

# Elementary Statistics

Brian E. Blank

January 19, 2016

FIRST PRESIDENT UNIVERSITY PRESS



# Chapter 2. Data—Numerical

## 2.1 Histograms

In this chapter we will concentrate on *numerical*, or, synonymously, *quantitative* variables. Remember: these are variables with values that are numbers that are used as numbers (as opposed to variables, such as *Gender*, that have values that are not numbers and variables, such as *Zip Code*, that have values that are numbers that are not used as numbers). How do we display numerical variables?

In Chapter 1 we learned two simple methods for visualizing the distributions of categorical variables: pie charts and bar charts. In both types of graphic, a geometric figure, be it a sector of a disk or a bar, is used to convey the frequency count of each possible category. Without modification, this method of illustration cannot usefully be applied to numerical variables. Consider, for example, beryllium-11 ( $^{11}\text{Be}$  in the literature of chemistry), which is an unstable isotope of beryllium-9. Suppose that we have an initial supply of beryllium-11 and conduct a statistical study in which the population consists of events: each event is the decay of one atom of beryllium-11 to stable form. Each such event gives rise to a case of the study. For each case, the numerical variable *Time of Decay* records the time elapsed between the start of the study and the radioactive decay that is the object of the case. Notice that the values of *Time of Decay* are numbers that can be used as numbers. For example, if we subtract the value of *Time of Decay* for one case from that of the next case, then we obtain a number that has a genuine numerical meaning, namely the time that has elapsed between two consecutive radioactive decays. In other words, *Time of Decay* is a numerical variable.

Suppose that we terminated our decay vigil after 26 seconds, and that 10,000 decays occurred in those 26 seconds. In other words, we recorded 10,000 observations of *Time of Decay* during our study. How do we depict the values of *Time of Decay*? Chances are, if we measured *Time of Decay* sufficiently accurately, then we recorded 10,000 different times. If so, then the frequency for each of the 10,000 values of *Time of Decay* would be 1. Do we draw 10,000 bars all with unit height? What would be the point?

What we do instead is to divide the 26 second interval containing all the values of *Time of Decay* into a convenient number, say 13, of subintervals, which, in the jargon of statistics, are called *bins* or *class intervals*. We then count the number of decays occurring in each bin. Suppose that our frequency counts of beryllium-11 decays are as follows:

Decay Time (sec)	[0,2)	[2,4)	[4,6)	[6,8)	[8,10)	[10,12)	[12,14)	[14,16)	[16,18)	[18,20)	[20,22)	[22,24)	[24,26)
Number of Decays	2	21	140	520	1207	1889	2132	1816	1208	646	282	104	33

Table 2.1.1: Beryllium-11 Decays

Remember that the notation  $[a, b)$  means that the interval contains the left endpoint  $a$  but not the right endpoint  $b$ . The “bar chart” of the frequencies given in Table 2.1.1 is presented in Figure 2.1.1.

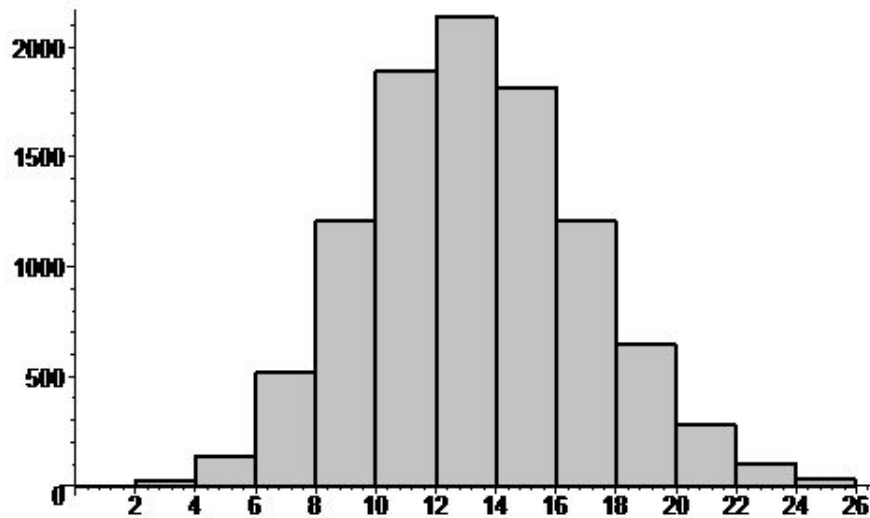


Figure 2.1.1: Number of Beryllium-11 Decays

The graph in Figure 2.1.1 is called a *histogram* or, a bit more precisely, a *frequency histogram* (because we can also graph *relative frequency histograms* and *percentage histograms*). The endpoints of the class intervals are called *class limits*. Specifically, a left endpoint is called a *lower class limit* and a right endpoint is called an *upper class limit*. In a histogram, each value of the numerical variable whose frequencies are plotted must fall into exactly one bin. One way to ensure that property is to choose the bins so that no data value equals any class limit. However, if the data set is large, it can be tedious to carry out these choices. A simpler solution is to adopt the common (but not universal) convention of placing a value in the larger bin if it equals the class limit that separates two consecutive bins. Thus, in the beryllium-11 example that we have been considering, we chose the bins so that each includes its left endpoint but excludes its right endpoint. By specifying the bins in this way, we have conformed to the convention of placing a value that fall on the upper limit of one class interval in the next class interval to the right. (For example, the value 16 does not belong to the class interval  $[14,16)$  but to  $[16,18)$ , the next higher bin.)

The width  $h = b - a$  of a class interval  $[a, b)$  is called the *class width* or *class size*. For most histograms, the class widths are the same for all class intervals. Some textbooks *require* the class intervals of a histogram to be equally wide. Other texts are more flexible, allowing that, in some exceptional circumstances, variable class widths can have advantages. Bear in mind, however, that variable class widths necessarily result in a violation of the Area Principle when the heights of the bars are plain vanilla frequencies. This problem can be averted by plotting the frequencies per unit width instead of the frequencies, but that complication can add its own confusion.

To ensure that the bins do not omit any data values, the lower class limit of the leftmost class interval must be less than or equal to the smallest data value  $x_{\min}$ , and the upper class limit of the rightmost class interval must be greater than or equal to the largest data value  $x_{\max}$ . The difference  $x_{\max} - x_{\min}$  is called the *range* of the data set. Usually the least lower class limit is chosen to be only slightly less than  $x_{\min}$ , if at all, and the greatest upper class limit is chosen to be only slightly greater than  $x_{\max}$ . As a result, the width of the histogram is only slightly greater than the range of the data.

Many textbooks suggest that the number  $n$  of bins should satisfy the inequalities  $5 \leq n \leq 20$ . This advice follows the *Goldilocks Principle*:  $n$  should be not too big and not too small. In fact, statisticians have devoted plenty of thought to the number of bins that should be used in a histogram, but they have not come to a consensus. For those who are interested, some of the proposals may be found in the next subsection.

## Class Width Formulas (Optional)

If, as is usually the case, the class intervals of a histogram share a common width  $h$ , and if  $n$  is the number of bins, then it follows that  $n \cdot h$  is not less than the range  $x_{\max} - x_{\min}$ . In general, these two numbers will not differ substantially. That is,  $n \cdot h \approx x_{\max} - x_{\min}$ . Consequently, choosing the number  $n$  of bins is equivalent to choosing the bin width  $h$ . As mentioned, this choice is somewhat arbitrary, but there are some rules of thumb that are in use. For example, the default in *Excel* is to assign a value to  $n$  according to the formula

$$n = \sqrt{N},$$

where  $N$  is the number of observations of the numerical variable. Some judgment must be used. For our beryllium-11 example, the choice  $n = \sqrt{10,000}$ , or  $n = 100$ , would be inconveniently large for our purposes.

In 1926, the American statistician Herbert Sturges proposed the following formula<sup>1</sup>:

$$n = \lceil 1 + \log_2(N) \rceil \approx \lceil 1 + 3.3219 \cdot \log_{10}(N) \rceil.$$

The notation  $\lceil x \rceil$  is read as “the ceiling of  $x$ ” and is equal to the integer obtained by rounding  $x$  up. For example,  $\lceil \pi \rceil$  is 4. In the beryllium-11 example under discussion, Sturges’s formula leads to  $n = \lceil 1 + 3.321928095 \cdot \log_{10}(10,000) \rceil = 15$ , a value that is quite close to the value  $n = 13$  that we used. Again, judgment decides the issue: the choice of 13 leads to conveniently simple integer endpoints for the bins.

There are several other formulas that have been proposed for the number  $n$  of bins, and in many of the alternative formulas, such as the Rice formula  $n = \lceil 2N^{1/3} \rceil$ , the value of  $n$  is approximately proportional to  $N^{1/3}$ .

## Histogram Conventions

Despite the fact that histograms are composed of bars, they are generally not considered to be bar charts. For example, one textbook<sup>2</sup> states,

“Although a bar chart and a histogram may look somewhat similar, they are not the same display. You can’t display categorical data in a histogram or quantitative data in a bar chart.”

A contrarian might argue these statements. The process of binning the values of a numerical variable, it might be claimed, creates categories: the values of the numerical variable are categorized according to the bins they fall in. This point of view is not far-fetched. *Age*, for instance, is a numerical variable, but binning the values of *Age* amounts to the creation of the categorical variable, *Age Group*. *Height* is a numerical variable, but here too, binning might be perceived as categorizing. For example, suppose that height is measured in inches and we choose the bins  $[0,67)$ ,  $[67,77)$ , and  $[77, \infty)$ . For the purposes of becoming a police officer in many jurisdictions, these bins are the categories, Too Short, Acceptable Height, and Too Tall. Supposing that 20 applicants to the police academy with heights between 57 and 87 inches are rejected. Suppose further that 7 of them were too short, 3 were too tall, and 10 were of an acceptable height but were rejected for other reasons (criminal record, weight not proportional to height, surly demeanor, intermittent explosive disorder, ...). Do we say that a frequency graph of this data set is a histogram or a bar chart? To some extent it depends on how we label the horizontal axis: see Figure 2.1.2, in which the data is illustrated by a histogram, and Figure 2.1.3, in which the data is displayed by a bar chart. Apart from the axis labeling, the two figures look very much the same.

<sup>1</sup>The expression advocated by Sturges for the number of classes is properly known as *Sturges’s Formula*, but near misses are common in the literature. The author retains textbooks the way hoarders hoard hoarded things. In his collection of statistics texts, there are references to both Sturges’ formula and Sturge’s formula.

<sup>2</sup>Stats: Data and Models, Richard de Veaux, Paul Velleman, Davis Bock, Pearson

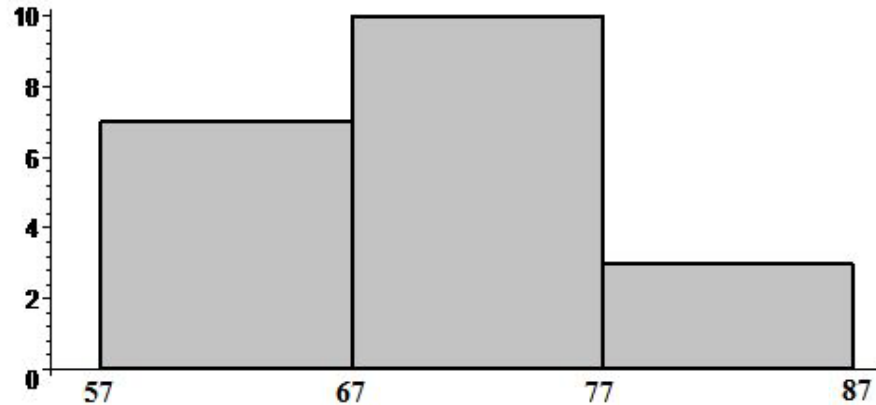


Figure 2.1.2: Histogram of Rejected Police Academy Applicants

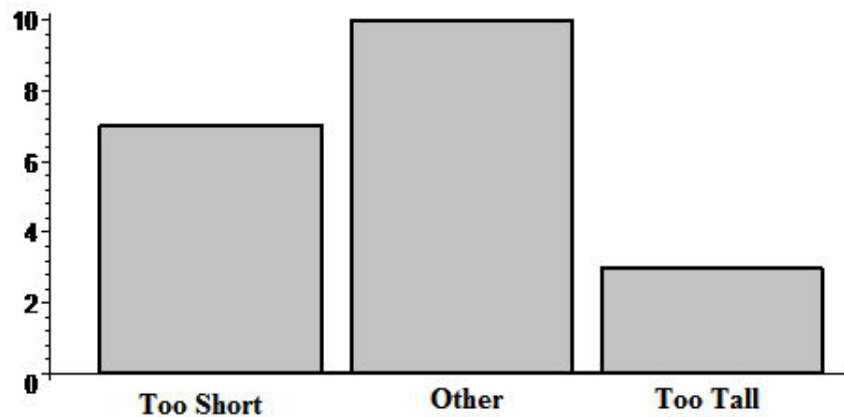


Figure 2.1.3: Bar Chart of Rejected Police Academy Applicants

It is worth using Figures 2.1.2 and 2.1.3 to point out a cosmetic difference between histograms and bar charts. In a histogram, the right side of every rectangle abuts the left side of the next rectangle, as can be seen in Figure 2.1.2. Thus, the only way there can be a gap between two rectangles of a histogram is if the bin(s) between them is(are) empty and therefore has(have) 0 height.<sup>3</sup> On the other hand, it is standard practice to position the rectangles of bar charts so that no two are contiguous, as in Figure 2.1.3. Furthermore, the rectangles of histograms are positioned with an inviolable order determined by the increasing order of the bin ranges. By contrast, the rectangles of a bar chart can be placed in any order.<sup>4</sup> For instance, in Figure 2.1.3, the last two bars might be interchanged so that the height rejections are grouped together. This swap cannot be done for the histogram in Figure 2.1.2 because the bin  $[67,77)$  must precede the bin  $[77,87)$ .

<sup>3</sup>One regrets that, in Statistics, it is often the case that examples can be found to contradict generic truths. At the time of this writing, Wolfram MathWorld, “the web’s most extensive mathematics resource” (according to Wolfram) illustrates its article on histograms with a histogram that separates its rectangles with gaps. <http://mathworld.wolfram.com/Histogram.html> The histogram shown on the cited web page was not actually created by Wolfram’s premiere product, *Mathematica*, which does produce correct histograms.

<sup>4</sup>Although the rectangles of a bar chart can be placed in any order, it is often desirable to arrange them from left to right in order of decreasing frequencies. That allows the viewer to take in at a glance the most significant categories. Such a bar chart is called a *Pareto chart*.

## Plotting Alternatives

Although histograms are the primary graphic for illustrating numerical variables, there are alternatives. One is called a *stem-and-leaf plot*. Consider the mileage data of Table 2.1.2 below.

