature in each of the cities?

12. The table below gives the ages of patients admitted to a hospital emergency room during a certain evening:

10	12	21	22	22	23	32	34	35	35
36	36	37	38	39	43	44	45	45	45
45	46	53	54	55	55	56	57	60	64

a) Construct a box plot for the data.

b) From your box plot, would you expect the mean age of the patients to be larger or smaller than the median age? Explain.

13. The following table shows the distribution of grades on a certain exam

Grade	# of Students
(40,50]	3
(50,60]	3
(60,70]	8
(70,80]	11
(80,90]	4
(90,100]	3

a) Using the table construct a labeled histogram.

b) From the table, estimate the medians, the quartiles and the mean grade.

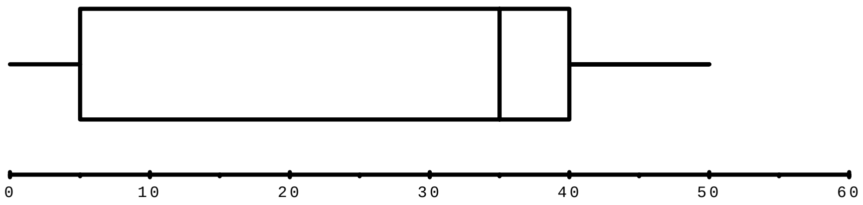
14. Prepare a frequency table for the signers' age data (Table 7.3). Use the table to estimate the mean and standard deviation for the data. Compare to the exact values $\overline{x} = 44.2$ and $s = 10.2$.

15. a) The table below from *Excel* summarizes the data on university expenditures per student for the complete listing of 201 schools given in the file *univ_exp.xls*. Using the table, find estimates for the median, quartiles, mean, standard deviation and variance of the complete data set.

Bins ($000)	Freq	Rel. Freq	Cumul. Freq
0	0	0.000	0.000
10	41	0.204	0.204
20	103	0.512	0.716
30	34	0.169	0.886
40	6	0.030	0.915
50	7	0.035	0.950
60	5	0.025	0.975
70	2	0.010	0.985
80	2	0.010	0.995
90	0	0.000	0.995
100	0	0.000	0.995
110	1	0.005	1.000

b) Using the table in a) prepare a relative frequency histogram for this data. How does the distribution compare with the histogram for the top 50 rated schools (Figure 7.3)?

16. Consider the box plot shown below:



a) What are the median, 1st quartile and 3rd quartile?

b) What percent of the data lie between 5 and 40?

c) What is the range of the data?

d) Would you expect the mean to be less than, equal to, or greater than the median? Explain.

17. The table below gives the average height (in meters) of mature specimens from 39 different species of oaks. (See file *acorn.xls*.)

Heights (in Meters) of Mature Oaks of 39 Different Species									
0.3	1	3	3	4	6	9	9	9	9
11	13	15	15	15	17	17	18	18	18
20	21	21	23	23	23	23	24	24	24
24	24	25	26	27	27	27	30	30	

a) Find the median, 1st quartile and 3rd quartile and draw a box plot for the data.

b) Construct a <u>relative frequency</u> histogram for the data.

c) Based on your answers to a) and b), would you expect the mean to be smaller or larger than the median? Explain without actually computing the mean.

18. Suppose the mean height of six boys is 64.3 inches. A 7th boy whose height is 60 inches joins the group. What is the average height of the seven boys?

19. Suppose you attain an average of 73 on three exams during the first half of a course. Five exams are scheduled for the second half. What must your average be on the remaining five exams to finish with an average of 80 for the entire eight exams?

20. Suppose a data collection S consists of the six numbers 7, 3, 5, 2, 1, 2.

a) Find a new data value which when adjoined to S will give a data set S* having a mean of 4.

b) Find the median of the original collection S.

c) Suppose any number whatsoever (not necessarily a whole number) is adjoined to the collection S, producing a new collection S'. Explain why the median for S' differs from the median of S by at most 1/2.

21. a) The file *oldFaith.xls* contains 230 observations of eruptions of Old Faithful. Each observation records the duration of the eruption (Column A) and the time to the next eruption (Column B). (The data in Table 7.1 were obtained by random sampling from column B). Use *Excel* to prepare a properly labeled histogram of the length of the eruptions. Does the histogram exhibit a bimodal shape?

b) Find the mean values of the duration of the eruption and the time between eruptions. Find the standard deviation for each. Test whether the conclusion of the Bell Curve Rule (Rule 7.1) holds for each of these variables.

22. Use *Excel* to prepare properly labeled box plots of the length of the eruption and the time to next eruption for the Old Faithful data in *oldFaith.xls*. Should the box plots be prepared on the same horizontal axis? Why or why not? Can you conclude anything regarding bimodality from the box plots?

23. The file *hibernat.xls* contains data on the longevity of 144 hamsters. Prepare a histogram and a box plot for the longevity. How would you characterize the shape of the distribution. Specifically, comment on the skewness, symmetry, and modality (peaks).

24. Use *Excel* to prepare a histogram of the lottery data in the file *lottery.xls*. On the basis of this data would you conclude that any of the possible pick-three numbers has an equal chance of being selected?

25. Using the file *Pi_digits.xls*, prepare a histogram of the first 5000 digits of the mathematical constant π. Does the histogram provide evidence in support of the (unproven) conjecture that each digit appears with a frequency of approximately 1/10 in the infinite decimal expansion of this number?

26. a) Use *Excel* to prepare a frequency distribution table and a histogram for the complete grade data in the file *grades.xls*. The histogram should have an appropriate title and the axes should be labeled informatively.

    b) Use *Excel* to find the mean and the median for the entire data set. Print out at most two pages, containing the table, the graph and the values of the mean and the median.

    c) What estimate could you give for the median, based on the frequency table? Explain.

    d) What estimate could you give for the mean, based on your frequency table? Explain.

    e) Using your frequency table, determine the fraction of students who received a grade less than or equal to 70. What fraction received a grade greater than 80?

    f) Explain how, by examining only the histogram, you might reasonably conclude that the mean is smaller than the median for this data.

27. Using the file *homers.xls*, prepare a histogram and box plot for Mark McGwire's 1998 home run data. Does the data appear to be skewed? If so, what might explain the skewness?

28. Using the file *vitaminc.xls*, prepare box plots (on the same axis) showing the survival data for each of the five types of cancers given in the study. What conclusions can you draw with regard to the effectiveness of ascorbate in the situation studied by the authors?

29. The file *ran_norm.xls* contains “random normal” numbers generated by the computer. Prepare a histogram and determine how well these numbers fit the Bell Curve Rule (Rule 7.1).