Chapter 3
Descriptive Statistics II: Numerical Summary Values

3.1 Numerical summary values for quantitative data

For many purposes a few well–chosen numerical summary values (statistics) will suffice as a description of the distribution of a quantitative variable. A **statistic** is a numerical characteristic of a sample. More formally, a statistic is a numerical quantity computed from the values of a variable, or variables, corresponding to the units in a sample. Thus a statistic serves to quantify some interesting aspect of the distribution of a variable in a sample. Summary statistics are particularly useful for comparing and contrasting the distribution of a variable for two different samples.

If we plan to use a small number of summary statistics to characterize a distribution or to compare two distributions, then we first need to decide which aspects of the distribution are of primary interest. If the distributions of interest are essentially mound shaped with a single peak (unimodal), then there are three aspects of the distribution which are often of primary interest. The first aspect of the distribution is its location on the number line. Generally, when speaking of the location of a distribution we are referring to the location of the "center" of the distribution. The location of the center of a symmetric, mound shaped distribution is clearly the point of symmetry. There is some ambiguity in specifying the location of the center of an asymmetric, mound shaped distribution and we shall see that there are at least two standard ways to quantify location in this context. The second aspect of the distribution is the amount of variability or dispersion in the distribution. Roughly speaking, we would say that a distribution exhibits low variability if the observed values tend to be close together on the number line and exhibits high variability if the observed values tend to be more spread out in some sense. The third aspect is the shape of the distribution and in particular the degree of skewness in the distribution.

As a starting point consider the **minimum** (smallest observed value) and **maximum** (largest observed value) as statistics. We know that all of the data values lie between the minimum and the maximum, therefore, the minimum and the maximum provide a crude quantification of location and variability. In particular, we know that all of the values of the variable are restricted to the interval from the minimum to the maximum; however, the minimum and the maximum alone tell us nothing about how the data values are distributed within this interval. If the distribution is reasonably symmetric and mound shaped, then the **midrange**, defined as the average of the minimum and the maximum, may provide a suitable quantification of the location of the center of the distribution. The median and mean, which are defined below, are generally better measures of the center of a distribution.

The **range**, defined as the distance from the minimum to the maximum can be used to quantify the amount of variability in the distribution. Note that the range is the positive number obtained by subtracting the minimum from the maximum. When comparing two distributions the distribution with the larger range will generally have more variability than the distribution with the smaller range; however, the range is very sensitive to extreme observations so that one or a few unusually large or small values can lead to a very large range.

We will now consider an approach to the quantification of the shape, location, and variability of a distribution based on the division of the histogram of the distribution into sections of equal area. This is equivalent to dividing the data into groups, each containing the same number of values. We will first use a division of the histogram into halves. We will then use a division of the histogram into fourths.

The median is used to quantify the location of the center of the distribution. In terms of area, the **median** is the number (point on the number line) with the property that the area in the histogram to the left of the median is equal to the area to the right of the median. Here and in the sequel we will use a lower case n to denote the sample size, *i.e.*, n will denote the number of units in the sample. In terms of the n observations, the **median** is the number with the property that at least $n/2$ of the observed values are less than or equal to the median and at least $n/2$ of the observed values are greater than or equal to the median.

A simple procedure for finding the median, which is easily generalized to fractions other that 1/2, is outlined below.

1. Arrange the data (observations) in increasing order from the smallest (obs. no. 1) to the largest (obs. no. n). Be sure to include all n values in this list, including repeats if there are any.
2. Compute the quantity $n/2$.

3a. If $n/2$ is not a whole number, round it up to the next largest integer. The observation at the location indicated by the rounded–up value in the ordered listing of the data is the median.

3b. If $n/2$ is a whole number, then we need to average two values to get the median. The two observations to be averaged are obs. no. $n/2$ and the next observation (obs. no. $n/2 + 1$) in the ordered listing of the data. Find these two observations and average them to get the median.

We can use the distance between the minimum and the median and the distance between the median and the maximum to quantify the amount of skewness in the distribution. The distance between the minimum and the median is the range of the lower (left) half of the distribution, and the distance between the median and the maximum is the range of the upper (right) half of the distribution. If the distribution is symmetric, then

these two distances (median – minimum) and (maximum – median) will be equal. If the distribution is skewed, then we would expect to observe a larger range (indicating more variability) for the half of the distribution in the direction of the skewness. Thus if the distribution is skewed to the left, then we would expect (median – minimum) to be greater than (maximum – median). On the other hand, if the distribution is skewed to the right, then we would expect (maximum – median) to be greater than (median – minimum).

Example. Weed seeds (revisited). Recall that this example is concerned with the number of weed seeds found in $n = 98$ quarter–ounce batches of grass. Since $98/2 = 49$, the median for this example is the average of observations 49 and 50. Referring to Table 6 of Chapter 2 we find that the minimum number of weed seeds is 0, the maximum is 7, and the median is 1, since observations 49 and 50 are each 1. The range for this distribution is $7 - 0 = 7$. Notice that the range of the right half of this distribution (maximum – median) $= 7 - 1 = 6$ is much larger than the range of the left half (median – minimum) $= 1 - 0 = 1$ confirming our observation that this distribution is strongly skewed to the right.

Example. Vole reproduction (revisited). Recall that this example is concerned with the number of babies in $n = 170$ litters of voles. Since $170/2 = 85$, the median for this example is the average of observations 85 and 86. Referring to Table 7 of Chapter 2 we find that the minimum number of babies is 1, the maximum is 11, and the median is 6, since observations 85 and 86 are each 6. The range for this distribution is $11 - 1 = 10$. Notice that the range of the right half of this distribution (maximum – median) $= 11 - 6 = 5$ is equal to the range of the left half (median – minimum) $= 6 - 1 = 5$ confirming our observation that this distribution is symmetric.

A more detailed quantification of the shape and variability of a distribution can be obtained from a division of the distribution into fourths. In order to divide a distribution into fourths, we need to specify three numbers or points on the number line. These statistics are called **quartiles**, since they divide the distribution into quarters. In terms of area, the **first quartile**, denoted by Q_1 (read this as Q sub one), is the number (point on the number line) with the property that the area in the histogram to the left of Q_1 is equal to one fourth and the area to the right of Q_1 is equal to three fourths. The **second quartile**, denoted by Q_2, is the median. The **third quartile**, denoted by Q_3, is the number (point on the number line) with the property that the area in the histogram to the left of Q_3 is equal to three fourths and the area to the right of Q_3 is equal to one fourth. In terms of the n observations, Q_1 is the number with the property that at least $n/4$ of the observed values are less than or equal to Q_1 and at least $3n/4$ of the observed values are greater than or equal to Q_1. Similarly, Q_3 is the number with the property that at least $3n/4$ of the observed values are less than or equal to Q_3 and at least $n/4$ of the observed values are greater than or equal to Q_3.

The method for finding the median given above is readily modified for finding the first and third quartiles. For Q_1, we simply replace $n/2$ by $n/4$ and replace the words 'the median' by Q_1. To find Q_3, use exactly the same method but count down from the largest value instead of counting up from the smallest value. Some calculators and computer programs use variations of the methods given above for finding Q_1 and Q_3. These variations may give slightly different values for Q_1 and Q_3.