Vehicle	EPA Combined City/Highway Mileage (mpg)
2014 Toyota 4Runner 2WD 6 cyl, 4.0 L, Automatic (S5)	19
2014 Toyota 4Runner 4WD 6 cyl, 4.0 L, Automatic (S5)	18
2014 Toyota Avalon 6 cyl, 3.5 L, Automatic (S6)	25
2014 Toyota Avalon Hybrid 4 cyl, 2.5 L, Auto(AV-S6)	40
2014 Toyota Camry 4 cyl, 2.5 L, Automatic (S6)	28
2014 Toyota Camry 6 cyl, 3.5 L, Automatic (S6)	25
2014 Toyota Camry Hybrid LE 4 cyl, 2.5 L, Automatic	41
2014 Toyota Camry Hybrid XLE/SE 4 cyl, 2.5 L, Automatic	40
2014 Toyota Corolla 4 cyl, 1.8 L, Automatic 4-spd	31
2014 Toyota Corolla 4 cyl, 1.8 L, (var. gear ratios)	32
2014 Toyota Corolla 4 cyl, 1.8 L, Manual 6-spd	31
2014 Toyota Corolla LE Eco 4 cyl, 1.8 L, Automatic	35
2014 Toyota Prius 4 cyl, 1.8 L, (var. gear ratios)	50
2014 Toyota Prius v 4 cyl, 1.8 L, (var. gear ratios)	42
2014 Toyota RAV4 4 cyl, 2.5 L, Automatic (S6)	26
2014 Toyota RAV4 AWD 4 cyl, 2.5 L, Automatic (S6)	25
2014 Toyota Venza 4 cyl, 2.7 L, Automatic (S6)	23
2014 Toyota Venza 6 cyl, 3.5 L, Automatic (S6)	22
2014 Toyota Venza 4WD 4 cyl, 2.7 L, Automatic (S6)	22
2014 Toyota Yaris 4 cyl, 1.5 L, Manual 5-spd	33
2014 Toyota Yaris 4 cyl, 1.5 L, Automatic 4-spd	32

**Table 2.1.2: Combined EPA Mileage: 2014 Toyota Fleet (Trucks and Land Cruisers excluded)**

Source: U.S. Department of Energy, <http://www.fueleconomy.gov/feg/bymake/Toyota2014.shtml>

First we will bin this data using the class intervals  $[15,19)$ ,  $[20,24)$ ,  $[25,29)$ ,  $[30,34)$ ,  $[35,39)$ ,  $[40,44)$ ,  $[45,49)$ , and  $[50,54)$ . In doing so, we will arrange the data in a way that leads not only to a histogram but to a stem-and-leaf plot.

class intervals	EPA Combined City/Highway Mileage (mpg)
$[50,54)$	50
$[45,49)$	
$[40,44)$	40, 40, 41, 42
$[35,39)$	35
$[30,34)$	31, 31, 32, 32, 33
$[25,29)$	25, 25, 25, 26, 28
$[20,24)$	22, 22, 23
$[15,19)$	18, 19

**Table 2.1.3: Combined EPA Mileage: 2014 Toyota Fleet (Trucks and Land Cruiser's excluded)—Binned**

From the arrangement of the data in Table 2.1.3, we can quickly record the frequencies: 2, 3, 5, 5, 1, 4, 0, 1 for the class intervals in ascending order. These counts lead to the histogram in Figure 2.1.4.

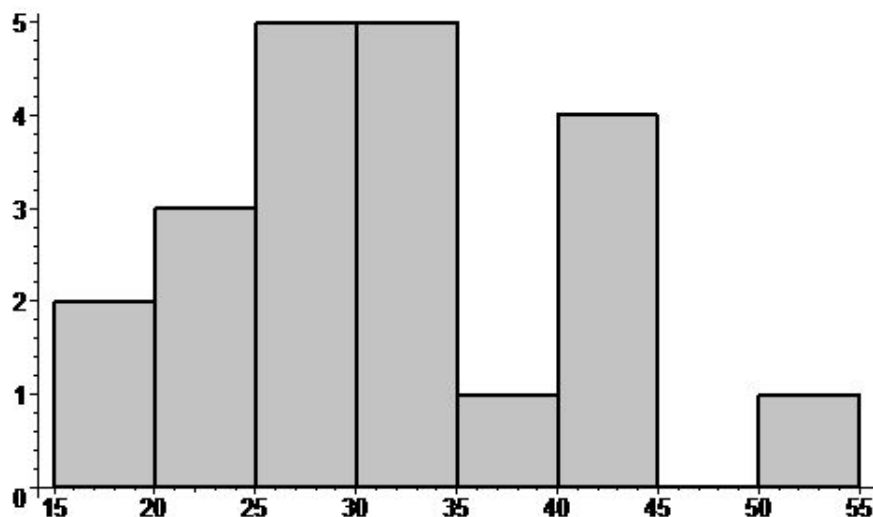


Figure 2.1.4: Histogram of EPA Combined Mileage for 2014 Toyota Fleet

Now we record the data of Table 2.1.3 in a new table, Table 2.1.4, which has the appearance of a histogram. Table 2.1.4 has the same number of rows as Table 2.1.3. For each row, the cell of the left column of Table 2.1.4 contains the first digit of the data in the right cell of the corresponding row of Table 2.1.3. These entries constitute the “stem” of the stem-and-leaf diagram we are creating. For each row of Table 2.1.4, the entry of the right cell is the sequence of second digits of the numbers in the right cell of the corresponding row of Table 2.1.3. These entries constitute the “leaf” of the stem-and-leaf diagram we are creating. The result is the stem-and-leaf diagram shown in Table 2.1.4.

5		0
4		
4		0 0 1 2
3		5
3		1 1 2 2 3
2		5 5 5 6 8
2		2 2 3
1		8 9

Table 2.1.4: Stem-and-Leaf Diagram of EPA Combined Mileage for 2014 Toyota Fleet

Now imagine that Table 2.1.4 is rotated  $90^\circ$  counterclockwise, or turn your head  $90^\circ$  clockwise: from this point of view the resulting figure has exactly the shape of the histogram that appears in Figure 2.1.4 (although the represented data decreases rather than increases from left to right). One advantage of a stem-and-leaf plot is that the details of the data are visible. Thus, we can see from Table 2.1.4 that three Toyotas have 25 mpg for their combined EPA mileage ratings. From histogram 2.1.4, all we can say about this point is that the number of Toyotas with a combined EPA rating of 25 is between 0 and 5.

Another alternative to a histogram is the *frequency polygon*. Suppose that the leftmost class interval is  $[x_0, x_1)$  and that the frequency count for this class is  $N_1$ . Instead of plotting a rectangle of height  $N_1$  and class interval  $[x_0, x_1)$  for the base, we plot the single point  $(m_1, N_1)$  where  $m_1 = (x_0 + x_1)/2$  is the midpoint of the class interval  $[x_0, x_1)$ . (The midpoint of a class interval is called a *class mark*.) This is repeated for the second bin, the third bin, and so on. Then consecutive plotted points are connected with line segments.



This is like a childhood activity, but instead of a sketch of a bunny rabbit appearing, a surrogate histogram (i.e. frequency polygon) materializes. In the beryllium-11 example, the plotted points would be (1,2), (3,21), (5,140), (7,520), (9,1207), (11,1889), (13,2132), (15,1816), (17,1208), (19,646), (21,282), (23,104), and (25,33). The resulting frequency polygon is graphed in Figure 2.1.5, superimposed on the histogram it replaces.

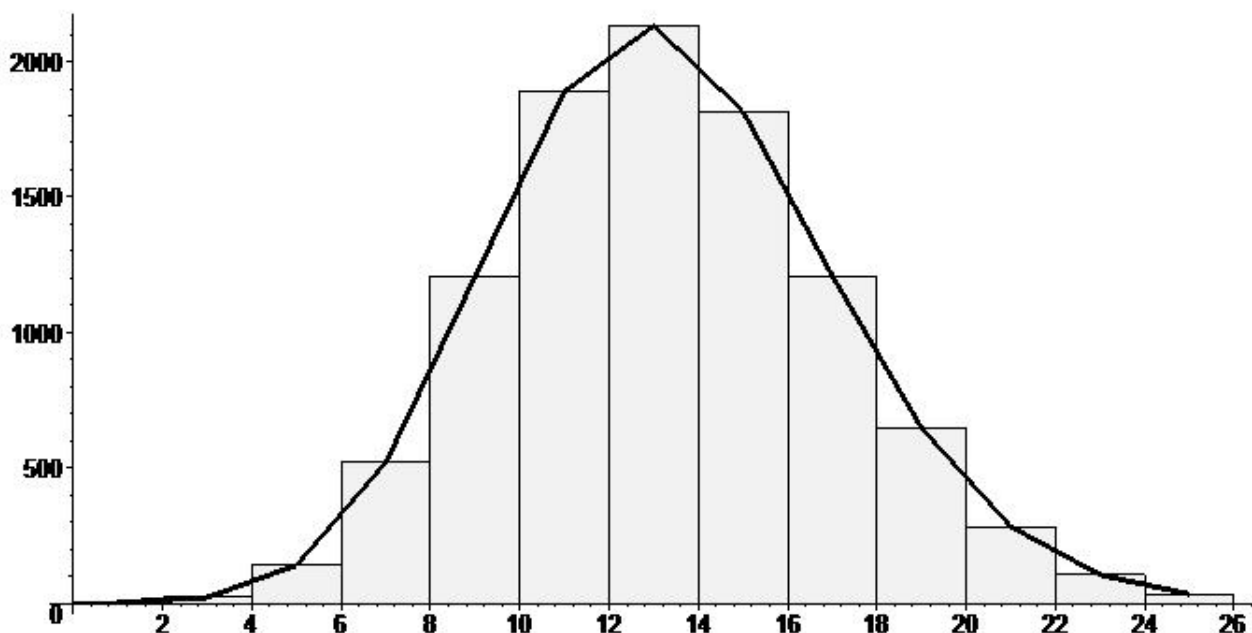


Figure 2.1.5: Frequency Polygon of EPA Combined Mileage for 2014 Toyota Fleet

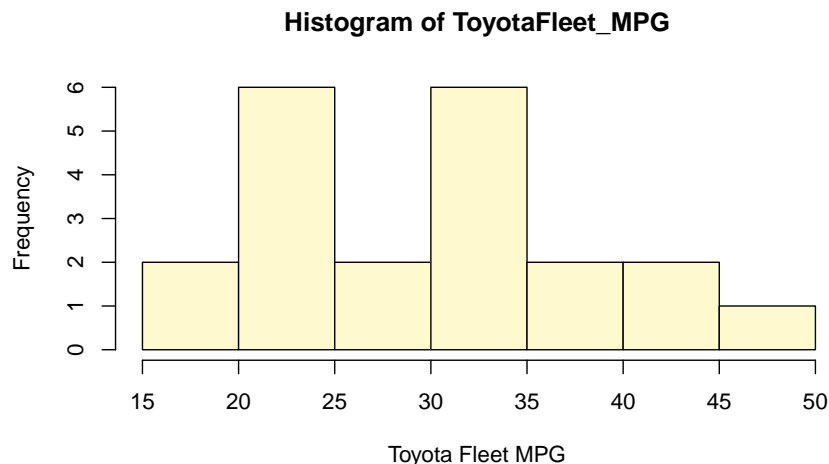
## Histograms in R (Optional)

It is very easy to graph histograms in R. All it takes is the command `hist(list)` where `list` is the sequence  $x_1, x_2, \dots, x_N$  of numbers, encoded as the vector  $c(x_1, \dots, x_N)$  in R, that constitutes the data set that is to be binned. In a less barebones call, optional parameters are included in the argument list of `hist`. Among the embellishments available to us, we can tint the bars, label the horizontal axis, and *suggest* the number of bins that are to be used. For example, the code

```
> ToyotaFleet_MPG <- c(19,18,25,40,28,25,41,40,31,32,31,35,50,42,26,25,23,22,22,33,32)
> hist(ToyotaFleet_MPG, breaks = 9, col = "lemonchiffon", xlab = "Toyota Fleet MPG")
```

begins by assigning the name “ToyotaFleet\_MPG” to the column of MPG values in Table 2.1.2. This step is merely one of convenience—in all future references to the data, we may use the short name that has been assigned rather than write out all the numbers in the data set. Here we have used R’s assignment operator `<-`. We usually will use an equality sign for assignment because it looks more natural, but `<-` has greater geek appeal and it appears to be a more popular choice. The second line of the code creates a histogram of the data. The bars of the histogram are shaded with the specified designer color `lemonchiffon`<sup>5</sup> and the horizontal axis is labeled “Toyota Fleet MPG”. An attempt to use 9 bins has been made, but, as you can see from the result of this code, Figure 2.1.6, the program rejected our choice in favor of bins that have simple integer class limits.

<sup>5</sup>See <http://research.stowers-institute.org/efg/R/Color/Chart/index.htm> for the extensive list of colors that are available in R.



**Figure 2.1.6: Histogram for 2014 Toyota Fleet (Produced using R)**

By comparing the data set with the frequency counts shown in Figure 2.1.6, you will also notice that when a data value equals a class limit, the default action is to place it in the *smaller* bin (contrary to the convention adopted in these notes). We can enforce our preferred choice by using the argument `right = FALSE`. To ensure our choice of the number of bins, we assign the sequence of desired class limits to `break` rather than the number of bins. Thus, the second line of the code

```
> ToyotaFleet_MPG <- c(19,18,25,40,28,25,41,40,31,32,31,35,50,42,26,25,23,22,22,33,32)
> nodes <- c(seq(15, 55, by = 5))
> hist(ToyotaFleet_MPG, breaks=nodes, right=FALSE, col="peachpuff", xlab="Toyota Fleet MPG")
```

assigns the name `nodes` to the sequence that begins at 5, is augmented by 5 for each new term in the sequence, and ends at 55. (This is the sequence of class limits that we used to produce Figure 2.1.4.) In the last line, a histogram is created using `nodes` as the sequence of bin endpoints. The argument `right=FALSE` means “It is false that the right endpoint of a bin belongs to the bin.” The resulting histogram, tinted with the lovely color peachpuff, is shown in Figure 2.1.7.

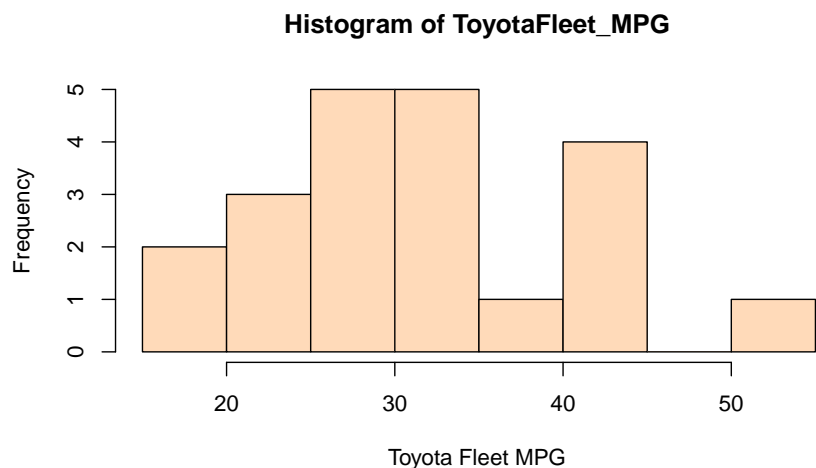
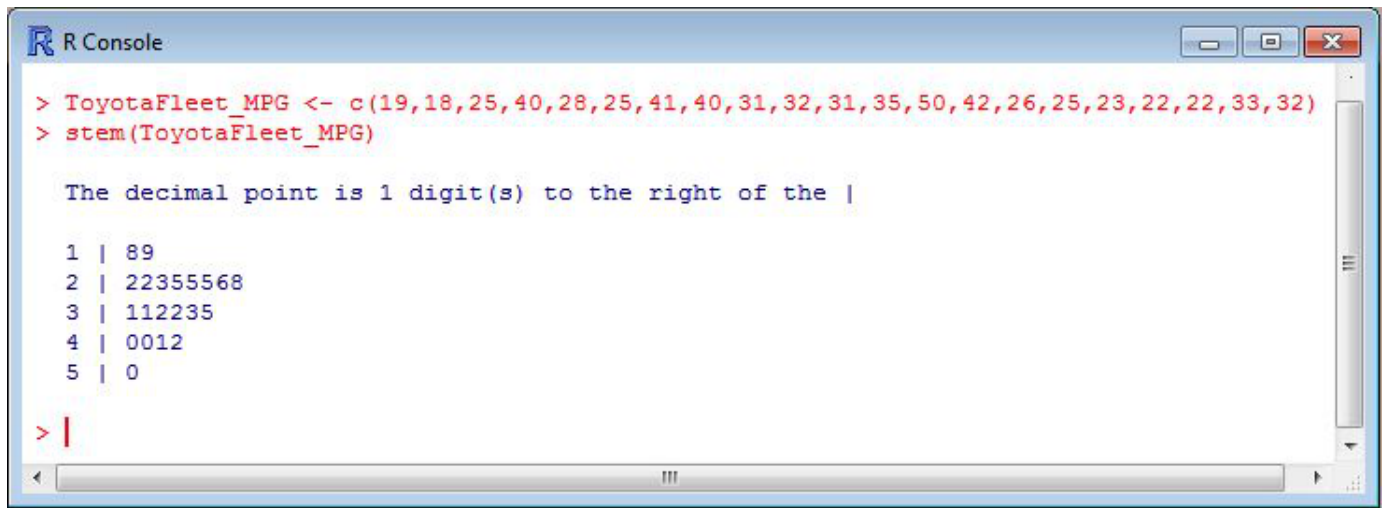


Figure 2.1.7: Histogram for 2014 Toyota Fleet (Produced using R)

## Stem and Leaf Plots in R (Optional)

It is very easy to graph stem and leaf plots in R. All it takes is the command `stem(list)` where `list` is the sequence  $x_1, x_2, \dots, x_N$  of numbers, encoded as the vector  $c(x_1, \dots, x_N)$  in R, that constitutes the data set that is to be stem-plotted. Figure 2.1.8, a screen capture of an R session, illustrates:



```

R Console
> ToyotaFleet_MPG <- c(19,18,25,40,28,25,41,40,31,32,31,35,50,42,26,25,23,22,22,33,32)
> stem(ToyotaFleet_MPG)

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

 1 | 89
 2 | 22355568
 3 | 112235
 4 | 0012
 5 | 0
> |

```

Figure 2.1.8: Stem and Leaf Plot for 2014 Toyota Fleet (R Session Screen Capture)

The returned information about the decimal point tells a person looking at the plot that the numbers are 18.0, 19.0, and so on. The following table for municipal solid waste (garbage, both landfill and recyclable) will involve numbers that do have decimal points.