Example. Weed seeds (revisited). Since $98/4 = 24.5$, the quartiles Q_1 and Q_3 for this example are the observations located at position 25 counting up for Q_1 and counting down for Q_3. Referring to Table 6 of Chapter 2 we find that $Q_1 = 0$ and $Q_3 = 2$. Notice that the range of the lower three–fourths of this distribution, Q_3 – minimum, is 2 while the range of the upper fourth, maximum – Q_3 is 5. This indicates that 75% (a large proportion) of the batches of grass have relatively few weed seeds, and the skewness in this distribution is due to the high amount of variability among the numbers of weed seeds in the 25% of the batches with between 2 and 7 weed seeds.

Previously we introduced the range as a measure of variability. An alternative measure of variability is provided by the interquartile range. The **interquartile range** (IQR) is the distance between the first quartile Q_1 and the third quartile Q_3, *i.e.*, the interquartile range is the positive number obtained by subtracting Q_1 from Q_3. Notice that the **interquartile range** is the range of the middle half of the distribution. The interquartile range is less sensitive to the presence of a few extreme observations in the data than is the range. For example, if there are one or two unusually large or unusually small values, then these values may have the effect of making the range much larger than it would be if these unusual values were not present. In such a situation, we might argue that the range is too large to be deemed an appropriate overall measure of the variability of the distribution. The interquartile range is not affected by a few unusual values, since it only depends on the middle half of the data. We could use the range of a larger part of the middle of the distribution, say the middle 75% or 90%, as a compromise between the range and the interquartile range.

The five summary statistics: the minimum (min), the first quartile (Q_1), the median (med), the third quartile (Q_3), and the maximum (max), constitute the **five number summary** of the distribution. Each of these five statistics provides a quantification of a particular aspect of the distribution. They quantify where the distribution begins, where the first quarter of the distribution ends, and so on. Furthermore, the distances between these five statistics can be used to quantify the shape (skewness) of the distribution.

The four distances: (Q_1 – min), (med – Q_1), (Q_3 – med), and (max – Q_3), are the ranges of the first, second, third, and fourth quarters of the distribution, respectively. These distances can be used to quantify the amount of variability in the corresponding parts of the distribution. Comparisons of appropriate pairs of these distances provide

indications of certain aspects of the shape of the distribution. The relationship between (med – Q_1) and (Q_3 – med) can be used to quantify the shape (skewness) of the middle half of the distribution. Since (Q_1 – min) and (max – Q_3) are the lengths of the tails (lower and upper fourths) of the distribution, the relationship between these numbers can be used to quantify skewness in the tails of the distribution.

Example. Cholesterol levels in Guatemalans. This example is taken from Devore and Peck, *Statistics*, 3 ed., (1997), Duxbury, p. 23. The original source is "The Blood Viscosity of Various Socioeconomic Groups in Guatemala" in *The American Journal of Clinical Nutrition*, Nov., 1964, 303–307. The Institute of Nutrition of Central America and Panama measured the serum total cholesterol levels for a group of 49 adult, low–income rural Guatemalans and for a group of 45 adult, high–income urban Guatemalans. The serum total cholesterol levels (in mg/dL) are provided in Table 1 and stem and leaf histograms are given in Figure 1.

Table 1. Guatemalan cholesterol data.

Rural group cholesterol levels (in mg/dL).

95	108	108	114	115	124	129	129	131	131
135	136	136	139	140	142	142	143	143	144
144	145	146	148	152	152	155	157	158	158
162	165	166	171	172	173	174	175	180	181
189	192	194	197	204	220	223	226	231	

Urban group cholesterol levels (in mg/dL).

133	134	155	170	175	179	181	184	188	189
190	196	197	199	200	200	201	201	204	205
205	205	206	214	217	222	222	227	227	228
234	234	236	239	241	242	244	249	252	273
279	284	284	284	330					

Before we compute any summary statistics consider the stem and leaf histograms in Figure 1. Based on these histograms we can see that both of these cholesterol level distributions are basically mound shaped with some skewness to the right. In the rural group there are four individuals with somewhat high cholesterol levels (220 or more); there is a gap of 16 separating the cholesterol levels of these individuals from the rest of the rural group. It is this group of four observations which causes the rural distribution to appear skewed to the right. The urban group has similar slightly unusual groups of cholesterol levels; one group having somewhat low levels and one having somewhat high levels. There is one unusually large value (330) in the urban group that we might consider an outlier, since there is a gap of 46 between 330 and the next largest value. (An outlier is an observation that is widely separated from the majority of a distribution.) We will need to

consider the implications of this outlier in our analysis of this example. It is also apparent that the people in the urban group tend to have higher cholesterol levels than the people in the rural group. There appears to be more variability among the cholesterol levels for the urban group. With the urban outlier there appears to be much more variability in the cholesterol levels of the urban group, and without it there appears to be slightly more variability in the urban group cholesterol levels. If we ignore the outlier, the urban group distribution appears to be essentially symmetric.

Figure 1. Guatemalan cholesterol stem and leaf histograms.

The stem represents tens and the leaf represents ones. (mg/dL)

Rural		Urban	
9	5	9	
10	88	10	
11	45	11	
12	499	12	
13	115669	13	34
14	0223344568	14	
15	225788	15	5
16	256	16	
17	12345	17	059
18	019	18	1489
19	247	19	0679
20	4	20	001145556
21		21	47
22	036	22	22778
23	1	23	4469
24		24	1249
25		25	2
26		26	
27		27	39
28		28	444
29		29	
30		30	
31		31	
32		32	
33		33	0

The five number summaries and the associated distances based on them are provided, for the rural group, for the entire urban group, and for the urban group omitting 330, in Table 2. The steps involving in computing the medians and quartiles, for the rural group

and the entire urban group, are outlined below. For the rural group there are $n = 49$ observations so that

(1) $49/2 = 24.5$, thus the median 152 is obs. no. 25, corresponding to the first 2 leaf in the 15 stem.

(2) $49/4 = 12.25$, thus the first and third quartiles are $Q_1 = 136$, the 13^{th} observation counting up, corresponding to the second 6 leaf in the 13 stem, and $Q_3 = 174$, the 13^{th} observation counting down, corresponding to the second 4 leaf in the 17 stem.

For the urban group there are $n = 45$ observations so that

(1) $45/2 = 22.5$, thus the median 206 is obs. no. 23, corresponding to the 6 leaf in the 20 stem.

(2) $45/4 = 11.25$, thus the first and third quartiles are $Q_1 = 196$, the 12^{th} observation counting up, corresponding to the 6 leaf in the 19 stem, and $Q_3 = 239$, the 12^{th} observation counting down, corresponding to the 9 leaf in the 23 stem.

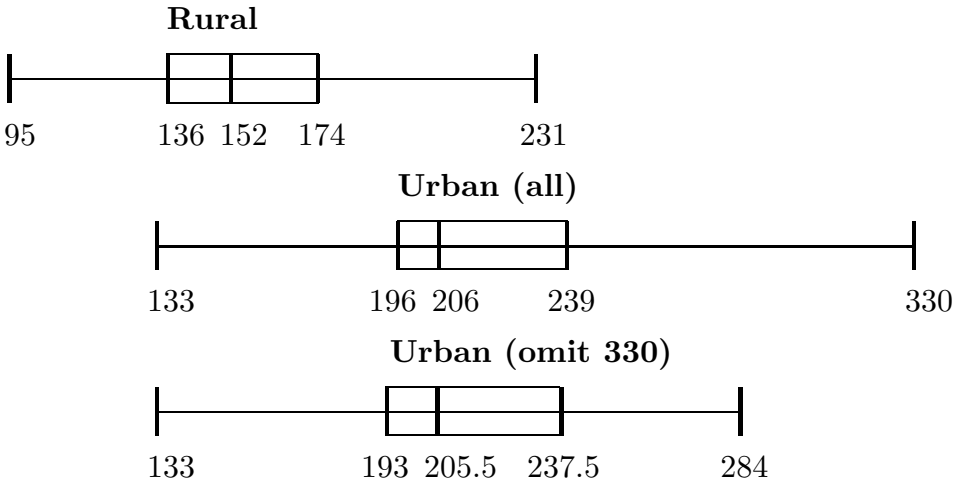
Table 2. Five number summaries with distances.