Food scraps	31.7
Glass	13.6
Metals	20.8
Paper	83.0
Plastics	30.7
Rubber, leather, textiles	19.4
Wood	14.2
Yard waste	32.6
Other	8.2

Table 2.1.5: Municipal Solid Waste (Millions of Tons), U.S. E.P.A., 2007

Figure 2.1.9, a screen capture of an R session, shows that the default stem and leaf plot is not very good.

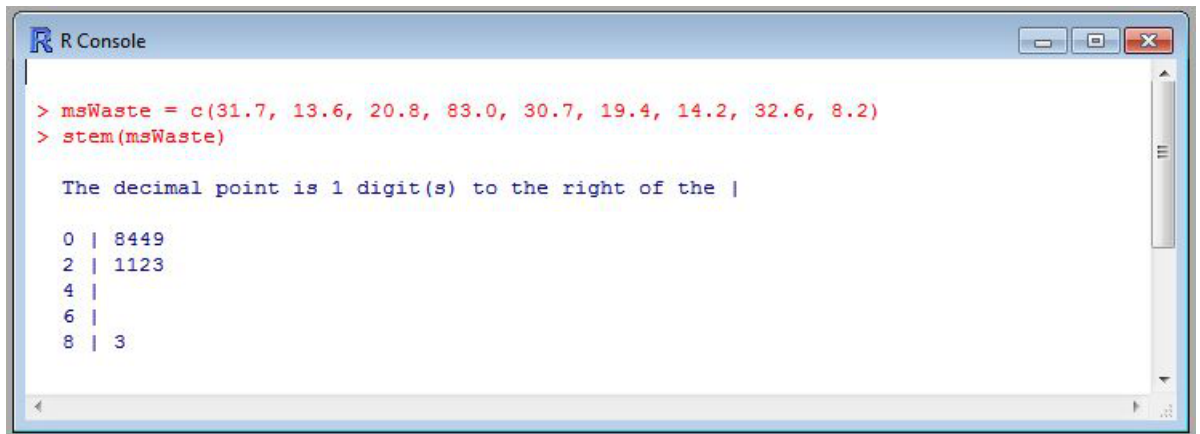


Figure 2.1.9: Stem and Leaf Plot for Municipal Solid Waste (R Session Screen Capture)

We can include the parameter `scale = ...` to do better. When the right side is filled in with 2, this argument will roughly double the height of the plot:

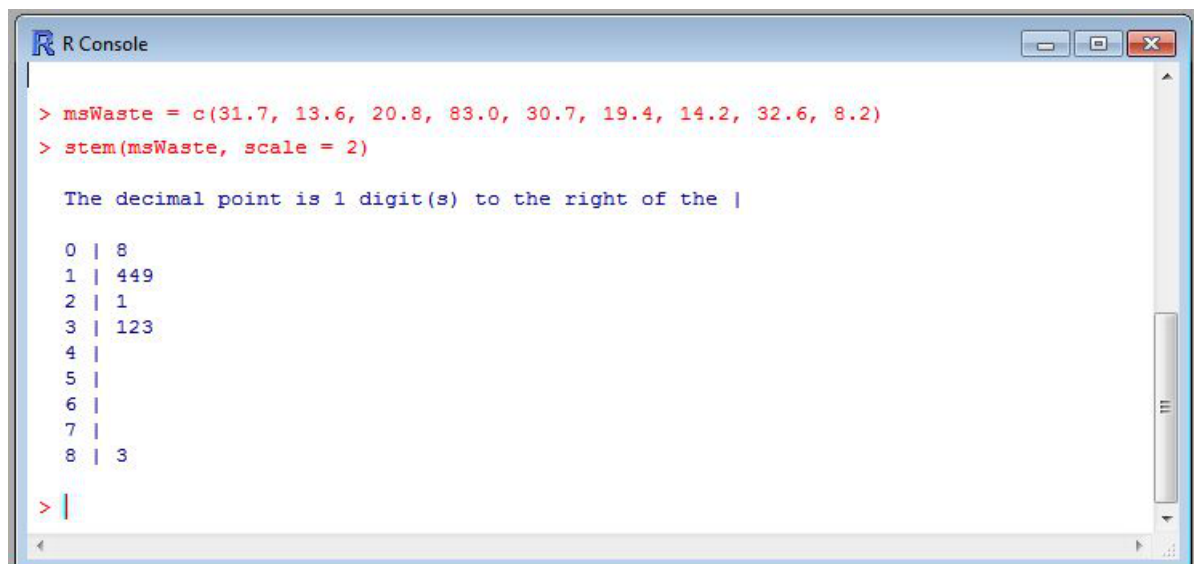


Figure 2.1.10: Stem and Leaf Plot for Municipal Solid Waste, Rescaled (R Session Screen Capture)

## 2.2 The Shape of a Distribution

The camel has a single hump;  
 The dromedary, two;  
 Or else the other way around.  
 I'm never sure. Are you?

The author is sure. It's the other way around, Ogden.

In this section we consider the peaks, the center, and the spread of a distribution. These three terms will not be given formal, precise definitions, and will be used only in their intuitive senses. That is not to say that there will not be any formal, precise definitions in this section. We will precisely define the term “mode” and then use it somewhat loosely to describe the peaks of a distribution. We will precisely define the “mean” and “standard deviation” of a distribution. They are, respectively, measures of the center and spread of a distribution. But they are not the only ones. In the next section we will define alternative measures of the center and spread.

## The Mode

Given a set of data (such as the observed values of a numerical variable), the value in the set that occurs the most frequently is called the *mode*. Thus, the mode of the observations 5, 1, 2, 2, 1, 4, 3, 2, 3, 6 is 2. Strictly speaking, referring to “the mode” rather than “a mode” is, in general, imprecise, because a data set may have more than one mode. Both 4 and 6 are modes of the data set 5, 7, 6, 5, 4, 4, 6, 3, 4, 6. The data set 5, 5, 5, 6, 6, 6, 7, 7, 7 has three modes. A data set with precisely one mode is said to be *unimodal*. A data set with precisely two modes is said to be *bimodal*. A data set with more than two modes is said to be *multimodal*.<sup>6</sup>

The definition of *mode* just given is the official one. The scope of the term, however, has been widened to encompass several related meanings. Consider the beryllium example of Section 2.1. The data set of 10,000 decays was not provided (let us give thanks), but, as was remarked, had the timings of the decays been measured with sufficient accuracy, there would have been 10,000 *different* numbers in the data set. In other words, the frequency of every value in the data set would be 1. No value in the data set would appear more frequently than the other values! What would be gained by talking about the modes of a data set consisting of 10,000 entries, each a mode? On the other hand, even though we were not furnished with the timing of any particular decay, we were presented with the frequency distribution of the set of all decays. Whether we consult Table 2.1.1 or Figure 2.1.1, we see that [12,14) is the bin in which the greatest number of decays occurred. We refer to this bin as the mode of the distribution. This usage of *mode* as the name for a range of values of a numerical variable is not precisely the same as the original usage of *mode* in which it stood for a particular value of the variable.

Until now, we have defined “mode” to refer to the greatest frequency among *all* values in the data set. In the language of calculus, a *global* maximum occurs at a mode. However, the term “mode” is also used to describe a bin that has the greatest frequency count when compared only with those of neighboring bins. In this sense, a bin might be a mode even if a distant bin has a greater frequency count. Phrased using the language of calculus, “mode” describes a bin at which a *local* maximum occurs. The terms *unimodal*, *bimodal*, and *multimodal* are extended to encompass these neologisms. Thus, Figure 2.1.1 of Section 2.1 is said to display a unimodal distribution. It is the dromedary of histograms. Figure 2.2.1, which presents a histogram of the age of onset of Hodgkin Lymphoma in the United States, is an illustration of a bimodal distribution: it is the bactrian camel of histograms.

---

<sup>6</sup>In his informative book, *Number: The Language of Science*, Tobias Dantzig noted, “The Bushmen of South Africa have no number words beyond ‘one,’ ‘two,’ and ‘many.’” More recently, it has been observed that members of the Pirahã, a small Amazonian tribe, use the “one-two-many” system of counting. See: *Numerical cognition without words: evidence from Amazonia*, P. Gordon, *Science* **306** no. 5695, 496-9. Focusing on important linguistic clues of this nature, anthropologists have theorized that modern statisticians are the descendants of a numeracy-challenged people not yet identified.

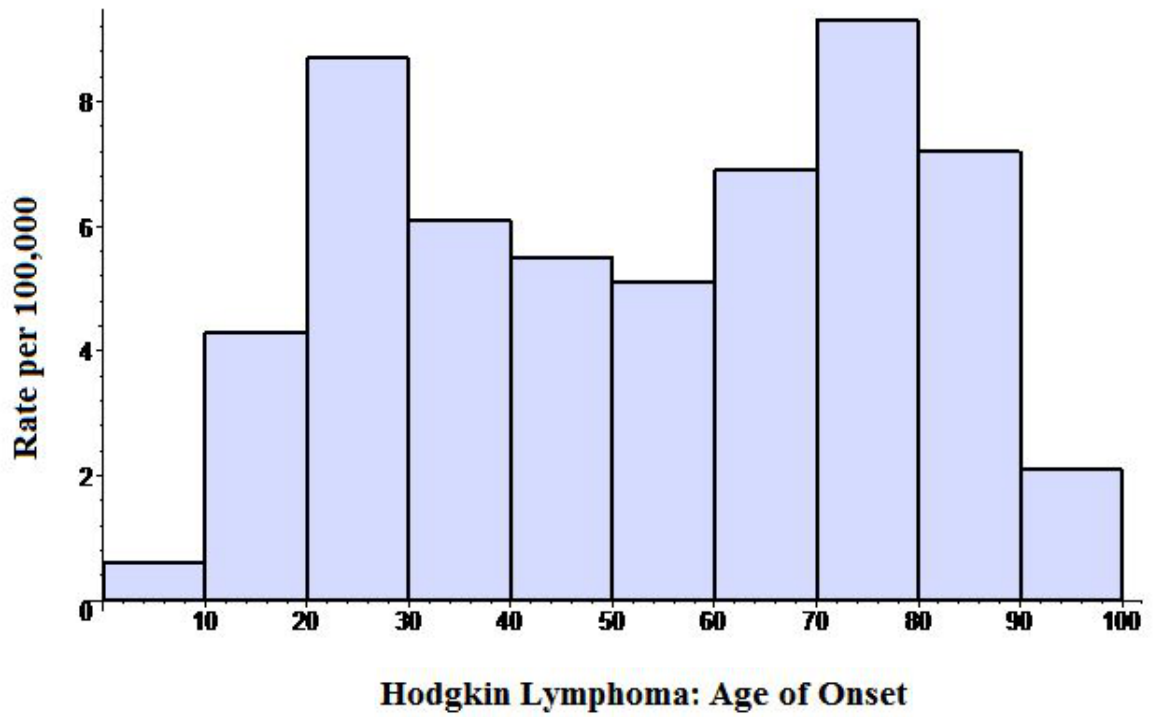
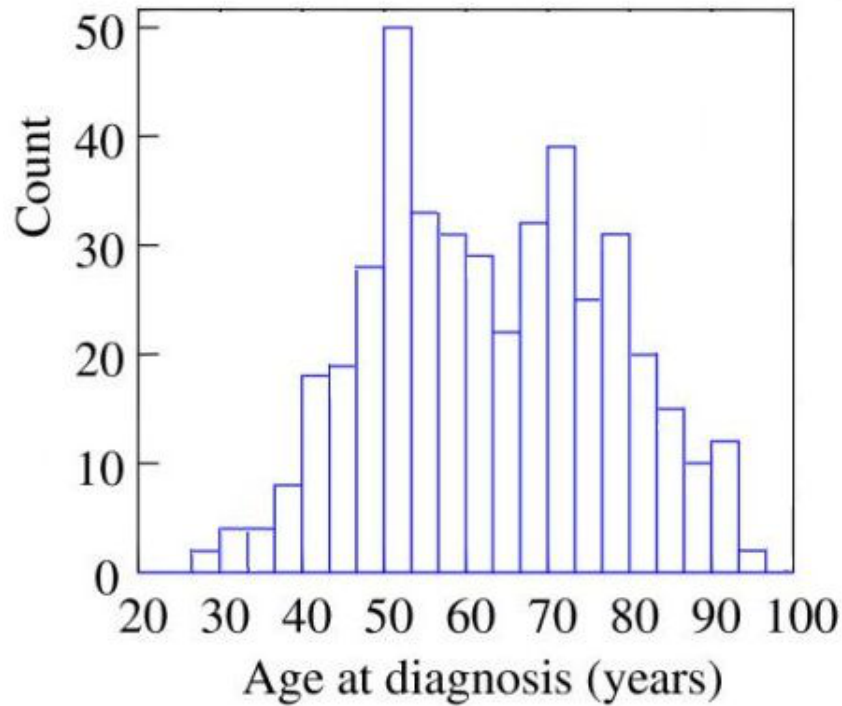


Figure 2.2.1: Age of Onset of Hodgkin Lymphoma: Rate per 100,000

Figure 2.2.2 below arose in an investigation during the years 1996–2002 of a possible link between the incidence of invasive breast cancer and the operation of a municipal solid waste incinerator in Besançon, France (Source: [http://openi.nlm.nih.gov/detailedresult.php?img=2267447\\_1476-072X-7-4-2&req=4](http://openi.nlm.nih.gov/detailedresult.php?img=2267447_1476-072X-7-4-2&req=4), Retrieved 09/02/2014)).



**Figure 2.2.2: Age at Diagnosis of Invasive Breast Cancer**

How many modes does the histogram in Figure 2.2.2 have? Nobody would disagree with the assertion that modes occur around the ages 50 and 70. What about the local maxima that occur in the late 70s and early 90s? If these are true modes, then the distribution is multimodal. But the maxima occurring there are not very pronounced. These maxima might even disappear if a different class width is chosen. Perhaps the distribution is really bimodal. The author leans to this judgment.

Figure 2.2.3 presents another situation that calls for careful consideration. It shows a histogram for a data set of 500 random two-digit integers.

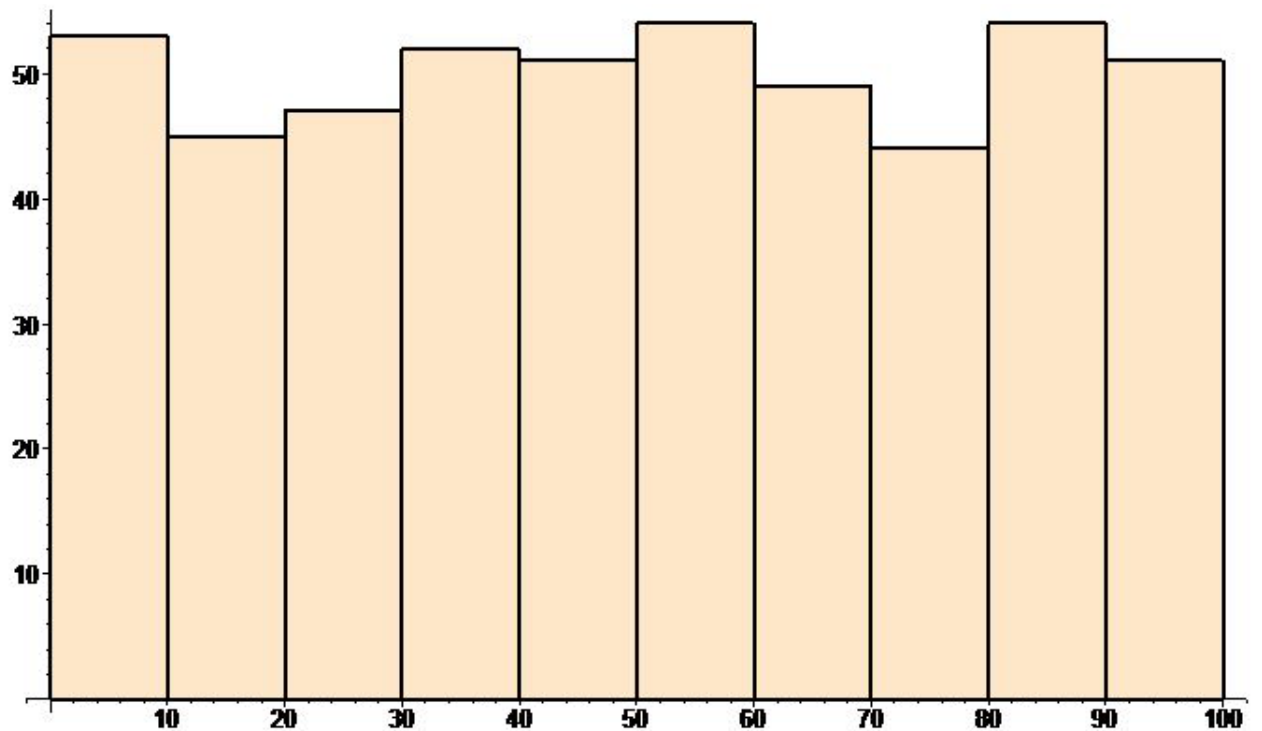


Figure 2.2.3: Histogram of 500 Random Two-Digit Numbers

Strictly speaking, four local maxima appear in the histogram. But these maxima do not stand out at all. They, along with the two local minima, seem to be insignificant peaks and valleys in what is an essentially flat distribution—in statistical jargon, a *uniform distribution*. We are justified in asserting that this distribution has no mode.

## Symmetry

Suppose that a histogram has a central bar, that the first bar to its left and the first bar to its right have equal heights, that the second bar to its left and the second bar to its right have equal heights, and so on. Then the histogram (and the distribution it represents) is said to be *symmetric*. If a vertical axis is drawn through the class mark of the central bar, and if the histogram is folded along the vertical axis, so that the left side is folded over onto the right, then it coincides exactly with the right side. Figure 2.2.4 shows a symmetric distribution of exam scores in a class of 110 students.



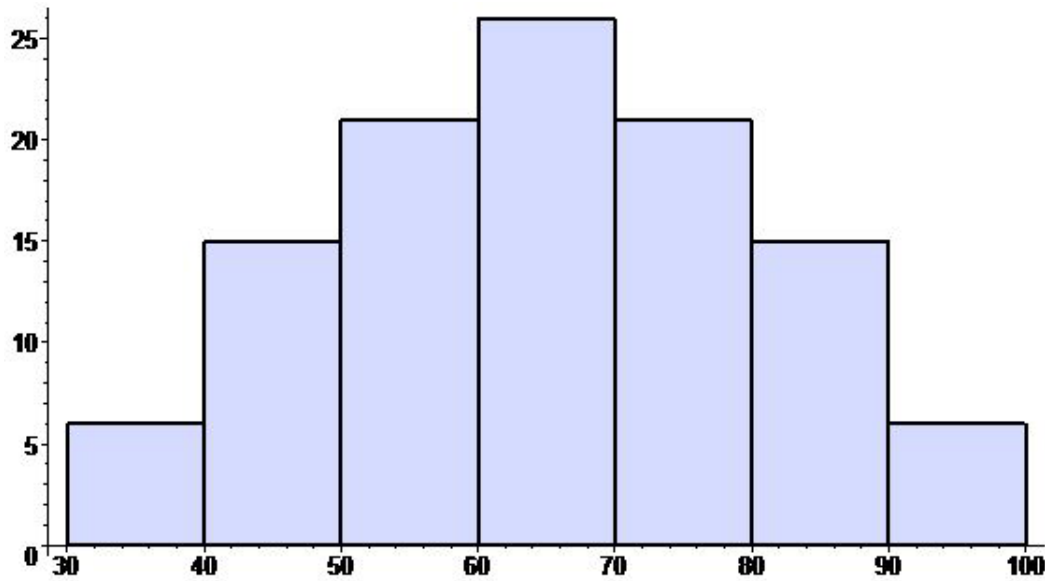


Figure 2.2.4: Histogram of Exam Scores (Bin width 10)

It should be noted that we must be a bit tolerant in deciding when a histogram is symmetric. Suppose that the histogram in Figure 2.2.4 is based on the following data:

Exam Score	[30,40)	[40,50)	[50,60)	[60,70)	[70,80)	[80,90)	[90,100)
Frequency	6	15	21	26	21	15	6

Table 2.2.1: Histogram of Exam Scores (Bin width 10)

Now suppose that we refine the data by reducing the class width from 10 to 5. In order to fit the data on the page, we will tabulate using the bin numbers from left to right, namely 1, 2, 3,  $\dots$  14, instead of the class intervals  $[30,35)$ ,  $[35,40)$ ,  $[40,45)$ ,  $\dots$   $[95,100)$ .

Bin	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Frequency	2	4	6	9	10	11	12	14	12	9	8	7	4	2

Table 2.2.2: Histogram of Exam Scores (Bin width 5)

The resulting histogram is shown in Figure 2.2.5.

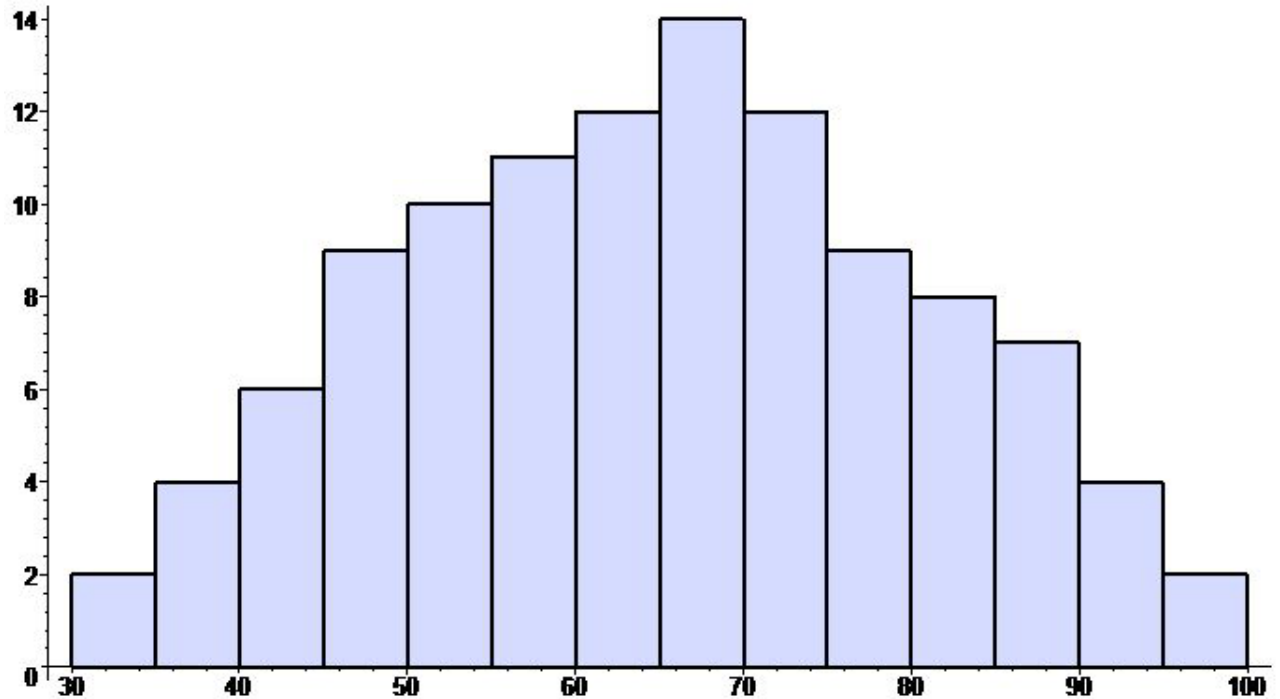


Figure 2.2.5: Histogram of Exam Scores (Bin width 5)

In contrast to the histogram in Figure 2.2.4, the symmetry of the histogram in Figure 2.2.5 is not perfect. To allow some real world imperfections, we use our judgment as to the meaning of “nearly equal” and relax the definition of *symmetry* as follows: If a histogram has a central bar, if the first bar to its left and the first bar to its right have nearly equal heights, if the second bar to its left and the second bar to its right have nearly equal heights, and so on., then the histogram (and the distribution it represents) is said to be *symmetric*. A histogram or distribution that is not symmetric is said to be *asymmetric*.

## Skewness

A *tail* of a distribution is the collection of bins, generally representing decreasing frequencies as the bins become farther from the center of the range, that extends to either side of a distribution. It is, one admits, a vaguely defined term, but, like the tail of a dog, be it long, short, bushy, or, curly, one tends to recognize one when one sees one. Figure 2.1.1 illustrates a distribution with both a left tail and a right tail. When a right tail extends further to the right than the left tail extends to the left, as in Figure 2.1.1, the distribution is said to be *skewed right* or *positively skewed* (because positive numbers stretch to the right on the number line). Similarly, if a left tail extends further to the left than the right tail extends to the right, then the distribution is said to be *skewed left* or *negatively skewed* (because negative numbers stretch to the left on the number line). Figure 2.2.6 shows a negatively skewed distribution.

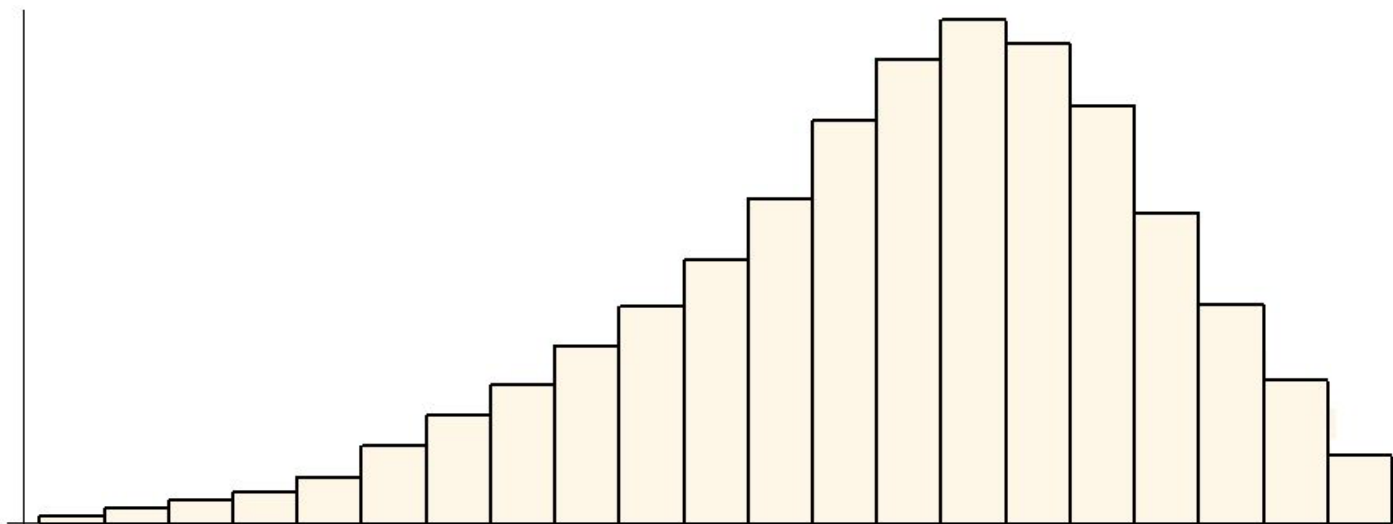


Figure 2.2.6: Histogram of a Negatively Skewed Distribution

## The Sample Mean

The *sample mean* of a data set  $x_1, x_2, x_3, \dots, x_N$ , which we are regarding as a sample drawn from a larger population, is denoted by  $\bar{x}$  and defined to be

$$\bar{x} = \frac{1}{N} (x_1 + x_2 + x_3 + \dots + x_N). \quad (2.2.1)$$

Outside of statistics, this number would more commonly be called the average value of the numbers in the data set. There is a shorthand notation for the sum  $x_1 + x_2 + x_3 + \dots + x_N$  that appears in the definition of the mean. The shorthand notation of which we speak, known as *sigma-notation* because it involves the Greek letter sigma ( $\Sigma$ ), can be used to express many sums. In sigma-notation, we denote the sum  $a_1 + a_2 + a_3 + \dots + a_N$  by  $\sum_{j=1}^N a_j$ . The “ $j$ ” in the notation is called the *index of summation*. Any other letter could have been used without changing the value of the sum. That is because we calculate  $\sum_{j=1}^N a_j$  by first replacing the  $j$  in  $a_j$  with the lower limit of summation, here specified to be 1, next replacing the  $j$  in  $a_j$  with 2, then with 3, and so on until we reach the upper limit of summation  $N$ . Finally, we add up all the expressions obtained by the replacements. Using sigma-notation, the formula for the mean is given by

$$\bar{x} = \frac{1}{N} \sum_{j=1}^N x_j. \quad (2.2.2)$$

## The Sample Variance

Well, that’s the news from Lake Wobegon<sup>7</sup>, where all the women are strong, all the men are good looking, and all the children are above average.

–Garrison Keillor

<sup>7</sup>See [http://en.wikipedia.org/wiki/Lake\\_Wobegon](http://en.wikipedia.org/wiki/Lake_Wobegon)

It may happen that all numbers in a data set  $x_1, x_2, x_3, \dots, x_N$  are equal. That is, it may happen that there is one number  $\xi$  such that  $x_1 = x_2 = x_3 = \dots = x_N = \xi$ . In this case, the mean  $\bar{x}$  is also equal to  $\xi$ . There is no variation or dispersion from the mean, and, by any standards, such a data set is especially uninteresting. In every other happenstance, there is at least one number in a data set that is above the mean, and at least one number that is below the mean. (It is an inviolable mathematical truth that numbers cannot all be above their average.) So, if a data set does not consist of one number repeated many times, then there is some variation from the mean in the set, and some of the variation comes from numbers above the mean and some of the variation comes from numbers below the mean. We desire a measure of the total dispersion of the numbers in a data set from their mean. We will call this measure, once we find a suitable formula for it, the *sample variance* of the data set. Now, one way to measure how far apart two numbers are is to calculate their difference. Thus, for each value  $x_j$  in the set we can calculate the difference  $x_j - \bar{x}$  and add up all these deviations from the mean:

$$\text{candidate sample variance} = (x_1 - \bar{x}) + (x_2 - \bar{x}) + \dots + (x_N - \bar{x}).$$

The problem with *candidate sample variance* is that some of its summands are positive and some are negative. When they are summed, there will be cancellations. The resulting sum will therefore be too small to accurately reflect the true, total variation.<sup>8</sup> For instance, if Ben gets 50 on each of two tests, and if Jerry gets 0 and 100, then the mean for both Ben and Jerry is 50. The candidate sample variance for Ben is  $(50 - 50) + (50 - 50)$ , or 0, and who can argue with that value for the variance? However, thanks to cancellation, the candidate sample variance for Jerry, namely  $(0 - 50) + (100 - 50)$ , is also 0,<sup>9</sup> and who would not believe that Jerry had the maximum possible variance!