Rural group. (mg/dL) n=49					
min:	95				
Q_1:	136	Q_1- min:	41		
med:	152	med - Q_1:	16	med - min:	57
Q_3:	174	Q_3- med:	22		
max:	231	max - Q_3:	57	max - med:	79
Urban group (all). (mg/dL) n=45					
min:	133				
Q_1:	196	Q_1- min:	63		
med:	206	med - Q_1:	10	med - min:	73
Q_3:	239	Q_3- med:	33		
max:	330	max - Q_3:	91	max - med:	124
Urban group (omit 330). (mg/dL) n=44					
min:	133				
Q_1:	193	Q_1- min:	60		
med:	205.5	med - Q_1:	12.5	med - min:	72.5
Q_3:	237.5	Q_3- med:	32		
max:	284	max - Q_3:	46.5	max - med:	78.5

Before we continue with our discussion of this example we will introduce a simple graphical display corresponding to the information in Table 2. We can use the five number summary values to form a simple graphical representation of a distribution known as a

box plot or a box and whiskers plot. A box plot does not convey as much information as a stem and leaf histogram but it does give a useful graphical impression of the shape of the distribution as well as its location and variability. Simple box plots for the Guatemalan cholesterol example are provided in Figure 2.

Figure 2. Box plots for cholesterol level.



Notice that each box plot has five vertical marks indicating the locations of the five number summary values. The box which extends from the first quartile to the third quartile and is divided into two parts by the median gives an impression of the distribution of the values in the middle half of the distribution. In particular, a glance at this box indicates whether the middle half of the distribution is skewed or symmetric and indicates the magnitude of the interquartile range (the length of the box). The line segments (whiskers) which extend from the ends of the box to the extreme values (the minimum and the maximum) give an impression of the distribution of the values in the tails of the distribution. The relative lengths of the whiskers indicate the contribution of the tails of the distribution to the symmetry or skewness of the distribution.

Returning to the cholesterol example first consider the shapes of the cholesterol distributions. We can use the distances, based on the five number summary, given in Table 2 to quantify the degree of skewness in these distributions. Comparing the distances for the rural group we find that max – med = 79 > 57 = med – min, Q_3 – med = 22 > 16 = med – Q_1, and Max – Q_3 = 57 > 41 = Q_1 – min. All of these comparisons support our contention that the cholesterol distribution for the rural group is skewed right. For the urban group, including the outlier, we have max – med = 124 > 73 = med – min, Q_3 – med = 33 > 10 = med – Q_1, and Max – Q_3 = 91 > 63 = Q_1 – min. All of these comparisons support our contention that the cholesterol distribution for the urban group is skewed right. If we omit the outlier (330) from the urban group we find that max –

med = 78.5 is only slightly larger than med – min = 72.5 suggesting that without the outlier the cholesterol distribution for the urban group is reasonably symmetric. Without the outlier the middle half of the distribution is still somewhat skewed right, since Q_3 – med = 32 > 12.5 = med – Q_1; but, the range of the left tail (lower fourth) Q_1 – min = 60 is now larger than the range of the right tail (upper fourth) Max – Q_3 = 46.5.

The fact that the median 152 for the rural group is much smaller than the median 206 (with the outlier) or 205.5 (without the outlier) of the urban group supports our contention that the people in the urban group tend to have higher cholesterol levels than the people in the rural group.

With the outlier the range 197 for the urban group is much larger than the range 137 for the rural group. If we omit the outlier, then the range for the urban group is 151 which is still larger than 137 but not by so much. On the other hand, if we consider the interquartile ranges, 38 for the rural group and 43 (44.5 without the outlier) for the urban group, we find that there is a similar amount of variability in the middle halves of these distributions. Hence, our contention that there is much more variability among the cholesterol levels of the urban Guatemalans depends very heavily on the cholesterol level of one individual. Whether we include this individual or not, we are justified in claiming that there is more variability among the cholesterol levels of the urban Guatemalans.

Based on our analysis of these cholesterol level distributions we might propose several hypotheses or conjectures about why these distributions differ as they do. First we might conjecture that the rural Guatemalans are probably more physically active and eat food which is lower in fat than the urban Guatemalans. This would cause the rural Guatemalans to tend to have lower cholesterol levels. Second, we might argue that there is less variability in the cholesterol levels of the rural Guatemalans because their lifestyles and eating habits are probably quite similar.

The approach that we have been using to form summary statistics is to select a single representative value from the observed values of the variable (or the average of two adjacent observed values) to quantify a particular aspect of the distribution. We have also considered statistics that are distances between two such representative values.

An alternative approach to forming a summary statistic is to combine all of the observed values to get a suitable statistic. The first statistic of this type that we consider is the mean. The **mean**, which is the simple arithmetic average of the n data values, is used to quantify the location of the center of the distribution. You could compute the mean by adding all n data values together and dividing this sum by n; however, it is better to use a calculator or a computer.

The sample mean is often denoted by the symbol $\overline{X}$ (read this as X bar). This is a convenient place for us to introduce some standard notation. It is standard practice to use a letter, such as X, to denote a variable and the values of the variable. You are

free to choose a letter with mnemonic value instead of the generic letter X; however, you should not use S or Z as these letters are reserved for special uses. If X denotes the variable of interest, then we will use $\overline{X}$ to denote the mean of the distribution of X. If we used a different letter, say Y, to denote the variable, then we would use $\overline{Y}$ to denote the corresponding mean. We will use function notation to denote the other statistics we defined above. That is, if X denotes the variable, then $\min(X)$, $Q_1(X)$, $\text{med}(X)$, $Q_3(X)$, $\max(X)$, $\text{range}(X)$, and $\text{IQR}(X)$ denote the minimum, the first quartile, the median, the third quartile, the maximum, the range, and the interquartile range, respectively. You should read these symbols as follows: read $\min(X)$ as the minimum of X, $Q_1(X)$ as the first quartile of X, and so on.

Recall that the median is the number (point on the number line) with the property that the area in the histogram to the left of the median is equal to the area to the right of the median. The mean is the number (point on the number line) where the histogram would balance. To understand what we mean by the balance point, imagine the histogram as being cut out of a piece of cardboard. The mean is located at the point along the number line side of this cutout where the histogram cutout would balance. These geometric characterizations of the mean and the median imply that when the distribution is symmetric the mean will be equal to the median. Furthermore, if the distribution is skewed to the right, then the mean (the balance point) will be larger than the median (to the right of the median). Similarly, if the distribution is skewed to the left, then the mean (the balance point) will be smaller than the median (to the left of the median).

The primary use of the mean, like the median, is to quantify the location of the center of a distribution and to compare the locations (centers) of two distributions. Since both the mean and the median can be used to quantify the location of the center of a distribution, it seems reasonable to ask which is more appropriate. If the distribution is approximately symmetric, then the mean and the median will be approximately equal. On the other hand, if the distribution is not symmetric, then the median is likely to provide a better indication of the center of the distribution. For example, if the distribution is strongly skewed to the right, then the mean may be much larger than the median and the mean may not be a good indication of the center of the distribution. For a specific problem it is a good idea to mark the locations of the mean and the median on a histogram of the distribution and consider which seems more reasonable as an indicator of the center of the distribution.

If the mean $\overline{X}$ is deemed suitable as a measure of the center of the distribution of X, then the deviations $(X - \overline{X})$ of the observed values of X from their mean $\overline{X}$ contain information about the amount of variability in the distribution. If there is little variability (the observed values of X are close together and they are close to the mean $\overline{X}$), then the deviations $(X - \overline{X})$ will tend to be small in magnitude (absolute value). On the other

hand, if there is a lot of variability (at least some of the observed values of X are far apart and they are not all close to the mean $\overline{X}$), then the deviations $(X-\overline{X})$ will tend to be large in magnitude. It is this observation which suggests that a summary statistic based on the distances between each of the observed values of the variable and their mean can be used to measure the variability in the distribution. The standard deviation is one such statistic. The **standard deviation** is the square root of the "average" of the squared deviations of the observed values of the variable from their mean. A formula for the standard deviation is given below; however, you should not use this formula to compute the standard deviation. Instead you should use a calculator or a computer to compute the standard deviation. In symbols, the **standard deviation** of the distribution of the variable X, denoted by S_X (read this as S sub X), is

$$S_X = \sqrt{\frac{\Sigma(X-\overline{X})^2}{n-1}}$$

In this formula the capital Greek letter sigma, Σ, represents the statement "the sum of ", and $(X-\overline{X})^2$ denotes the square of the distance from the observed value X to the mean $\overline{X}$. Therefore, the expression under the square root sign in the formula is the "average" of the squared deviations of the observed values of the variable from their mean as mentioned above. The reason for the square root is so that the standard deviation of X and the variable X are in the same units of measurement.

The standard deviation is positive, unless there is no variability at all in the data. That is, unless all of the observations are exactly the same, the standard deviation is a positive number. The standard deviation is a very widely used measure of variability. Unfortunately, the standard deviation does not have a simple, direct interpretation. The important thing to remember is that larger values of the standard deviation indicate that there is more variability in the data. A closely related measure of variability is the **variance** which is simply the square of the standard deviation, *i.e.*, the variance of the distribution of X is $S_X^2 = \sum(X-\overline{X})^2/(n-1)$.

There are quotation marks around the word average in the definition of the standard deviation because we divided by $n-1$ even though there are n squared deviations in the average. The reason for this is that, in a sense, there are only $n-1$ individual pieces of information contained in the collection of n deviations from the mean. It is readily verified that the sum of the deviations from the mean (not the sum of their squares) is equal to zero, *i.e.*, $\Sigma(X-\overline{X}) = 0$. This is the algebraic version of the fact that the mean is the balance point of the distribution. Because of this fact, if we know the values of any $n-1$ of the deviations, then we can determine the value of the remaining deviation. This is the sense in which there are only $n-1$ individual pieces of information contained in the collection of n deviations from the mean; and is the reason that we divide by $n-1$.

The means, medians, standard deviations, ranges, and interquartile ranges for the Guatemalan cholesterol level distributions are given in Table 3. Because all three of these distributions are somewhat skewed to the right, we find that in all three cases the mean is larger than the median. Notice the effects of excluding the outlier from the urban group on these statistics. First consider the mean and the median; excluding this outlier has essentially no effect on the median but has an appreciable effect on the mean. This illustrates the sensitivity of the mean to extreme observations. Next consider the three measures of variability. As we noted above, excluding the outlier has a large effect on the range but little effect on the interquartile range. As with the mean, excluding the outlier has an appreciable effect on the standard deviation. This illustrates that, like the mean, the standard deviation is also sensitive to extreme observations.

In this example, if we base our comparisons of the location and the amount of variability in these distributions on the mean and standard deviation we reach essentially the same conclusions as we did when using the five number summary.

Table 3. Summary statistics for the cholesterol example.

group	mean	median	std. dev.	range	IQR
rural	157.02	152	31.75	137	38
urban (all)	216.87	206	39.92	197	43
urban (omit 330)	214.30	205.5	36.42	151	42

Example. EPA mileage values for subcompact cars. Table 4 contains the EPA mileage values and some related information for 56 subcompact car model/engine combinations. This information was obtained from the June 2000 edition of the model year 2000 fuel economy guide provided on the DOT/EPA web site `www.fueleconomy.gov`. If there were two or more listings for the same car model/engine combination, then only one value was included. In particular, if mileage values were provided for a particular car model/engine combination with both automatic and manual transmissions, then only the mileage value for the manual transmission was included. The car models listed in the EPA fuel economy guide are grouped into size classes based on the combined passenger and cargo volume of the car. For example, subcompact cars have combined volumes between 85 and 99 cubic feet and compact cars have combined volumes between 100 and 109 cubic feet. For this example we will consider the two mileage values (city and highway) as response variables. The other variables in Table 4 might serve as potentially interesting explanatory variables. In this example a car model/engine combination is a unit and we have a pair of responses, city mileage and highway mileage, for each car model. We will first consider the distributions of city and highway mileage values separately, ignoring the fact that we have pairs of mileage values for each model.

Table 4. Model year 2000 subcompact car EPA mileage values.

city denotes city mileage in miles per gallon
hiwy denotes highway mileage in miles per gallon
trans denotes transmission type (automatic or manual) and number of gears
displ denotes engine displacement in liters
cyl denotes number of cylinders
drv denotes front, rear, or all wheel drive

manufacturer	model	city	hiwy	trans	displ	cyl	drv
Acura	Integra	25	31	(M5)	1.8	4	F
Acura	Integra(DOHC/VTEC)	25	30	(M5)	1.8	4	F
Bentley	Azure	11	16	(A4)	6.8	8	R
Bentley	Continental SC	11	16	(A4)	6.8	8	R
Bentley	Continental T	11	16	(A4)	6.8	8	R
BMW	323CI	20	29	(M5)	2.5	6	R
BMW	328CI	21	29	(M5)	2.8	6	R
Chevrolet	Camaro	19	30	(M5)	3.8	6	R
Chevrolet	Camaro	18	27	(M6)	5.7	8	R
Chevrolet	Cavalier	24	34	(M5)	2.2	4	F
Chevrolet	Cavalier	23	33	(M5)	2.4	4	F
Chevrolet	Metro	39	46	(M5)	1	3	F
Chevrolet	Metro	36	42	(M5)	1.3	4	F
Ferrari	Ferrari 456 MGT/MGTA	10	16	(M6)	5.5	12	R
Ford	Escort ZX2	25	33	(M5)	2	4	F
Ford	Mustang	20	29	(M5)	3.8	6	R
Ford	Mustang	17	25	(M5)	4.6	8	R
Ford	Mustang(4 Valve)	17	24	(M5)	4.6	8	R
Honda	Civic	32	37	(M5)	1.6	4	F
Honda	Civic(VTEC)	30	35	(M5)	1.6	4	F
Honda	Civic(DOHC/VTEC)	26	31	(M5)	1.6	4	F
Honda	Prelude	22	27	(M5)	2.2	4	F
Hyundai	Tiburon	23	32	(M5)	2	4	F
Jaguar	XK8	18	25	(A5)	4	8	R
Jaguar	XKR	16	23	(A5)	4	8	R
Lexus	SC 300/SC 400	19	23	(A4)	3	6	R
Lexus	SC 300/SC 400	18	25	(A5)	4	8	R
Mercedes–Benz	CLK320	21	29	(A5)	3.2	6	R
Mercedes–Benz	CLK430	18	25	(A5)	4.3	8	R
Mitsubishi	Eclipse	23	31	(M5)	2.4	4	F
Mitsubishi	Eclipse	20	28	(M5)	3	6	F
Mitsubishi	Mirage	33	40	(M5)	1.5	4	F
Mitsubishi	Mirage	28	36	(M5)	1.8	4	F

This table is continued on the next page.