The solution is to make all contributions to variance nonnegative so that there are no cancellations. Using the absolute value of each summand of *candidate variance* would do the job, but it would introduce a problem: the absolute value function is not differentiable at one point of its domain, and that failure would be enough to prevent statisticians from using the differential calculus to study *candidate sample variance*. Another way to make each summand of *candidate variance* nonnegative so that there are no cancellations is to square it. Squaring is a differentiable function, so the application of calculus is not disabled by using

$$\text{candidate sample variance} = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_N - \bar{x})^2.$$

There is still a problem with *candidate sample variance*: it can become a huge number just by taking a very large sample size. To understand what we mean by this, suppose that  $0.001 \leq |x_j - \bar{x}| \leq 0.002$  for every member  $x_j$  in our data set. Clearly, there is very little variation in this data set. But suppose that there are one trillion (1,000,000,000,000) values in our data set. Then,

$$\text{candidate sample variance} \geq \underbrace{(0.001)^2 + (0.001)^2 + \dots + (0.001)^2}_{\text{one trillion summands}} = 10^{12} \cdot 10^{-6} = 1,000,000.$$

We do not want variance to inflate just because we have a large sample size. A natural solution to this problem is to divide by the sample size:

$$\text{candidate sample variance} = \frac{1}{N} \left( (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_N - \bar{x})^2 \right).$$

There remains a problem with our latest *candidate sample variance*, but this problem is not so easily detected. To understand the essence of the remaining problem, let us return to the beryllium-11 example. We know the half-life of beryllium-11, that is, the time it takes the mass of a sample to be reduced to half the initial mass.

<sup>8</sup>In fact, the sum of all deviations from the mean is always 0, as will be shown in the optional subsection *Degrees of Freedom* that is found later in this section.

<sup>9</sup>As remarked in the preceding footnote, the sum of all deviations from the mean is always 0, so there is no surprise that the calculation for Jerry results in the same number as for Ben.

That knowledge together with the general law of radioactive decay determines a theoretical distribution for the time until decay of a beryllium-11 isotope. That theoretical distribution has a mean, which we will denote by  $\mu$ , and a variance, which we will denote by  $\sigma^2$ . Because the parameters  $\mu$  and  $\sigma^2$  are theoretical numbers that pertain to all beryllium-11 isotopes, and not only the isotopes that decayed during our study, we call  $\mu$  and  $\sigma^2$  the **population mean** and **population variance** respectively. We expect that, if the sample size is sufficiently large, then the sample mean  $\bar{x}$  will closely approximate the population mean  $\mu$ , and it does. We also want *candidate sample variance* to closely approximate the population variance  $\sigma^2$  when the sample size is large, and our latest stab at *candidate sample variance* fulfills our wish. However, it also tends to consistently underestimate the value it should approximate. And this tendency to underestimate exists generally—it is a defect that is *not* particular to the beryllium-11 decay distribution. In a more advanced statistics course, it is shown that this flaw of estimation can be avoided by replacing  $N$  with  $N - 1$  in the denominator of *candidate sample variance*. That is, we define the variance, which is denoted by the symbol  $s^2$ , as follows:

$$\text{sample variance} = s^2 = \frac{1}{N-1} \left( (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_N - \bar{x})^2 \right), \quad (2.2.3)$$

or, using sigma-notation,

$$s^2 = \frac{1}{N-1} \sum_{j=1}^N (x_j - \bar{x})^2. \quad (2.2.4)$$

## Standard Deviation

You will note that sample variance does not scale in an optimal way. Suppose, for example, that the data set  $x_1, x_2, x_3, \dots, x_N$  consists of distances measured in feet (ft). Then the units of variance are square feet (ft<sup>2</sup>). If we switch to inches, then the numerical values of the data are increased by a factor of 12, but the numerical value of the sample variance is increased by a factor of 12<sup>2</sup>, or 144. To obtain a quantity that reflects the spread of the data from the mean but which scales the same way as the data, we square-root the sample variance. The resulting quantity, which is called the **sample standard deviation**<sup>10</sup>, is denoted by  $s$  (what else?) and is defined by

$$\text{sample standard deviation} = s = \sqrt{\text{sample variance}} = \sqrt{\frac{1}{N-1} \sum_{j=1}^N (x_j - \bar{x})^2}. \quad (2.2.5)$$

## Degrees of Freedom of the Sample variance and Sample Standard Deviation

Earlier in this section we observed that the sums of the deviations for hypothetical test-takers Ben and Jerry were both zero. We now observe that this is true for any data set. Let  $\bar{x}$  denote the mean of  $x_1, x_2, \dots, x_N$ . If we multiply both sides of equation (2.2.2) by  $N$ , then we obtain

$$\sum_{j=1}^N x_j = N \bar{x}. \quad (2.2.6)$$

---

<sup>10</sup>Sir Francis Galton (1822–1911) introduced the notion of *standard deviation* in the early 1870s. However, he used the deviation from the *median* rather than the mean. The definition of “median” will be given in the next section.

We use this equation to calculate the sum of all (non-squared) deviations from the mean as follows:

$$\begin{aligned}
 \sum_{j=1}^N (x_j - \bar{x}) &= \left( \sum_{j=1}^N x_j \right) - \left( \sum_{j=1}^N \bar{x} \right) \\
 &= N\bar{x} - \sum_{j=1}^N \bar{x} \quad \text{using equation (2.2.6)} \\
 &= N\bar{x} - N\bar{x} \quad \text{adding } N \text{ copies of the constant } \bar{x} \\
 &= 0.
 \end{aligned}$$

One way to interpret this result is that the last deviation from the mean is determined by the values of all the preceding deviations:

$$\sum_{j=1}^{N-1} (x_j - \bar{x}) + (x_N - \bar{x}) = \sum_{j=1}^N (x_j - \bar{x}) = 0,$$

so

$$x_N - \bar{x} = - \sum_{j=1}^{N-1} (x_j - \bar{x}). \quad (2.2.7)$$

This equation for  $x_N - \bar{x}$  reveals an important distinction between the data set  $x_1, x_2, x_3, \dots, x_{N-1}, x_N$  and the deviations  $x_1 - \bar{x}, x_2 - \bar{x}, x_3 - \bar{x}, \dots, x_{N-1} - \bar{x}, x_N - \bar{x}$  from the mean. No value in the data set  $x_1, x_2, x_3, \dots, x_{N-1}, x_N$  is predetermined. There are  $N$  observations and so there are  $N$  “degrees of freedom.” By contrast, the set  $x_1 - \bar{x}, x_2 - \bar{x}, x_3 - \bar{x}, \dots, x_{N-1} - \bar{x}, x_N - \bar{x}$  of deviations from the mean has only  $N - 1$  degrees of freedom: once  $N - 1$  deviations from the mean are specified, the  $N^{\text{th}}$  is determined by equation (2.2.7). We say that the sample variance and the sample standard deviation have  $N - 1$  **degrees of freedom**. You will note that the formulas for sample variance and sample standard deviation both use this number in their denominators.

## Transformations—What They Change, What They Do Not Change

One common way to transform a data set is to add one fixed number to every member of the data set. The number that is added may be negative, in which case the transforming operation might be considered to be the subtraction of a positive number. For example, we could transform a set of temperature measurements in degrees centigrade to a set of temperature measurements in degrees Kelvin by adding  $-273.15$  to each centigrade measurement (or, equivalently, subtracting  $+273.15$  from each). Such a transformation is called a **translation**.

Another common way to transform a data set is to divide every member of the data set by a fixed nonzero number  $\lambda$ . Equivalently, we can think of the transforming operation as multiplication by  $1/\lambda$ . For example, if a set consists of changes of temperatures in degrees centigrade, we could, by either multiplying by  $5/9$  or dividing by  $9/5$ . Such a transformation is called a **scaling**.

Suppose  $c$  is a constant and that the translation  $T$  defined by

$$x \xrightarrow{T} T(x) = x + c$$

is applied to a data set  $X$  consisting of numbers  $x_1, x_2, \dots, x_N$  with mean  $\bar{x}$  and standard deviation  $s_X$ . Let  $y_j = T(x_j) = x_j + c$  for every  $j$ ,  $1 \leq j \leq N$ . Then the mean  $\bar{y}$  and standard deviation  $s_Y$  of the data set  $Y$  consisting of numbers  $y_1, y_2, \dots, y_N$  are given by

$$\bar{y} = \bar{x} + c \quad \text{and} \quad s_Y = s_X. \quad (2.2.8)$$

Proofs of the formulas in line (2.2.8) are routine. We have

$$\bar{y} = \frac{1}{N} \sum_{j=1}^N y_j = \frac{1}{N} \sum_{j=1}^N (x_j + c) = \frac{1}{N} \sum_{j=1}^N x_j + \frac{1}{N} \sum_{j=1}^N c = \bar{x} + \frac{1}{N} N \cdot c = \bar{x} + c.$$

The  $j^{\text{th}}$  deviation from the mean in the Y data set is given by  $y_j - \bar{y}$ , or  $(x_j + c) - (\bar{x} + c)$ , or  $x_j - \bar{x}$ , which is the  $j^{\text{th}}$  deviation from the mean in the X data set. Since the sets X and Y have the same deviations from their means, they have the same standard deviation.

Suppose  $\lambda$  is a nonzero constant and that the scaling  $S$  defined by

$$x \xrightarrow{S} S(x) = \frac{x}{\lambda}$$

is applied to a data set X consisting of numbers  $x_1, x_2, \dots, x_N$  with mean  $\bar{x}$  and standard deviation  $s_X$ . Let  $y_j = S(x_j) = x_j/\lambda$  for every  $j$ ,  $1 \leq j \leq N$ . Then the mean  $\bar{y}$  and standard deviation  $s_Y$  of the data set Y consisting of numbers  $y_1, y_2, \dots, y_N$  are given by

$$\bar{y} = \frac{\bar{x}}{\lambda} \quad \text{and} \quad s_Y = \frac{s_X}{\lambda}. \quad (2.2.9)$$

Proofs of the formulas in line (2.2.9) are routine. We have

$$\bar{y} = \frac{1}{N} \sum_{j=1}^N y_j = \frac{1}{N} \sum_{j=1}^N \frac{x_j}{\lambda} = \frac{1}{\lambda} \frac{1}{N} \sum_{j=1}^N x_j = \frac{1}{\lambda} \bar{x}.$$

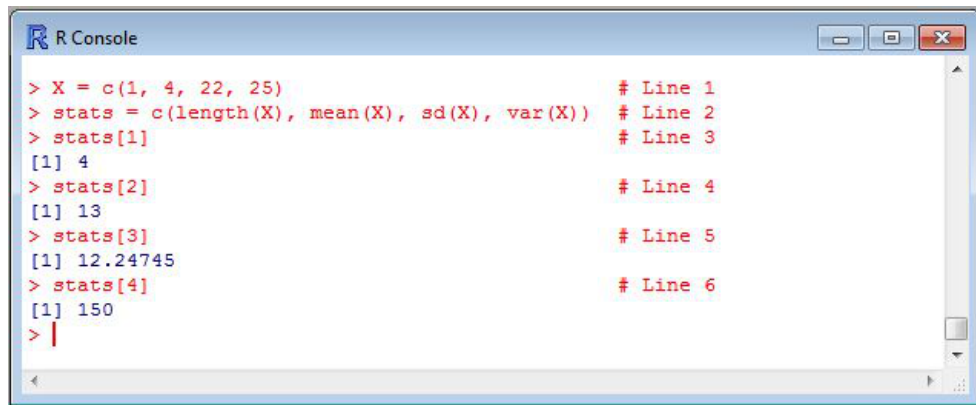
The calculation of the standard deviation  $s_Y$  will proceed via the calculation of the variance  $s_Y^2$  to avoid the inclusion of square roots:

$$\begin{aligned} s_Y^2 &= \frac{1}{N-1} \sum_{j=1}^N (y_j - \bar{y})^2 \\ &= \frac{1}{N-1} \sum_{j=1}^N \left( \frac{x_j}{\lambda} - \frac{\bar{x}}{\lambda} \right)^2 \\ &= \frac{1}{\lambda^2} \frac{1}{N-1} \sum_{j=1}^N (x_j - \bar{x})^2 \\ &= \frac{1}{\lambda^2} s_X^2. \end{aligned}$$

## Sample Mean, Sample Variance, Sample Standard Deviation in R (Optional)

The calculation of sample mean, sample variance, and sample standard deviation of a data set  $X$  with data values  $u, v, w, \dots$  in R could not be easier. We first convert the data set to a vector in R using the code `c(u,v,w,...)`. For convenience, we would ordinarily name the vector using R's assignment operation `leftName <- rightObject`, which assigns the object on the right to the name on the left. As we have mentioned, the sign of equality, namely `=`, can be used instead of `<-`, but this alternative seems to have less geek-appeal. Once `X <- c(u,v,w,...)` or `X = c(u,v,w,...)` has been entered, we use the commands `mean(X)`, `var(X)`, and `sd(X)` to calculate the sample mean, sample variance, and sample standard deviation of  $X$ . If  $X$  is a large data set, which is to say that the sample size  $N$  is large, then it is often convenient to use R to determine the sample size. The code `N = length(X)` counts the size of the data set and assigns it to `N`.

The following R session is a detailed illustration. Each line is commented with non-code that follows the sharp symbol `#`: in this example our comments are nothing more than line number references. The lines in blue are responses by R, where called for.



```

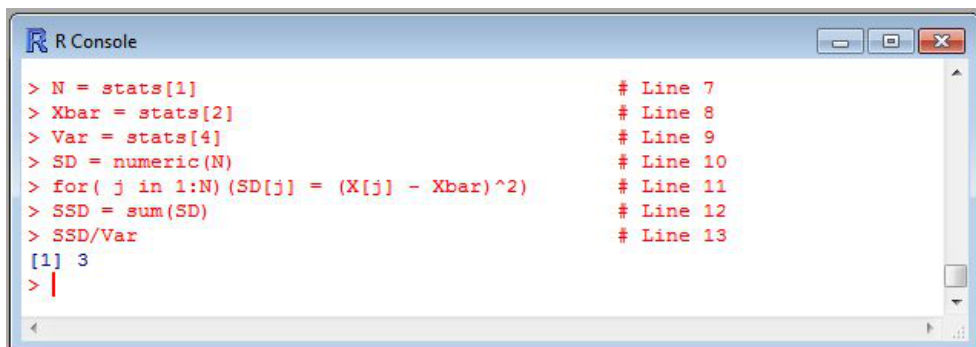
R Console
> X = c(1, 4, 22, 25) # Line 1
> stats = c(length(X), mean(X), sd(X), var(X)) # Line 2
> stats[1] # Line 3
[1] 4
> stats[2] # Line 4
[1] 13
> stats[3] # Line 5
[1] 12.24745
> stats[4] # Line 6
[1] 150
> |

```

Figure 2.2.7: Screen Capture of an R Session

In line 1, the data set 1, 4, 22, 25 is converted to a vector in R and assigned the name `X`. In Line 2, we create a vector, which we name `stats`, the first entry of which is the size of the sample, the second entry the mean, the third the standard deviation, and the fourth the variance. In general, `v[j]` picks out the  $j^{\text{th}}$  entry of vector `v`. Thus, Lines 3, 4, 5, and 6 cause R to respond with the size of the data set `X`, namely 4, the mean of `X`, namely 13, the standard deviation 12.2475, and the variance 150.