Table 4. Model year 2000 subcompact car EPA mileage values (continued from the preceding page).

manufacturer	model	city	hiwy	trans	displ	cyl	drv
Pontiac	Firebird/TransAm	19	30	(M5)	3.8	6	R
Pontiac	Firebird/TransAm	18	27	(M6)	5.7	8	R
Pontiac	Sunfire	24	34	(M5)	2.2	4	F
Pontiac	Sunfire	23	33	(M5)	2.4	4	F
Rolls–Royce	Corniche	11	16	(A4)	6.8	8	R
Saab	Saab 9-3 Conv.	22	29	(M5)	2	4	F
Saab	Saab 9-3 Viggen Conv.	20	29	(M5)	2.3	4	F
Saturn	SC	28	40	(M5)	1.9	4	F
Saturn	SC(DOHC)	27	38	(M5)	1.9	4	F
Subaru	Impreza AWD	23	29	(M5)	2.2	4	A
Subaru	Impreza AWD	21	28	(M5)	2.5	4	A
Suzuki	Esteem	30	37	(M5)	1.6	4	F
Suzuki	Esteem	28	35	(M5)	1.8	4	F
Suzuki	Swift	36	42	(M5)	1.3	4	F
Toyota	Solara Conv.	23	30	(A4)	2.2	4	F
Toyota	Solara Conv.	19	26	(A4)	3	6	F
Toyota	Celica	28	34	(M5)	1.8	4	F
Toyota	Celica	23	32	(M6)	1.8	4	F
Volkswagen	Cabrio	24	31	(M5)	2	4	F
Volkswagen	New Beetle	25	31	(M5)	1.8	4	F
Volkswagen	New Beetle	24	31	(M5)	2	4	F
Volvo	C70 Conv.	20	26	(M5)	2.3	5	F
Volvo	C70 Conv.	19	26	(A4)	2.4	5	F

The stem and leaf histograms of Figure 3 summarize the distributions of the EPA city and highway gas mileage values for the $n = 56$ model year 2000 subcompact car models. Notice that each of these distributions includes five unusually low mileage values. Five car models have city mileage values of 10 or 11 mpg and five car models have highway mileage values of 16 mpg. It turns out that the five car models with the lowest city mileage values are also the five car models with the lowest highway mileage values. In both distributions there is a large separation between the five low mileage values and the mileage values of the 51 other subcompact car models. Before we proceed with our examination of this example we need to look at the original data, including all relevant information about the car models, to see why these five car models have such low mileage values.

Figure 3. Stem and leaf histograms for model year 2000 subcompact car EPA mileage values.

The stem represents tens and the leaf represents ones. (mpg)

City	Highway
1 01111	
1	
1	
1 677	1 66666
1 8888899999	1
2 00000111	2
2 223333333	2 33
2 44445555	2 45555
2 67	2 666777
2 8888	2 889999999
3 00	3 0000111111
3 23	3 22333
3	3 44455
3 66	3 677
3 9	3 8
	4 00
	4 22
	4
	4 6

From Table 4 we find that the five subcompact car models with the lowest city and highway mileage values are: three Bentley models, one Ferrari model, and one Rolls–Royce model. This group of car models contains four ultra–luxury models and one high performance sports car. Since these five car models do not fit in with the usual conception of a subcompact car, we will remove them from the data. Thus the remainder of this discussion is restricted to the collection of $n = 51$ subcompact car models remaining after removing the five car models discussed above.

We will first make some observations based on these stem and leaf histograms. The city and highway mileage distributions both appear to be skewed to the right. This indicates that, for both the city and highway mileage values, there tends to be more variability among the larger mileage values than among the lower mileage values. Each of these mileage histograms has a single peak. The peak in the city mileage histogram is located near the lower end of the distribution while the peak in the highway mileage distribution is more centrally located. The locations of these peaks and the mound shapes of these distributions indicate that, for subcompact cars, the car mileage values tend to be clustered around the low 20's and the highway mileage values tend to be clustered around the upper

20's and lower 30's. As we would expect, the highway mileage distribution is located higher on the number line than is the city mileage distribution indicating that these subcompact cars tend to get higher mileage on the highway than they do in the city.

Table 5. Subcompact car EPA mileage summary statistics, excluding the five unusual car models.

statistic	**city**	**highway**
n	51	51
min	16	23
Q_1	19	27
med	23	30
Q_3	26	34
max	39	46
range	23	23
IQR	7	7
Q_1- min	3	4
med $-Q_1$	4	3
Q_3- med	3	4
max $-Q_3$	13	12
mean	23.53	31.12
std dev	5.27	5.18

We will now quantify and expand on our observations about the subcompact car mileage distributions. Relevant summary statistics are given in Table 5. In the discussion below, we will use C to denote the city mileage of a subcompact car model and H to denote the highway mileage of a subcompact car model.

First consider the shapes of the subcompact car mileage distributions. For the city mileage distribution we see that: $\max(C) - \text{med}(C) = 16 > 7 = \text{med}(C) - \min(C)$, $\max(C) - Q_3(C) = 13 > 3 = Q_1(C) - \min(C)$, and $\overline{C} = 23.53 > 23 = \text{med}(C)$. All of these comparisons support our contention that the city mileage distribution is skewed to the right. Notice that $Q_3(C) - \text{med}(C) = 3$ which is approximately equal to $\text{med}(C) - Q_1(C) = 4$; this suggests that the middle half of this distribution is reasonably symmetric. For the highway mileage distribution we see that: $\max(H) - \text{med}(H) = 16 > 7 = \text{med}(H) - \min(H)$, $\max(H) - Q_3(H) = 12 > 4 = Q_1(H) - \min(H)$, and $\overline{H} = 31.12 > 30 = \text{med}(H)$. All of these comparisons support our contention that the highway mileage distribution is skewed to the right. Notice that $Q_3(H) - \text{med}(H) = 4$ which is approximately equal to $\text{med}(H) - Q_1(H) = 3$; this suggests that the middle half of this distribution is also reasonably symmetric.

Next consider the locations of the subcompact car mileage distributions. The median city mileage $\text{med}(C) = 23$ is less than the median highway mileage $\text{med}(H) = 30$ and the mean city mileage $\overline{C} = 23.53$ is less than the mean highway mileage $\overline{H} = 31.12$. Both of these comparisons support our contention that the city mileages of subcompact cars tend to be lower than the highway mileages of subcompact cars. Notice that there is some overlap of the city mileages and the highway mileages indicating that some subcompact cars have city mileage values that are higher than the highway mileage values of some subcompact cars and *vice versa*.

Finally consider the variability in these subcompact car mileage distributions. The facts that: $\text{range}(C) = 23 = \text{range}(H)$, $\text{IQrange}(C) = 7 = \text{IQrange}(H)$, and $S_C = 5.27$ is approximately equal to $S_H = 5.18$, all support the contention that the variability in subcompact car city mileage values is about the same as the variability in subcompact car highway mileage values.

In our comparison of the city and highway mileages for subcompact cars, we ignored the fact that we actually have pairs of city and highway mileage values for each of the 51 car models. If we want to know how the highway mileage of a subcompact car model relates to its city mileage, then we need to base our comparison on the paired city and highway mileages. One way to do this is to consider the difference between the highway mileage and the city mileage for a car model. For each car model we will determine this difference value by subtracting its city mileage value from its highway mileage value. The highway minus city mileage differences for the $n = 51$ subcompact car models are given in Table 6. The 51 difference values in Table 6 are listed (reading across a row and then going to the next row) in the same order as the 51 city and highway mileage values are listed in Table 4, skipping the five unusual car models. A stem and leaf histogram for these differences is given in Figure 4 and the difference summary statistics are given in Table 7.

In the discussion below, we will use D to denote the difference $D = H - C$ between the highway mileage of a subcompact car model and its city mileage. From the stem and leaf histogram the shape of the subcompact car mileage difference distribution appears to be mound shaped and slightly skewed to the right. The facts: $\max(D) - \text{med}(D) = 5 > 3 = \text{med}(D) - \min(D)$, $\max(D) - Q_3(D) = 3 > 2 = Q_1(D) - \min(D)$, $Q_3(D) - \text{med}(D) = 2 > 1 = \text{med}(D) - Q_1(D)$, and $\overline{D} = 7.59 > 7 = \text{med}(D)$, all support our contention that the mileage difference distribution is slightly skewed to the right. This distribution has a single peak (mode) at 7 indicating that for these subcompact car models it is most common for the highway mileage value to exceed the city mileage value by 7 mpg. For the majority of these subcompact car models the mileage difference is fairly close to 7 mpg. However, there are a few car models for which this mileage difference is a good bit larger. For example, the car model with the largest highway minus city mileage difference is the Saturn SC model without DOHC for which the mileage difference is 12 mpg.

Table 6. Model year 2000 subcompact car EPA mileage differences (highway - city,) excluding the five unusual car models.

6	5	9	8	11	9	10	10	7	6
8	9	7	8	5	5	5	5	9	7
7	4	7	8	7	8	8	7	8	11
9	10	10	7	9	12	11	6	7	7
7	6	7	7	6	9	7	6	7	6
7									

Figure 4. Stem and leaf histogram for subcompact car EPA mileage differences (highway - city), excluding the five unusual car models.

The stem represents ones and the leaf represents tenths (mpg).

4	0
5	00000
6	0000000
7	0000000000000000
8	0000000
9	0000000
10	0000
11	000
12	0

Table 7. Subcompact car EPA mileage difference (highway - city) summary statistics, excluding the five unusual car models.

min =	4	Q_1−min =	2
Q_1 =	6	med−Q_1 =	1
med =	7	Q_3−med =	2
Q_3 =	9	max −Q_3 =	3
max =	12		
range =	8	mean =	7.59
IQR =	3	std dev =	1.80

Using the mean mileage difference $\overline{D} = 7.59$, we conclude that, on the average, the highway mileage of a subcompact car model is 7.59 mpg larger than its city mileage. Based on the median mileage difference $\text{med}(D) = 7$, we would conclude that the highway mileage of a subcompact car model is 7 mpg larger than its city mileage; with half of the models having a difference less than 7 and half having a difference larger than 7.

In this example, the difference between the highway and city mileage means, $\overline{H} - \overline{C} = 31.12 - 23.53 = 7.59$, is equal to the mean mileage difference, $\overline{D} = 7.59$; and the difference between the highway and city mileage medians, $\text{med}(H) - \text{med}(C) = 30 - 23 = 7$, is equal to the median mileage difference, $\text{med}(D) = 7$. In paired data situations like this the difference of the two means is always equal to the mean of the differences. On the other hand, the difference between the two medians does not always equal the median of the differences.

It is interesting to note that the five car models which we excluded as outliers due to their unusually low city and highway mileage values would not be unusual in terms of their mileage differences (one 6 and four 5's).

3.2 Modified box plots.

In this section we will consider a modified box plot designed to provide more information about the tails of the distribution. In the modified box plot a more complex method, which provides an indication of extreme observations, is used to construct the whiskers.

The simple box plot defined in Section 3.1 has a box, which extends from the first quartile Q_1 to the third quartile Q_3 divided into two parts by a line at the median, representing the middle of the distribution, and two whiskers, extending from the ends of the box to the most extreme values (the minimum and the maximum), representing the tails of the distribution.

Observations located near the ends of a distribution are said to be extreme. The whiskers in a simple box plot indicate the range of the lower and upper tails and the locations of the most extreme values but do not provide details about the behavior of observations near the extremes of the distribution. An extreme observation which is widely separated from the majority of the observations is said to be an **outlier**. Outliers deserve special consideration, since they may represent interesting exceptional cases or they may represent errors made in recording the data. Note that, depending on the spacing of the observations, extreme observations may or may not be considered outliers. Outliers are easy to spot in a stem and leaf histogram but not in a simple box plot.

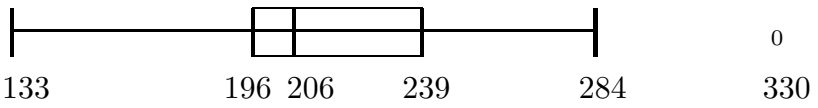
Before we can construct a modified box plot we need to quantify what we mean by an extreme observation. We will use multiples of the interquartile range (IQR) to distinguish between two types of extreme observations. First notice that the IQR is the range of the middle half of the data and the length of the box in the box plot. An observation which is much more than the IQR below the first quartile Q_1 or much more than the IQR above the third quartile Q_3 might reasonably be classified as an extreme observation. We will classify observations which are more than $1.5 \times \text{IQR}$ but less than $3 \times \text{IQR}$ below the first quartile or above the third quartile as somewhat extreme. That is, an observation between $Q_1 - 3 \times \text{IQR}$ and $Q_1 - 1.5 \times \text{IQR}$ or between $Q_3 + 1.5 \times \text{IQR}$ and $Q_3 + 3 \times \text{IQR}$ is said to be

somewhat extreme. We will classify observations which are more than $3 \times \text{IQR}$ below the first quartile or above the third quartile as very extreme. That is, an observation below $Q_1 - 3 \times \text{IQR}$ or above $Q_3 + 3 \times \text{IQR}$ is said to be very extreme.

The quantities $Q_1 - 1.5 \times \text{IQR}$ and $Q_3 + 1.5 \times \text{IQR}$ are known as the lower and upper inner fences, and the quantities $Q_1 - 3 \times \text{IQR}$ and $Q_3 + 3 \times \text{IQR}$ are known as the lower and upper outer fences. To construct a modified box plot we first find the five number summary values and the lower and upper fences. To construct the upper whisker we first draw a line from the upper end of the box, Q_3, extending to the largest observation which is less than the upper inner fence $Q_3 + 1.5 \times \text{IQR}$. We then indicate observations beyond the upper inner fence using two symbols; one symbol, such as a 0, is used for observations between the upper inner fence and the upper outer fence and another symbol, such as a *, is used for observations beyond the upper outer fence. The lower whisker is constructed in an analogous fashion.