Figure 2.2.8 is a continuation of the R session shown in Figure 2.2.7. Its purpose is to show some additional functionality of R, to indicate looping capabilities, and to verify that the denominator R uses for both the variance and standard deviation of a sample of size  $N$  is  $N - 1$ .



```

R Console
> N = stats[1] # Line 7
> Xbar = stats[2] # Line 8
> Var = stats[4] # Line 9
> SD = numeric(N) # Line 10
> for(j in 1:N) (SD[j] = (X[j] - Xbar)^2) # Line 11
> SSD = sum(SD) # Line 12
> SSD/Var # Line 13
[1] 3
> |

```

Figure 2.2.8: Screen Capture of an R Session, Continued from Figure 2.2.7

Lines 7, 8, and 9 assign conveniently concise names for the sample size, sample mean, and sample variance. Line 10 creates an empty vector  $\langle \quad, \quad, \quad, \quad \rangle$  with  $N = 4$  entries. Line 10 also names this vector `SD`. Think of this vector as an empty egg carton that we will fill with four eggs. Well, numbers actually. Line 11 does that, one at a time, by making four assignments. The first is  $SD[1] = (X[1] - \bar{X})^2 = (1 - 13)^2 = 144$ , the second is  $SD[2] = (X[2] - \bar{X})^2 = (4 - 13)^2 = 81$ , the third is  $SD[3] = (X[3] - \bar{X})^2 = (22 - 13)^2 = 81$ , and the fourth is  $SD[4] = (X[4] - \bar{X})^2 = (25 - 13)^2 = 144$ . After this line, the vector `SD` is equal to  $\langle 144, 81, 81, 144 \rangle$ . Line 12 then sums these four vector entries and assigns the result, 450, to `SSD`, the Sum of the Square Deviations from the mean. Now, if `d` is the denominator R uses to calculate the variance, then it follows that  $\text{Var} = \text{SSD}/d$ , or  $d = \text{SSD}/\text{Var}$ . Line 13 calculates the ratio `SSD/Var` and returns the value 3, which is, indeed,  $N - 1$ .



## 2.3 Quartiles and the Five-Number Summary

There are five numbers associated with a data set  $x_1, x_2, x_3, \dots, x_N$  that provide us with a decent overview of the distribution of the data set. Taken together, they form the legendary **five-number summary** of the data set. If the size  $N$  of the data set is large, then obtaining a useful sense of the distribution with only five numbers amounts to very good compression ratio. Better yet, we have already encountered two of these numbers: the minimum value  $x_{\min}$  and the maximum value  $x_{\max}$ . Seems like smooth sailing: only three more numbers to go. They are denoted  $Q_1, Q_2, Q_3$ : the middle number,  $Q_2$ , competes with  $\bar{x}$  as a measurement of the center of the distribution, and  $Q_3 - Q_1$  competes with  $s_X$  as a measurement of the spread from the center.

### Percentiles

Suppose that  $0 \leq p \leq 100$ . Roughly speaking, the  $p^{\text{th}}$ -percentile of a data set  $x_1, x_2, x_3, \dots, x_N$  is a number that separates the lower  $p\%$  of values in the data set from the upper  $(100 - p)\%$ . Defining such a number precisely is less straightforward than one might think and there is no standard way to do so. Several different definitions are in fairly common use—one statistical program lets the user choose from five different formulas.

In order to define percentiles, it will be convenient to have a notation for the data set  $x_1, x_2, x_3, \dots, x_N$  when it is reordered from smallest value to largest value. To that end, we let  $x_{[1]}, x_{[2]}, x_{[3]}, \dots, x_{[N]}$  be a rearrangement of the data so that

$$x_{[1]} \leq x_{[2]} \leq x_{[3]} \leq \dots \leq x_{[N]}.$$

Notice that  $x_{[1]} = x_{\min}$  and  $x_{[N]} = x_{\max}$ .

**Example 1.** For the data set 29, 23, 26, 23, 28, 26, 27, 28, 34, 33, 20, 22, 29, 33 what are  $N$  and  $x_{[j]}$  for  $1 \leq j \leq N$ ?

**Solution.** The data set 29, 23, 26, 23, 28, 26, 27, 28, 34, 33, 20, 22, 29, 33 reordered from smallest to largest is 20, 22, 23, 23, 26, 26, 27, 28, 28, 29, 29, 33, 33, 34. Notice that, whichever order we use, the data set has 14 terms (even though some of the terms are repeated, with the result that fewer than 14 different numbers compose the set). Thus,  $N = 14$ . Reading the terms of the ordered list from left to right, we see that  $x_{[1]} = 20, x_{[2]} = 22, x_{[3]} = 23, x_{[4]} = 23, x_{[5]} = 26, x_{[6]} = 26, x_{[7]} = 27, x_{[8]} = 28, x_{[9]} = 28, x_{[10]} = 29, x_{[11]} = 29, x_{[12]} = 33, x_{[13]} = 33, \text{ and } x_{[14]} = 34$ .

We define the  $p^{\text{th}}$ -**percentile** to be the number  $x_{[m]}$  where  $m$  is the smallest index such that at least  $p\%$  of the values in the data set are less than or equal to  $x_{[m]}$ . This definition is commonly used, but there are other formulas that are also commonly used. In other words, there is *no* standard definition of *percentile*. Kind of surprising, eh? Notice that, in *this* definition, the  $p^{\text{th}}$ -percentile is *always* a member of the data set. In some of the other definitions out there, a percentile need not necessarily be a number in the data set.

There is actually an explicit formula for the index  $m$  of the  $p^{\text{th}}$ -percentile  $x_{[m]}$ :

$$m = \left\lceil \frac{p \cdot N}{100} \right\rceil.$$

The definition of the ceiling function that appears in this formula is given in Section 2.1. Basically it means, that if the number inside the upside down Ls is an integer, then leave it alone, but if it has a fractional part, then round *up* to the next larger integer.

**Example 2.** What is the 25<sup>th</sup>-percentile of the data set 29, 23, 26, 23, 28, 26, 27, 28, 34, 33, 20, 22, 29, 33 of the preceding example? What is the 50<sup>th</sup>-percentile?

**Solution.** In this example,  $N = 14$  and  $x_{[1]} = 20, x_{[2]} = 22, x_{[3]} = 23, x_{[4]} = 23, x_{[5]} = 26, x_{[6]} = 26, x_{[7]} = 27, x_{[8]} = 28, x_{[9]} = 28, x_{[10]} = 29, x_{[11]} = 29, x_{[12]} = 33, x_{[13]} = 33,$  and  $x_{[14]} = 34$ . For the 25<sup>th</sup>-percentile, we calculate

$$m = \left\lceil \frac{25 \cdot 14}{100} \right\rceil = \left\lceil \frac{7}{2} \right\rceil = \lceil 3.5 \rceil = 4.$$

Therefore, the 25<sup>th</sup>-percentile is  $x_{[4]}$ , or 23. For the 50<sup>th</sup>-percentile, we calculate

$$m = \left\lceil \frac{50 \cdot 14}{100} \right\rceil = \left\lceil \frac{14}{2} \right\rceil = \lceil 7 \rceil = 7.$$

Therefore, the 50<sup>th</sup>-percentile is  $x_{[7]}$ , or 27.

## Quartiles

For every data set (with at least 4 members), there are three numbers,  $Q_1, Q_2,$  and  $Q_3$  that divide the values of the data set into approximately equal fourths. (If the sample size  $N$  is not an integer multiple of 4, such as 4, 8, 12, 16, and so on, then exactly equal fourths is not possible.) Roughly speaking,  $Q_1$  is approximately the 25<sup>th</sup>-percentile,  $Q_2$  is approximately the 50<sup>th</sup>-percentile, and  $Q_3$  is approximately the 75<sup>th</sup>-percentile. Remember that percentiles are always members of the data set (in our definition), but we will not define the quartiles in a way that requires them to be members of the data set. As with percentiles, statistics does not provide a standard way of defining quartiles—there are several different definitions in common use.

We will define  $Q_2$  first, because the definitions of  $Q_1$  and  $Q_3$  depend on it. The name for  $Q_2$  is *middle quartile* or, synonymously and more commonly, *median*. For us, the terms *middle quartile* and *median* will *not* be synonymous with the term “50<sup>th</sup>-percentile”. According to the definition we have adopted, the 50<sup>th</sup>-percentile, like all other percentiles, is a member of the data set. In the definition of the median that we will soon describe, the median might be the average of two consecutive members of the ordered data set. If those two members of the data set are unequal, then the median will not be a member of the data set. If so, it cannot equal the 50<sup>th</sup>-percentile.

If a data set  $x_1, x_2, x_3, \dots, x_N$  has an odd number  $N$  of members, then the ordered rearrangement has a (unique) middle value, and we define  $Q_2$  to be that middle value. If  $N$  is not large, then this easiest done by inspection. For example, if  $N = 7$ , then at a glance we see that  $x_{[4]}$  is the middle value of the ordered list

$$x_{[1]} \quad x_{[2]} \quad x_{[3]} \qquad x_{[4]} \qquad x_{[5]} \quad x_{[6]} \quad x_{[7]}.$$

Inspection is not feasible when  $N$  is large, so it is useful to have a formula. Set  $k = \lfloor N/2 \rfloor$ . The floor symbol is used like the ceiling symbol that we have encountered, but, whereas we round up to calculate a ceiling, we round *down* to calculate the floor. Thus,  $\lfloor 3 \rfloor = 3$  and  $\lfloor 3.5 \rfloor = 3$ . It is not difficult to see that, with this definition of  $k$ , we have  $N = 2k + 1$  when  $N$  is an odd number. For example, if  $N = 7$ , then  $k = \lfloor 7/2 \rfloor = \lfloor 3.5 \rfloor = 3$ , and  $N = 2 \cdot 3 + 1 = 2 \cdot k + 1$ . The middle value of the data set is  $x_{[k+1]}$ .

If a data set  $x_1, x_2, x_3, \dots, x_N$  has an even number  $N$  of members, then the ordered rearrangement has *two* “middle values”. They are not exactly in the middle: one is the term on the left side closest to the middle and the other is the term on the right side closest to the middle. We define  $Q_2$  to be the average of the two middle values. If  $N$  is not large, then this easiest done by inspection. For example, if  $N = 8$ , then at a glance we see that  $x_{[4]}$ , the last term of the left half of the data, and  $x_{[5]}$ , the first term of the right half of the data, are the two middle values of the ordered list:

$$x_{[1]} \quad x_{[2]} \quad x_{[3]} \quad x_{[4]} \qquad x_{[5]} \quad x_{[6]} \quad x_{[7]} \quad x_{[8]}.$$

Inspection is not feasible when  $N$  is large, so it is useful to have a formula. As in the previous case, set  $k = \lfloor N/2 \rfloor$ . In this case  $N/2$  is an integer, so the floor symbol does not change anything—there is no rounding

to be done—and can be disregarded. For  $N = 8$ , we have  $k = \lfloor 8/2 \rfloor = \lfloor 4 \rfloor = 4$ , and  $N = 8 = 2 \cdot 4 = 2k$ . It is easy to see that  $N = 2k$  is always true when  $N$  is even. The two middle values are  $x_{[k]}$  and  $x_{[k+1]}$ . As a result,  $Q_2 = (x_{[k]} + x_{[k+1]})/2$ .

The number  $Q_1$ , which, roughly speaking, separates the lowest 25% of the data from the highest 75% of the data, is called the **lower quartile**. The number  $Q_3$ , which, roughly speaking, separates the lowest 75% of the data from the highest 25% of the data, is called the **upper quartile**. If the sample size  $N$  is even, i.e., if  $N = 2k$ , then the ordered data list splits into two halves, each with  $k$  members. We define  $Q_1$  to be the median of the left half and  $Q_3$  to be the median of the right half.

If the sample size  $N$  is odd, i.e., if  $N = 2k + 1$ , then the ordered data list contains  $k$  members to the left of the middle value  $x_{[k+1]}$  and  $k$  members to the right. The question becomes, What do we do with the middle value itself? Apparently, one strategy is to discard it. That strategy does not fit in the author's game plan. The middle value is a genuine data value, not an average. The author retains data values the way hoarders hoard hoarded things. Our strategy is to include the middle value in both halves. Thus, we define  $Q_1$  to be the median of the left half

$$x_{[1]} \quad x_{[2]} \quad x_{[3]} \quad \cdots \quad x_{[k+1]}$$

and  $Q_3$  to be the median of the right half.

$$x_{[k+1]} \quad x_{[k+2]} \quad x_{[k+3]} \quad \cdots \quad x_{[2k+1]}$$

The terms for the four subsets of the data set that are separated by the quartiles are **bottom quartile**, **lower middle quartile**, **upper middle quartile**, and **top quartile**. The difference  $Q_3 - Q_1$  is called the **interquartile range**, which is abbreviated as *IRQ*. The nondecreasing sequence  $x_{\min}, Q_1, Q_2, Q_3, x_{\max}$  is called the **five number summary**.

## Four Prototype Quartile Calculations

According to the definition of the median that we have adopted, the median is the middle value of an ordered data set with an odd number of terms or the average of the two middle values if the ordered data set has an even number of terms. Because the upper and lower quartiles are defined to be medians of a subset of the full data set, they too are either middle values or averages. Thus, every one of the three quartiles is a middle value, which we will abbreviate by mv, or an average, which we will abbreviate by avg. Because of the symmetric definitions of  $Q_1$  and  $Q_3$ , we see that  $Q_1$  and  $Q_3$  are both middle values or both averages. In the four examples of this subsection, we will see that the way the three quartiles are obtained is determined by the remainder that results when the population size is divided by 4.

It is a simple fact of arithmetic that when we divide an integer  $N$  by 4, we obtain a remainder equal to 0, 1, 2, or 3. Using the values 8, 9, 10, 11, 12, 13, 14, and 15 to illustrate, we have  $8 = 2 \cdot 4 + 0$ ,  $9 = 2 \cdot 4 + 1$ ,  $10 = 2 \cdot 4 + 2$ ,  $11 = 2 \cdot 4 + 3$ ,  $12 = 3 \cdot 4 + 0$ ,  $13 = 3 \cdot 4 + 1$ ,  $14 = 3 \cdot 4 + 2$ , and  $15 = 3 \cdot 4 + 3$ . We will use the values  $N = 8$ ,  $N = 9$ ,  $N = 10$ , and  $N = 11$  for our examples. The behavior that will be seen for these values of  $N$  are representative of the behaviors that occur for all values of  $N$ .

**Example 3.** Calculate the quartiles for the presorted data set 13, 17, 18, 20, 24, 28, 30, 31.

**Solution.** In this example,  $N = 8 = 4 \cdot 2 + 0 = 4 \cdot \ell + 0$  with  $\ell = 2$ . Because  $N$  is even, the ordered data splits into two equal-sized halves, 13, 17, 18, 20 and 24, 28, 30, 31, with the median  $Q_2$  being the average of the values in each half closest to the middle:  $Q_2 = (20 + 24)/2 = 22$ . The lower quartile  $Q_1$  is the median of the numbers in the lower half, namely the average of 17 and 18, or 17.5. The upper quartile  $Q_3$  is the median of the numbers in the upper half, namely the average of 28 and 30, or 29. For the purposes of filling in the table that follows these examples, we remark that all three quartiles were obtained by averaging:  $Q_1$ : avg,  $Q_2$ : avg,  $Q_3$ : avg.

**Example 4.** Calculate the quartiles for the presorted data set 13, 17, 18, 20, 24, 28, 30, 31, 35.

**Solution.** In this example,  $N = 9 = 4 \cdot 2 + 1 = 4 \cdot \ell + 1$  with  $\ell = 2$ . Because  $N$  is odd, the ordered data splits into two equal-sized halves separated by a middle value:

$$13, 17, 18, 20, \quad 24, \quad 28, 30, 31, 35.$$

The median  $Q_2$  is the middle value, 24. To calculate  $Q_1$  and  $Q_3$  when there is a middle value, another copy of the middle value is included in the data set so that the augmented set has an even number of terms and splits into equal-sized halves:

$$13, 17, 18, 20, 24, \quad 24, 28, 30, 31, 35.$$

The lower quartile  $Q_1$  is the median of the numbers in the lower half, namely the middle value 18. The upper quartile  $Q_3$  is the median of the numbers in the upper half, namely the middle value 30. For the purposes of filling in the table that follows these examples,, we remark that all three quartiles were obtained by identifying middle values:  $Q_1$ : mv,  $Q_2$ : mv,  $Q_3$ : mv.