Consider the urban cholesterol level distribution for the Guatemalan cholesterol example. In this example we have $Q_1 = 196$, $Q_3 = 239$, $\text{IQR} = 43$, $1.5 \times \text{IQR} = 64.5$, and $3 \times \text{IQR} = 129$. The inner fences are $196 - 64.5 = 131.5$ and $239 + 64.5 = 303.5$. There are no observed cholesterol levels below the lower inner fence but there is one observed cholesterol level, 330 mg/dL, above the upper inner fence. The largest cholesterol level between Q_3 and the upper inner fence is 284 mg/dL. The outer fences are 67 and 368. There are no observed cholesterol levels outside the outer fences. The modified box plot for this cholesterol level distribution is given in Figure 5. In this example we would say that the one cholesterol level (330) marked as somewhat extreme is an outlier, since it is fairly widely separated from the other values.

Figure 5. Modified box plot for (all) urban cholesterol levels.



3.3 Numerical measures of relative position

There are many situations when we might wish to quantify the position of a particular value of a variable relative to a sample of values. For example, when presented with the results of a standardized test, we would like to know where our score stands relative to the scores of everyone else who took the test. We will discuss two different ways to quantify the relative position of a particular value of a variable.

The first measure of the relative position of a particular value X is the percentile rank of X which quantifies the location of X in an ordered listing of all of the values in the sample. The **percentile rank** of a particular value X is the percentage of the values in

the sample that are less than or equal to the particular value X. More specifically, if m of the n observed values in the sample are less than or equal to the particular value, then the percentile rank of the particular value is $(m/n)100\%$. Reports of scores on standardized tests often include the actual score and its percentile rank. The percentile rank of an individual's test score indicates how the individual performed on the test relative to the group by providing the percentage of the group that scored no higher than the individual.

Notice that the five number summary values, the minimum, Q_1, the median, Q_3, and the maximum, are the $0^{th}, 25^{th}, 50^{th}, 75^{th}$, and 100^{th} percentiles of the distribution. Therefore, the use of the five number summary values to describe a distribution is an example of the use of selected percentiles to describe a distribution.

Consider the relative standing, in the Guatemalan cholesterol example, of a hypothetical individual with a cholesterol level of 210 mg/dL. Using Table 1 or Figure 1 we find that: The percentile rank of 210 in the rural group is 91.84% ($45/49 = .9184$), since 45 of the 49 rural Guatemalans have cholesterol levels of 210 or less; and, the percentile rank of 210 in the urban group is 51.11% ($23/45 = .5111$), since 23 of the 45 urban Guatemalans have cholesterol levels of 210 or less. Almost all of the rural Guatemalans have cholesterol levels of 210 or less; thus it is clearly unusual for a rural Guatemalan to have a Cholesterol level which is higher than 210. On the other hand, roughly half of the urban Guatemalans have cholesterol levels of 210 or less.

We can also use this percentile rank idea to quantify the difference in location between these cholesterol level distributions. For example, 81.63% of the rural Guatemalans have cholesterol levels of 188 or less, while 80% of the urban Guatemalans have cholesterol levels above 188.

The second measure of the relative position of a particular value X is the Z–score of X which quantifies the location of X relative to the mean $\overline{X}$ of the sample in terms of the standard deviation S_X of the sample. Since the Z–score is based on $\overline{X}$ and S_X, the Z–score is only appropriate when $\overline{X}$ and S_X are appropriate measures of the center and variability in the sample, respectively. We will develop the Z–score in two stages.

First, we need a measure of the location of X relative to the center of the distribution as determined by the mean $\overline{X}$. The deviation, $X - \overline{X}$, of X from the mean $\overline{X}$ is such a measure. The deviation $X - \overline{X}$ is the signed distance from the particular value X to the mean $\overline{X}$. If $X - \overline{X}$ is negative, then X is below (smaller than) the mean. If $X - \overline{X}$ is positive, then X is above (larger than) the mean. In summary, the sign of the deviation $X - \overline{X}$ indicates the location of X relative to the mean $\overline{X}$; and the magnitude of the deviation $|X - \overline{X}|$ is the distance from X to the mean $\overline{X}$, measured in the units of measurement used for the observation X.

Second, we want a measure of the location of X relative to the mean $\overline{X}$ which takes the amount of variability in the data into account. We will obtain such a measure by

using the standard deviation S_X of the sample to standardize the deviation $X - \overline{X}$. Given a particular value X, the sample mean $\overline{X}$, and the sample standard deviation S_X, the **Z–score** corresponding to X is

$$Z = \frac{X - \overline{X}}{S_X}.$$

The sign of the Z–score indicates the location of X relative to the mean $\overline{X}$ and the magnitude of the Z–score is the distance from X to the mean $\overline{X}$ in terms of standard deviation units. For example, if $Z = 2$, then X is two standard deviation units above the mean ($X = \overline{X} + 2S_X$), and, if $Z = -2$, then X is two standard deviation units below the mean ($X = \overline{X} - 2S_X$).

Returning to the Guatemalan cholesterol example and the relative position of an individual with a cholesterol level of 210, let R denote the cholesterol level of a rural Guatemalan and let U denote the cholesterol level of an urban Guatemalan. The rural cholesterol mean is $\overline{R} = 157.02$ mg/dL and the rural cholesterol standard deviation is $S_R = 31.75$ mg/dL. The urban cholesterol mean is $\overline{U} = 216.87$ mg/dL and the urban cholesterol standard deviation is $S_U = 39.92$ mg/dL. The raw deviation of a cholesterol level of 210 from the rural mean is $210 - \overline{R} = 52.98$ mg/dL. Since this quantity is positive, we see that a cholesterol level of 210 mg/dL exceeds the rural mean by 52.98 mg/dL. The raw deviation of a cholesterol level of 210 from the urban mean is $210 - \overline{U} = -6.87$ mg/dL. Since this quantity is negative, we see that a cholesterol level of 210 mg/dL is 6.87 mg/dL below the urban mean.

Standardizing these raw deviations yields a Z–score of $52.98/31.75 = 1.67$ for a rural cholesterol level of 210 mg/dL and a Z–score of $-6.87/39.92 = -.17$ for an urban cholesterol level of 210 mg/dL. Notice that these Z–scores are unitless numbers (number of standard deviation units from the mean) which are directly comparable. Therefore, a rural cholesterol level of 210 mg/dL is 1.67 standard deviation units above the rural mean cholesterol level and an urban cholesterol level of 210 mg/dL is .17 standard deviation units below the urban mean cholesterol level. In terms of standard deviation units, we see that 210 mg/dL is about 10 times as far away from the mean cholesterol level for the rural group as it is for the urban group. In other words, when taking variability into account we find that it is much more unusual for a rural Guatemalan to have a cholesterol level of 210 than it is for an urban Guatemalan to have a cholesterol level of 210.

The remainder of this section is devoted to two interesting results which establish a connection between Z–scores and percentages. The first result, the $68\% - 95\% - 99.7\%$ rule, is an approximate rule not a mathematical fact. Strictly speaking, this rule only applies to distributions that are unimodal (single peaked), mound shaped, and symmetric. A formal statement of this rule is provided below.

The 68%-95%-99.7% rule. *For a distribution that is unimodal (has a single peak), mound shaped, and reasonably symmetric:*
i) Approximately 68% of the observed values will be within one standard deviation unit of the mean. That is, approximately 68% of the observed values will have a Z–score that is between -1 and 1.
ii) Approximately 95% of the observed values will be within two standard deviation units of the mean. That is, approximately 95% of the observed values will have a Z–score that is between -2 and 2.
iii) Approximately 99.7% of the observed values will be within three standard deviation units of the mean. That is, approximately 99.7% of the observed values will have a Z–score that is between -3 and 3. Notice that this indicates that almost all of the observed values will be within three standard deviations of the mean.