**Example 5.** Calculate the quartiles for the presorted data set 9, 13, 17, 18, 20, 24, 28, 30, 31, 35.

**Solution.** In this example,  $N = 10 = 4 \cdot 2 + 2 = 4 \cdot \ell + 2$  with  $\ell = 2$ . Because  $N$  is even, the ordered data splits into two equal-sized halves, 9, 13, 17, 18, 20 and 24, 28, 30, 31, 35, with the median  $Q_2$  being the average of the values in each half closest to the middle:  $Q_2 = (20 + 24)/2 = 22$ . The lower quartile  $Q_1$  is the median of the numbers 9, 13, 17, 18, 20 in the lower half, namely the middle value 17. The upper quartile  $Q_3$  is the median of the numbers 24, 28, 30, 31, 35 in the upper half, namely the middle value 30. For the purposes of filling in the table that follows these examples, we remark that the median was obtained by averaging, whereas the upper and lower quartiles were obtained by identifying middle values:  $Q_1$ : mv,  $Q_2$ : avg,  $Q_3$ : mv.

**Example 6.** Calculate the quartiles for the presorted data set 9, 13, 17, 18, 20, 24, 28, 30, 31, 35, 36.

**Solution.** In this example,  $N = 11 = 4 \cdot 2 + 3 = 4 \cdot \ell + 3$  with  $\ell = 2$ . Because  $N$  is odd, the ordered data splits into two equal-sized halves separated by a middle value:

$$9, 13, 17, 18, 20, \quad 24, \quad 28, 30, 31, 35, 36.$$

The median  $Q_2$  is the middle value, 24. To calculate  $Q_1$  and  $Q_3$  when there is a middle value, another copy of the middle value is included in the data set so that the augmented set has an even number of terms and splits into equal-sized halves:

$$9, 13, 17, 18, 20, 24 \quad 24, 28, 30, 31, 35, 36.$$

The lower quartile  $Q_1$  is the median of the numbers in the lower half, which has an even number of terms. Therefore,  $Q_1$  is the average  $(17+18)/2$  of the two middle values, or 17.5 The upper quartile  $Q_3$  is the median of the numbers in the upper half, the average  $(30+31)/2$  of the two middle values, or 30.5 For the purposes of filling in the table that follows, we remark that the upper and lower quartiles were obtained by averaging, whereas the median was obtained by identifying a middle value:  $Q_1$ : avg,  $Q_2$ : mv,  $Q_3$ : avg.

In the table that follows, we summarize our observations.

$N$	$Q_1$	$Q_2$	$Q_3$
$4 \cdot \ell + 0$	avg	avg	avg
$4 \cdot \ell + 1$	mv	mv	mv
$4 \cdot \ell + 2$	mv	avg	mv
$4 \cdot \ell + 3$	avg	mv	avg

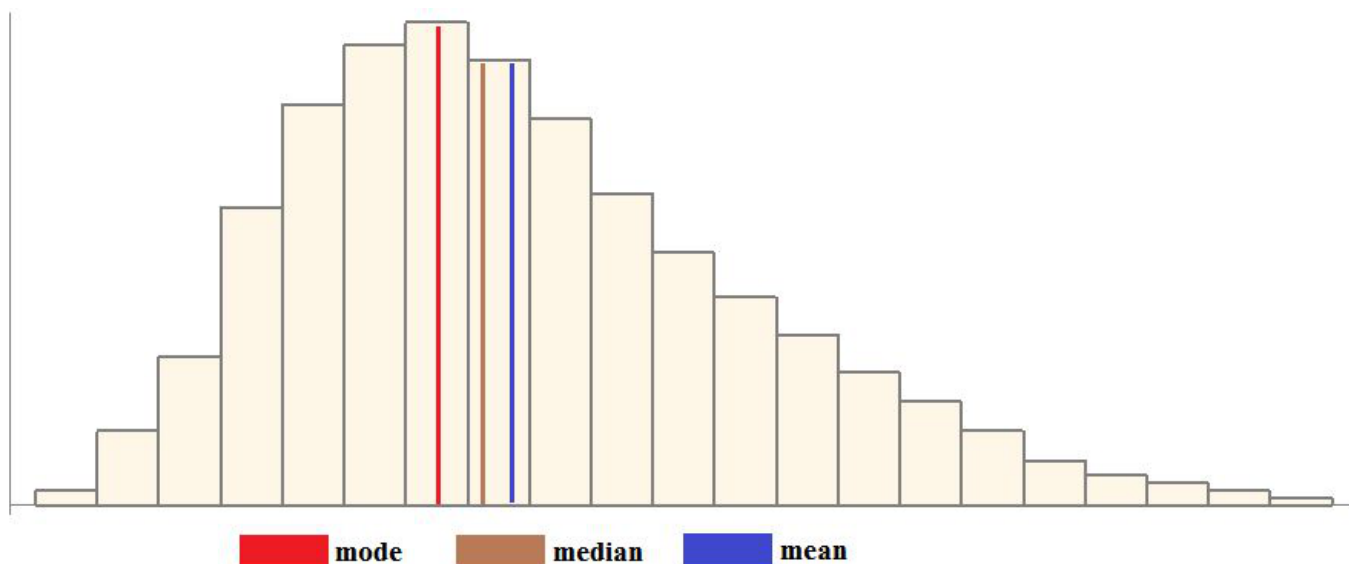
**Table 2.3.1: Quartile Calculation and the Remainder of  $N$  Upon Division by 4**

## Skewness, the Mean and the Median

In a skewed distribution, the mean is farther out in the long tail than is the median.

–David S. Moore, George P. McCabe<sup>11</sup>

In an *archetypal* unimodal distribution that is skewed left (as shown in Figure 2.2.6), the mean, median, and mode occur from left to right in alphabetical order. In an *archetypal* unimodal distribution that is skewed right, the mode, median, and mean occur from left to right in reverse alphabetical order. Figure 2.3.1 illustrates the second of these assertions, sometimes cited as the Mean-Median-Mode Inequality.



**Figure 2.3.1: Histogram of a Positively Skewed Unimodal Distribution**

Let us understand the basis of the Mean-Median-Mode Inequality, using a positively skewed unimodal distribution (as in Figure 2.3.1) for our discussion. The location of the (roughly central) mode speaks for itself. Because it is longer, we expect the right tail to have a greater effect on the mean and on the median than the left tail. The numbers to the right on the horizontal axis are the large values of the data set. The contribution of these numbers (as summands) to the numerator of the mean, tends to make the mean larger and, geometrically speaking, push it to the right. These large values that compose the right tail tend to push the median to the right, but to a lesser extent. That is because the median is about frequencies, frequencies are measured by the vertical axis, and the bars for the large values are short. Thus, the large data values that prominently effect the mean occur infrequently and have little effect on the median.

It should be emphasized that the Mean-Median-Mode Inequality is merely a rule of thumb. As in the quote with which this subsection opened, this rule of thumb is often stated without any qualifications as if it were a universal truth. It is not. Expect the Mean-Median-Mode Inequality to hold for a distribution that has a histogram similar to those of Figures 2.2.6 and 2.3.1. Otherwise, do not take it for granted that the Mean-Median-Mode Inequality holds.

<sup>11</sup>*Introduction to the Practice of Statistics*, Third Edition, David S. Moore, George P. McCabe, W.H. freeman and Co., p.43

## Outliers

Roughly speaking, an *outlier* is a data value that is distant *from the main group* of data values. So say we. The handbook of the National Institute of Standards and Technology says, “An *outlier* is an observation that lies an abnormal distance from other values in a random sample from a population.” Wikipedia says pretty much the same thing: “An *outlier* is an observation point that is distant from other observations.”

Clearly these definitions are somewhat vague. In the NIST definition, how do we decide what is “abnormal”? The NIST handbook acknowledges that its definition “leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal.” In the Wikipedia definition, how distant from other points must an observation be in order to be deemed “distant”? Wikipedia acknowledges, “Determining whether or not an observation is an outlier is ultimately a subjective exercise.”

Even worse, the NIST and Wikipedia definitions allow logical complications. Suppose we have a data set with 1000 observations, of which 990 are between 0 and 50 and 10 are between 77 and 78. The observations in this small cluster do not satisfy either the NIST or the Wikipedia definition of *outlier* (because each one is close to nine other data values). Nevertheless, most statisticians would consider them to be outliers.

There is no universally accepted mathematical criterion for categorizing an isolated data value as an outlier. Several have been proposed (and we will describe one of them later in this section), but, ultimately, classifying a data value as an outlier is a judgment call. Outliers should always be investigated—frequently outliers are due to error and it is always beneficial to root out errors. Standard scientific practice is to report outliers. If an outlier is to be disregarded from an analysis, then that decision should be explained.<sup>12</sup>

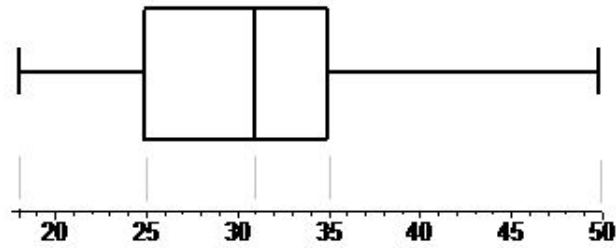
## Box and Whisker Plot

A *box and whisker plot* or *boxplot* uses a box and two whiskers to display some of the salient features of a numerical distribution. We will illustrate this graphing procedure with the EPA mpg data for the 2013 Toyota fleet, which we considered in Section 2.1: 18, 19, 22, 22, 23, 25, 25, 25, 26, 28, 31, 31, 32, 32, 33, 35, 40, 40, 41, 42, 50. The five-number summary for this set is  $x_{\min} = 18$ ,  $Q_1 = 25$ ,  $Q_2 = 31$ ,  $Q_3 = 35$ ,  $x_{\max} = 50$ . We draw a horizontal axis that spans the data set. Next, we draw equal length vertical segments at each of the three quartiles. The outer segments, i.e., the segments at  $Q_1$  and  $Q_3$ , form the vertical sides of a box. How tall should the vertical segments be? That is an aesthetic judgment. A box that looks like a shoe box aligned heel-to-toe horizontally is common.

Nearly one third of the way done! With the box out of the way, we start growing whiskers. For the whiskers, draw two vertical “fences”, one 1.5 IQRs to the left of the left side of the box, and one 1.5 IQRs to the right of the right side of the box. Draw these lightly so they can be erased—they are merely guidelines and not part of the finished boxplot. In our example,  $\text{IQR} = 35 - 25 = 10$ , so the left fence is at  $x = 25 - 1.5 \cdot 10 = 10$  and the right fence is at  $x = 35 + 1.5 \cdot 10 = 50$ . Next, extend a left whisker (a horizontal line segment) that starts half-way up the left side of the box and stretches horizontally to the data value that, among all data values to the left of the box, is the one farthest from the box but still within the fenced region. In our example, the horizontal line extends leftward from  $x = 25$  to  $x = 18$ . Draw the right whisker analogously. In our example, the right whisker extends horizontally to the right from  $x = 35$  to  $x = 50$ . (We remark that had the value 50 been 51, it would have been outside the fence. In that case the whisker would have extended only to  $x = 42$ , which would have been the greatest data value inside the fence.) Short vertical lines are usually added to the ends of the whiskers. See Figure 2.3.2.

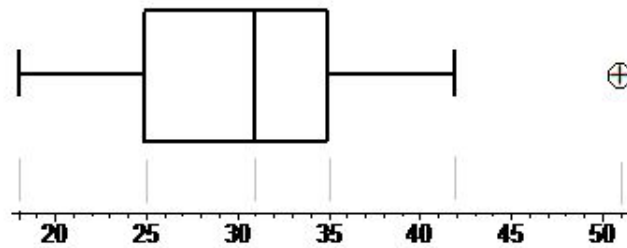
---

<sup>12</sup>In the older literature, outliers that were questionable observations were routinely discarded without comment. Some statisticians nowadays continue to discard without comment outliers that can be traced to error. Other statisticians maintain that all outliers must be reported.



**Figure 2.3.2: Box and Whisker Plot of Toyota EPA Combined Mileage, 2013 Fleet**

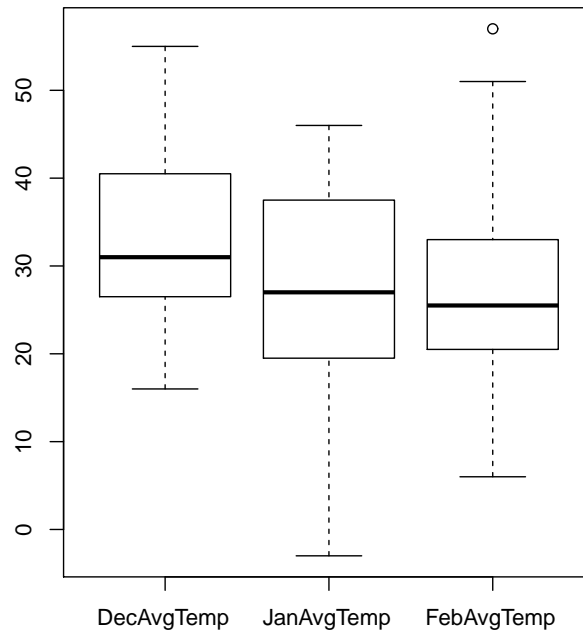
For the last step, in creating a boxplot, outliers are plotted by using symbols such as a small circle or an  $\oplus$ . Different symbols may be used in the same boxplot (without signifying any particular difference between the data values they represent). There are no outliers in our Toyota mileage example, but if the data value 50 had been 51, then it would have been outside the fence, and arguably far enough away from the nearest data value 42 to be considered an outlier. The boxplot of this hypothetical distribution appears in Figure 2.3.3.



**Figure 2.3.3: Box and Whisker Plot of Hypothetical Toyota EPA Combined Mileage, 2013 Fleet**

There are a number of variations on the basic boxplot theme. The boxplot we have described is called the *Tukey boxplot*. We have drawn our boxplots horizontally only because the graphics in these notes fit better with landscape orientation rather than portrait orientation. Vertically oriented boxplots work just as well and are extremely common.

Boxplots are particularly effective when a variable is tabulated for every value of a group, a boxplot is created for every value of the group, and the boxplots are placed side-by-side for comparison. For example, we might want to compare EPA combined mileage for the fleets of Toyota, Ford, and General Motors. Or we might want to compare weather features by month. During the winter of 2013-2014, the winter of the polar vortex, which month was coldest: December, January, or February? Boxplots for these months can be inspected in Figure 2.3.4.



**Figure 2.3.4:** Boxplots for the Average Daily Temperatures in St. Louis for Dec 2013, Jan 2014, Feb 2014

### Outliers on the Other Sides of the Fences

In the previous subsection, we described fences drawn at the values  $Q_1 - 1.5 \times \text{IQR}$  and  $Q_3 + 1.5 \times \text{IQR}$ . These are sometimes called the *lower inner fence* and the *upper inner fence* respectively. Fences drawn at  $Q_1 - 3 \times \text{IQR}$  and  $Q_3 + 3 \times \text{IQR}$  are called the *lower outer fence* and *upper outer fence* respectively.

As our discussion of boxplots suggests, observations that are *not* within the inner fences are *suspected* outliers. In fact, if such an observation is within the outer fences, it is, according to the NIST handbook, considered to be a *mild outlier*. If an observation is not within the outer fences, then it is, according to the NIST handbook, considered to be an *extreme outlier*. It need hardly be mentioned that these definitions are not universally adopted.

### Five-Number Summaries and Boxplots in R (Optional)

To calculate the five-number summary of a data set  $u, v, w, \dots$  in R, assign the data set to a name such as `X` by means of the code `X <- c(u,v,w,...)`, and then call on the `fivenum` command with argument `X`: `fivenum(X)`. For this command, R uses the same definitions for the three quartiles  $Q_1, Q_2, Q_3$  as are found in these notes, so there will be no surprises. The command `summary(X)` can also be used, but the method of calculating  $Q_1$  and  $Q_3$  is *not* the one described in these notes. For example, if `X` is the data set 1,2,3,4,7,9, then the command `summary(X)` prompts this output:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.000	2.250	3.500	4.333	6.250	9.000



There is no need to discuss the method of calculation that `summary` uses for  $Q_1$  and  $Q_3$ : we have already made our choice of method and we will stick to it.