When it is applicable, the $68\% - 95\% - 99.7\%$ rule, can be used to determine the relative position of a particular value of a variable based on the corresponding Z–score. Notice that this rule indicates that a fairly large proportion (68%) of the sample will lie within one standard deviation of the mean; a very large proportion (95%) of the sample will lie within two standard deviations of the mean; and, almost all (99.7%) of the sample will lie within three standard deviations of the mean.

The rural and urban cholesterol distributions are both unimodal, mound shaped, and reasonably symmetric. For the rural group we find that 34 of the 49 cholesterol levels (69.39%) are within one standard deviation of the mean; 46 of the 49 cholesterol levels (93.88%) are within two standard deviation of the mean; and, all 49 cholesterol levels (100%) are within three standard deviation of the mean. For the urban group we find that 34 of the 45 cholesterol levels (75.56%) are within one standard deviation of the mean; 42 of the 45 cholesterol levels (93.33%) are within two standard deviation of the mean; and, all 45 cholesterol levels (100%) are within three standard deviation of the mean. Notice that the $68\% - 95\% - 99.7\%$ rule works better for the rural group cholesterol distribution, since it is more symmetric than the urban group cholesterol distribution. We would get better agreement of the urban group cholesterol levels with the $68\% - 95\% - 99.7\%$ rule if we excluded the outlier.

Example. Heights of adult males in the United Kingdom. The heights (in inches) of 8585 adult males born in the United Kingdom (including the whole of Ireland) are summarized in Table 8. This example is taken from Kendall and Stuart, *The Advanced Theory of Statistics, vol.1*, Griffin, (1977), 8. The data are from the *Final Report of the Anthropometric Committee to the British Association*, (1883), 256.

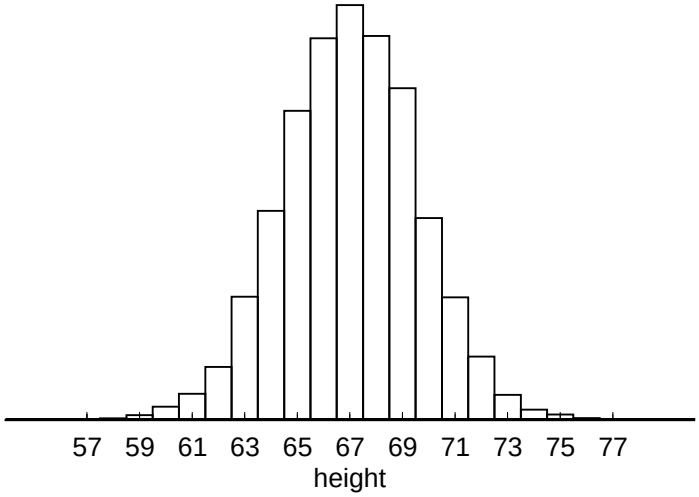
The histogram for the distribution of the 8585 heights of adult males for the United Kingdom height example in Figure 6 is unimodal, mound shaped, and symmetric. The sample mean height for this sample is 67.02 inches and the height standard deviation is

2.57 inches. For this example we find that 5835 of the 8585 heights (67.97%) are within one standard deviation of the mean height; 8307 of the 8585 heights (96.76%) are within two standard deviation of the mean height; and, 8542 of the 8585 heights (99.5%) are within three standard deviation of the mean height. Hence, the $68\% - 95\% - 99.7\%$ rule is quite accurate in its predictions for this UK height example.

Table 8. UK male heights.

height	frequency
57	2
58	4
59	14
60	41
61	83
62	169
63	394
64	669
65	990
66	1223
67	1329
68	1230
69	1063
70	646
71	392
72	202
73	79
74	32
75	16
76	5
77	2
total	8585

Figure 6. Histogram for UK heights.



The second result, Chebyshev's rule, is a mathematical fact that is true for any distribution. Unfortunately, the universal applicability of Chebyshev's rule forces its conclusions to be of more theoretical than practical interest. That is, the conclusions of Chebyshev's rule are valid for any distribution; but, they are often so imprecise that they are of limited practical use.

Chebyshev's rule. *For any distribution:*
i) At least 75% of the observed values will be within two standard deviation units of the mean. That is, at least 75% of the observed values will have a Z–score that is between -2 and 2.
ii) At least 89% of the observed values will be within three standard deviation units of the mean. That is, at least 89% of the observed values will have a Z–score that is between -3 and 3.
iii) In general, given a number $k > 1$*, at least* $[1 - (1/k^2)]100\%$ *of the observed values will be within k standard deviation units of the mean, i.e., at least this percentage of the observed values will have a Z–score that is between -k and k.*

3.4 Summary

In this chapter we introduced numerical summary values (statistics) and discussed the use of such statistics to quantify certain aspects of a distribution and to compare two distributions. Most of our attention focused on the shape of a distribution, the location of the distribution on the number line, and the amount of variability in the distribution. We began by defining the five number summary (minimum, Q_1, median, Q_3, maximum) which partitions the distribution into fourths. We then demonstrated how the five number summary and related statistics, such as the range and interquartile range, can be used

to summarize a distribution and to compare and contrast two distributions. A simple graphical representation of a distribution, the box plot, based on the five number summary was also introduced. We also defined the mean (a measure of location) and the standard deviation (a measure of variability).

Shape (skewness) Comparisons of the distances among the five number summary values can be used to assess and quantify skewness in a distribution as indicated below.

1. To quantify overall skewness in the distribution: compare (median – minimum) to (maximum – median).
2. To quantify skewness in the middle of the distribution: compare (median – Q_1) to (Q_3 – median).
3. To quantify skewness in the tails of the distribution: compare (Q_1 – minimum) to (maximum – Q_3).

Location The median and the mean are used to quantify the location of the center of a distribution on the number line. Recall that the median indicates the point which divides the distribution into halves (the histogram has equal area on each side of the median) while the mean indicates the point at which the distribution balances (the histogram has its center of gravity at the mean). If the distribution is symmetric, then the mean and the median are equal and either will suffice as a measure of the center of the distribution. If the distribution is heavily skewed, then the median is generally preferred over the mean as a measure of the center of the distribution. When comparing two distributions which have more or less the same shape either the median or the mean will suffice for comparing the locations of the distributions. But, when comparing distributions with different shapes the median is generally preferred over the mean for comparing the locations of the distributions.

Variability The range (maximum – minimum), interquartile range ($Q_3 - Q_1$), and standard deviation are used to quantify the variability in a distribution. For each of these statistics a larger value indicates more variability.

In Section 3.3 we discussed the use of percentile ranks and Z–scores to quantify the relative position of a particular value relative to the distribution of a sample. These ideas and in particular the Z–score, which indicates the location of a value relative to the mean in terms of standard deviation units, will reappear when we discuss inference in later chapters.

3.5 Exercises

For each of the examples in Section 1.2 which involve quantitative variables:

1. Determine the following summary statistics: the five number summary (minimum, Q_1, median, Q_3, maximum), the range, the interquartile range, the mean, and the standard deviation. (See the notes below for the examples with two or more groups.)

2. Discuss the distribution of the variable(s) using the summary statistics of question 1 to lend quantitative support to your discussion.

Notes:

For the examples with two or more groups (DiMaggio and Mantle, gear tooth strength), find the indicated summary statistics and compare and contrast the distributions of the variable for the two (or more) groups.

For the paired data examples (wooly–bear cocoons, homophone confusions) find the differences for each pair of data values and find the indicated summary statistics and describe the distribution of the differences.