In R, a boxplot for a data set `X <- c(u,v,w, ...)` can be obtained by simply entering `boxplot(X)`, but this call does not result in the Tukey boxplot: the whiskers extend to the smallest and largest values of the data set `X`. For a Tukey boxplot, an additional parameter must be specified: `boxplot(X, range = 1.5)`. The default orientation for a boxplot in R is vertical. If you desire a horizontal boxplot, then include a parameter to overrule the default: `boxplot(X, range = 1.5, horizontal=TRUE)`. Because boxplots are so useful for comparing the values of a variable when they are split into groups, it might be worthwhile to show the R code that we used to create Figure 2.3.4. In order to fit the lines on the page, we suppressed many of the daily averages and replaced them with ellipses.

```
> DecAvgTemp <- c(46,49,53,55,31,20, ...,39,42,49,31,21,34)
> JanAvgTemp <- c(36,16,17,37,19,-3, ...,43,16,12,24,36,34)
> FebAvgTemp <- c(33,21,19,25,17,6, ...,20,24,34,25.5,25.5,25.5)
> data <- data.frame(DecAvgTemp,JanAvgTemp,FebAvgTemp)
> boxplot(data, range=1.5)
```

## 2.4 Chebyshev's Inequality

In this section we will assume that every data set `X` under consideration contains at least two distinct observations. In other words, the standard deviation  $s_X$  of `X` satisfies  $s_X > 0$ , and so we may fearlessly divide by  $s_X$ .

If we know the size  $N$ , the mean  $\bar{x}$ , and the standard deviation  $s_X$  of a data set `X`, then we can infer some simple facts about values of `X`. For example, let

$$\rho = \sqrt{\frac{N-1}{N}} \times s_X.$$

Then there must be at least one observation that does not lie inside the open interval  $(\bar{x} - \rho, \bar{x} + \rho)$ . Otherwise, all the absolute deviations  $|x_j - \bar{x}|$  from the mean would satisfy  $|x_j - \bar{x}| < \rho$ , or  $(x_j - \bar{x})^2 < \rho^2$ , or

$$(x_j - \bar{x})^2 < \frac{N-1}{N} \times s_X^2,$$

and, as a result,

$$s_X^2 = \frac{1}{N-1} \sum_{j=1}^N (x_j - \bar{x})^2 < \frac{1}{N-1} \sum_{j=1}^N \frac{N-1}{N} \times s_X^2 = \frac{1}{N-1} N \frac{N-1}{N} \times s_X^2 = s_X^2,$$

which is impossible. For example, in a data set of size 10 with mean 0 and standard deviation 1, at least one observation must have absolute value greater than or equal to  $\sqrt{9/10} = 0.948683\dots$

### Chebyshev's Upper Bound for the Number of Peripheral Observations

Suppose that `X = x1, x2, ..., xN` is a data set with mean  $\bar{x}$  and standard deviation  $s_X$  (as given by formulas (2.2.2) and (2.2.5)). Let  $\lambda$  be a fixed positive number—other than requiring that  $\lambda > 0$ , there is no restriction in what  $\lambda$  might be: it might be less than 1 but it might be greater than 1, it might be a positive integer, but it might not be an integer, it might be a rational number like  $3/2$ , but it might be an irrational number like  $\sqrt{2}$ . We will use  $\lambda$  as a measure of whether a data value is located *centrally* or *peripherally*. To be specific, we will say that  $x_j$  is centrally located if it is less than  $\lambda$  standard deviations away from the mean. That is,

$x_j$  is centrally located if  $|x - \bar{x}| < \lambda \cdot s_X$ . On the other hand, if  $x_j$  is at least  $\lambda$  standard deviations away from the mean, which is to say that  $|x - \bar{x}| \geq \lambda \cdot s_X$ , then we say that  $x_j$  is peripherally located.

Let  $\mathcal{I} = \{1, 2, 3, \dots, N\}$ . This is the set of all indices of the data set  $X$ . Let  $\mathcal{I}_c(\lambda)$  be the subset of  $\mathcal{I}$  consisting of all those  $j$  in  $\mathcal{I}$  for which  $x_j$  is centrally located. Let  $\mathcal{I}_p(\lambda)$  be the subset of  $\mathcal{I}$  consisting of all those  $j$  in  $\mathcal{I}$  for which  $x_j$  is peripherally located. Let  $N_c(\lambda)$ , respectively  $N_p(\lambda)$ , be the number of centrally, respectively peripherally, located data values. Because a data value is either central or peripheral, it is clear that  $N = N_c(\lambda) + N_p(\lambda)$ . Additional reflection suggests that the value of  $s$  must place another constraint on the value of  $N_p(\lambda)$ . For each peripheral datum  $x_j$ , the absolute deviation  $|x_j - \bar{x}|$  is no less than  $\lambda \cdot s_X$ , and so the square deviation is no less than  $\lambda^2 \cdot s_X^2$ . If we sum only the square deviations from peripheral data, then we obtain at least  $N_p(\lambda) \cdot \lambda^2 \cdot s_X^2$ . Thus,

$$N_p(\lambda) \cdot \lambda^2 \cdot s_X^2 \leq \sum_{j \in \mathcal{I}_p(\lambda)} (x_j - \bar{x})^2 \leq \sum_{j=1}^N (x_j - \bar{x})^2 = (N - 1) s_X^2 < N s_X^2.$$

If we divide both sides of this inequality by  $N \cdot \lambda^2 \cdot s_X^2$ , then we obtain

$$\frac{N_p(\lambda)}{N} < \frac{1}{\lambda^2}. \quad (2.4.1)$$

Inequality (2.4.1) is one member of a family of similar inequalities that Pafnuty Lvovich Chebyshev, a Russian mathematician of many accomplishments, derived in 1867. In words, Chebyshev's Inequality states,

If  $\lambda$  is any positive number and  $X$  is any data set, then the fraction of points in  $X$  that lie at least  $\lambda$  standard deviations from the mean of  $X$  is less than  $1/\lambda^2$ .

It will be noted that, although we assumed only that  $\lambda > 0$ , the left size of inequality (2.4.1) can obviously not exceed 1. Therefore, inequality (2.4.1) is only of interest for  $\lambda$  satisfying  $\lambda \geq 1$ .

**Example 1.** Suppose that a data set  $X$  of 16 observations has mean  $\bar{x} = 36.4375$  and standard deviation  $s_X = 15.57335$ . How many of the observations can be at least 1.25 standard deviations from the mean?

**Solution.** Let  $N_p(1.25)$  be the number of observations that are at least 1.25 standard deviations from the mean. Then inequality (2.4.1) tells us that  $N_p(1.25) < 16/(1.25)^2$ , or  $N_p(1.25) < 10.24$ . Because  $N_p(1.25)$  is an integer, we infer that  $N_p(1.25) \leq 10$ : there are ten or fewer observations that are at least 1.25 standard deviations from the mean. Calculating  $\bar{x} - 1.25 \times s_X = 16.9708$  and  $\bar{x} + 1.25 \times s_X = 55.9042$ . Thus, there are at most ten observations that are either less than or equal to 16.9708 or greater than or equal to 55.9042.

The given information for this example was obtained from the data set  $X = 10, 11, 13, 25, 28, 29, 33, 40, 43, 43, 48, 48, 48$ . The three observations 10, 11, 13 are the only ones that are at least 1.25 standard deviations from the mean. We can see from this example that the upper bound provided by inequality (2.4.1) need not be sharp.

**Example 2.** Suppose that a data set  $X$  of 32 observations has mean  $\bar{x} = 10$  and standard deviation  $s_X = 3$ . How many of the observations can be less than or equal to 4?

**Solution.** Observe that  $4 = 10 - 6 = 10 - 2 \times 3 = \bar{x} - 2 \times s_X$ . An observation less than or equal to 4 is therefore at least 2 standard deviations from the mean. According to inequality (2.4.1), the fraction  $N_p(2)/32$  of observations that distance from the mean satisfies  $N_p(2)/32 < 1/(2)^2$ , or  $N_p(2) < 8$ . Thus, at most seven of the observations can be less than or equal to 4.

## Chebyshev's Lower Bound for the Number of Central Observations

We retain the notation that  $N_c(\lambda)$  represents the number of observations  $x_j$  that satisfy  $|x - \bar{x}| < \lambda \cdot s_X$ . Then  $N_c(\lambda) = N - N_p(\lambda)$ , or  $N_c(\lambda)/N = 1 - N_p(\lambda)/N$ . Using inequality (2.4.1), and remembering that multiplying both sides of an inequality by a negative number reverses the inequality, we obtain

$$\frac{N_c(\lambda)}{N} > 1 - \frac{1}{\lambda^2}. \quad (2.4.2)$$

Inequality (2.4.2) is equivalent to inequality (2.4.1) and it might also be called Chebyshev's Inequality (but rarely is). In words, it states,

If  $\lambda$  is any positive number and  $X$  is any data set, then the fraction of observations in  $X$  that are less than  $\lambda$  standard deviations from the mean of  $X$  is greater than  $1 - 1/\lambda^2$ .

**Example 3.** Fill in the blanks with a percentage:

In data set  $X$ , at least \_\_\_\_\_ % of the observations are within  $\sqrt{2}$  standard deviations of the mean.

In data set  $X$ , at least \_\_\_\_\_ % of the observations are within 1.5 standard deviations of the mean.

In data set  $X$ , at least \_\_\_\_\_ % of the observations are within 2 standard deviations of the mean.

In data set  $X$ , at least \_\_\_\_\_ % of the observations are within 3 standard deviations of the mean.

In data set  $X$ , at least \_\_\_\_\_ % of the observations are within 5 standard deviations of the mean.

In data set  $X$ , at least \_\_\_\_\_ % of the observations are within 10 standard deviations of the mean.

**Solution.** If we replace  $\lambda$  with  $\sqrt{2}$  in inequality (2.4.2), then  $1 - 1/\lambda^2 = 1/2$ , and, as a result, we see that at least 50% of the observations are within  $\sqrt{2}$  standard deviations of the mean. Similarly, at least 55.55% of the observations are within 1.5 standard deviations of the mean, at least 75% of the observations are within 2 standard deviations of the mean, at least 88.88% of the observations are within 3 standard deviations of the mean, at least 96% of the observations are within 5 standard deviations of the mean, and at least 99% of the observations are within 10 standard deviations of the mean.

## Exercises

1. A data set is binned as in the following table:

Bin	[0,10)	[10,20)	[20,30)	[30,40)
Frquency	4	6	10	4
Mean per Bin	5	14	24	34

The last row of the table gives, for each bin, the average of the data values that fall inside that bin. What is the mean of the entire data set?

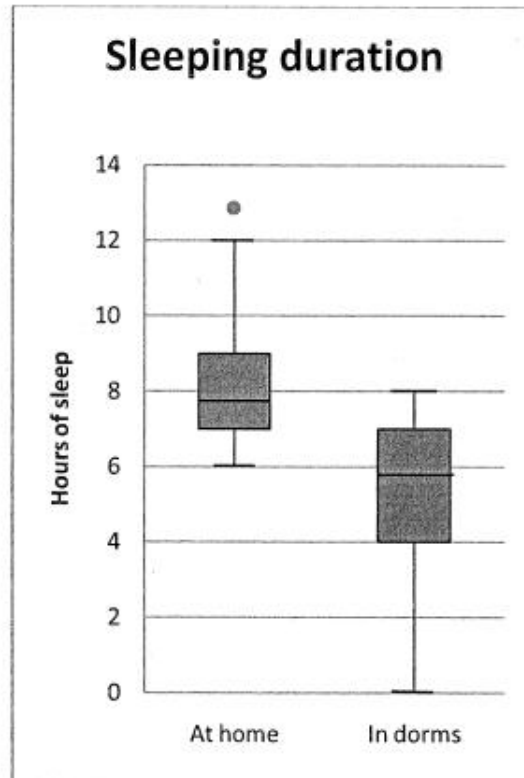
2. What is the interquartile range for the data set consisting of all the integers, 126, 127, 128,  $\dots$ , 223, 224, 225, from 126 to 225 inclusive? (Washington University exam, Fall 2010)
3. Consider the following stem-and-leaf plot:

10		0 3 5 6
9		0 2 2 4 4 7 9
8		0 1 2 3 4 5 6 7 8
7		4 5 5 6 6 7 7 9
6		0 1 2 3 6 9
5		0 6 8 8
4		0 0 7
3		4 4 4
2		2 5
1		9
0		2

- a) What are the modes? This is a confusing question—the answer depends on the meaning of *mode*. Identify two modes, one of each type that the definition allows.
- b) Which of the adjectives, uniform, unimodal, bimodal, multimodal, symmetric, negatively skewed, positively skewed, describe the distribution?
- c) What is the class width of the implicit histogram that results from the stem-and-leaf plot?
- d) What are the quartiles and IQR of this distribution?
- e) Draw a box-and-whiskers plot of the distribution.
- f) Draw a frequency polygon of the distribution.
4. A computer recorded the defects of a product manufactured by a new process. The mean number of defects per manufactured item was 100 and standard deviation was 10. An investigation of these suspiciously large numbers revealed that a programming bug caused a 0 to be appended at the right to every defect count (so that, for example, 13 defects were recorded as 130). Can the correct mean and standard deviation be determined from the incorrect values?
- A) No.
- B) Yes; The correct mean is 1 and the correct standard deviation is 1.
- C) Yes; The correct mean is 10 and the correct standard deviation is 1.
- D) Yes; The correct mean is 10 and the correct standard deviation is 10.
- E) Yes; The correct mean is 100 and the correct standard deviation is 100.
- F) Yes; The correct mean is 1000 and the correct standard deviation is 100.

(Washington University exam, Fall 2010)

5. Suppose that a quantitative variable takes the values 1, 1, 2, 3, 3. Use appropriate numbers to complete the following sentences about the whiskers of its vertical Tukey boxplot: The lower fence is at \_\_\_\_\_ and the upper fence is at \_\_\_\_\_. The bottom whisker extends from \_\_\_\_\_ (at the bottom edge of the box) to \_\_\_\_\_, and the top whisker extends from \_\_\_\_\_ (at the top edge of the box) to \_\_\_\_\_. (Washington University exam, Fall 2010)
6. The five-number summary of a data set  $X$  is 10, 24, 46, 60, 80. A histogram is created by using  $x_{[1]}$  as the lowest class limit and setting the class width to be  $\text{IQR}/3$ . There are no empty bins.
- a) What is the class width?
- b) Counting from the left, what is the first class mark?
- c) How many bins are there?
- d) What is the largest class limit?
- e) In what bin, counting from the left, does the median fall (whether or not it is a datum)?
- f) In a Tukey boxplot of  $X$ , how long are the whiskers?
7. In a survey of undergraduates designed to study the average number of hours of sleep per night, the data was split into two groups: one during the holidays at home, and one during term in school. Boxplots for the two groups are shown in the figure below:



Consider the following conclusions:

- i) The maximum sleep duration overall was 12 hours.
- ii) The distribution of sleep hours at home is skewed to the right.
- iii) The distribution of sleep hours in the dorm is skewed to the right.
- iv) The minimum sleep duration at home was 6 hours.

Which is the correct assessment of the correctness of the preceding conclusions?

- A) Every one of (i), (ii), (iii), and (iv) is true.
- B) Only (ii) is true.
- C) Only (iii) is true.
- D) Only (iii) and (iv) are true.
- E) Only (ii) and (iv) are true.
- F) Only (iii) is false.

(Washington University exam, Fall 2010)

8. Which of these variables is most likely to be bimodal? (Washington University exam, Spring 2014)
  - A) eye color
  - B) number of TV sets at home
  - C) hours of homework last week

- D) number of cigarettes smoked daily
- E) head circumference

9. (Washington University exam, Spring 2014)

The advantage of making a stem-and-leaf display instead of a dotplot is that a stem-and-leaf display:

- A) is for quantitative data, while a dotplot shows categorical data.
  - B) preserves the individual data values.
  - C) shows the shape of the distribution better than a dotplot.
  - D) satisfies the area principle.
  - E) none of these
10. An application of Chebyshev's Inequality to a particular data set  $X$  of size 25 tells us that the fraction of observations in the open interval  $(18,38)$  is greater than  $16/25$ . What are  $\bar{x}$  and  $s_X$ ?