2 Summarizing and Displaying Data

SW Chapter 2

Suppose we have a collection of n individuals, and we measure each individuals response on one quantitative characteristic, say height, weight, or systolic blood pressure. For notational simplicity, the collected measurements are denoted by $Y_1,\ Y_2, ..., Y_n$, where n is the **sample size**. The order in which the measurements are assigned to the place-holders $Y_1,\ Y_2, ..., Y_n$ is irrelevant.

Among the numerical summary measures computed by Minitab are the **sample mean** $\overline{Y}$ and the **sample standard deviation** s. The sample mean is a measure of **central location**, or a measure of a "typical value" for the data set. The standard deviation is a measure of **spread** in the data set. These summary statistics should be familiar to you. Let us consider a simple example to refresh your memory on how to compute them. Suppose we have a sample of $n = 8$ children with weights (in pounds): 5, 9, 12, 30, 14, 18, 32, 40. Then

$$\begin{aligned} \overline{Y} \;=\; \frac{\sum_i Y_i}{n} \;&=\; \frac{Y_1 + Y_2 + \cdots + Y_n}{n} \\ &=\; \frac{5 + 9 + 12 + 30 + 14 + 18 + 32 + 40}{8} \;=\; \frac{160}{8} \;=\; 20. \end{aligned}$$

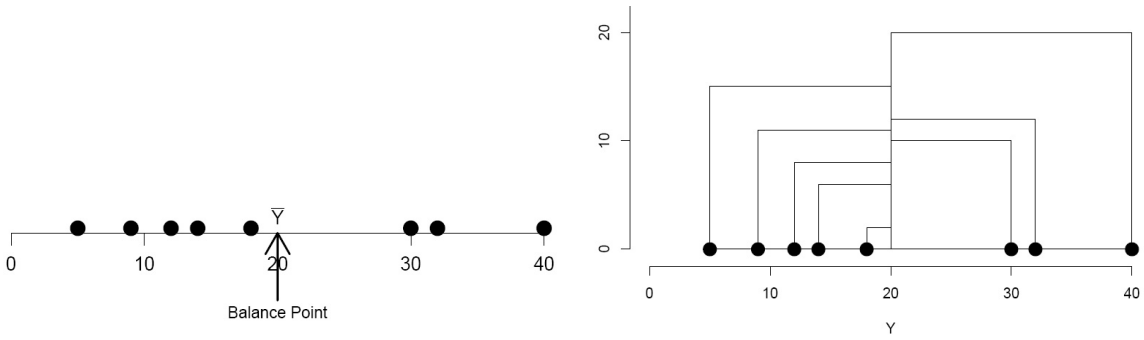
The sample standard deviation is the square root of the sample variance

$$\begin{aligned} s^2 \;=\; \frac{\sum_i (Y_i - \overline{Y})^2}{n-1} \;&=\; \frac{(Y_1 - \overline{Y})^2 + (Y_2 - \overline{Y})^2 + \cdots + (Y_k - \overline{Y})^2}{n-1} \\ &=\; \frac{(5-20)^2 + (9-20)^2 + \cdots + (40-20)^2}{7} = 156.3, \end{aligned}$$

that is, $s = \sqrt{s^2} = 12.5$. Summary statistics have well-defined units of measurement, for example, $\overline{Y} = 20lb,\ s^2 = 156.3lb^2$, and $s = 12.5lb$. The standard deviation is often used instead of s^2 as a measure of spread because s is measured in the same units as the data.

REMARK: If the divisor for s^2 was n instead of $n-1$, then the variance would be the average squared deviation observations are from the center of the data as measured by the mean.

The following graphs should help you to see some physical meaning of the sample mean and variance. If the data values were placed on a "massless" ruler, the balance point would be the mean (20). The variance is basically the "average" (remember n-1 instead of n) of the total areas of all the squares obtained when squares are formed by joining each value to the mean. In both cases think about the implication of unusual values (**outliers**). What happens to the balance point if the 40 were a 400 instead of a 40? What happens to the squares?



The **sample median** M is an alternative measure of central location. The measure of spread reported along with M is the **interquartile range**, $IQR = Q_3 - Q_1$, where Q_1 and Q_3 are the first and third quartiles of the data set, respectively. To calculate the median and interquartile range, order the data from lowest to highest values, all repeated values included. The ordered weights are

$$5 \;\; 9 \;\; 12 \;\; 14 \;\; 18 \;\; 30 \;\; 32 \;\; 40.$$

The median M is the value located at the half-way point of the ordered string. There is an even number of observations, so M is defined to be half-way between the two middle values, 14 and 18. That is, $M = .5(14 + 18) = 16lb$. To get the quartiles, break the data into the lower half: 5 9 12 14, and the upper half: 18 30 32 and 40. Then

$$Q_1 = \text{first quartile} = \text{median of lower half of data} = .5(9{+}12){=}10.5lb,$$

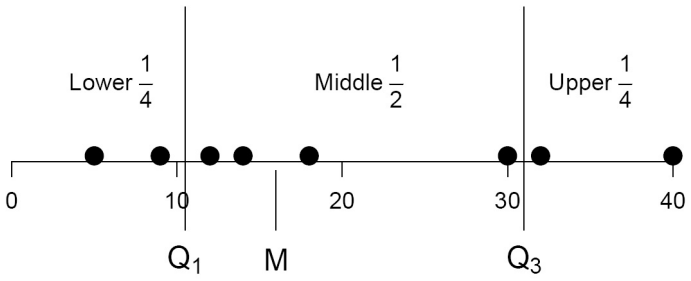
and

$$Q_3 = \text{third quartile} = \text{median of upper half of data} = .5(30{+}32) = 31lb.$$

The interquartile range is

$$IQR = Q_3 - Q_1 = 31 - 10.5 = 20.5lb.$$

The quartiles, with M being the second quartile, break the data set roughly into fourths. The first quartile is also called the 25^{th} percentile, whereas the median and third quartiles are the 50^{th} and 75^{th} percentiles, respectively.. The IQR is the **range** for the middle half of the data.



Suppose we omit the largest observation from the weight data:

5 9 12 14 18 30 32.

How do M and IQR change? With an odd number of observations, there is a unique middle observation in the ordered string which is M. Here $M = 14lb$. It is unclear which half the median should fall into, so M is placed into both the lower and upper halves of the data. The lower half is 5 9 12 14, and the upper half is 14 18 30 32. With this convention, $Q_1 = .5(9 + 12) = 10.5$ and $Q_3 = .5(18 + 30) = 24$, giving $IQR = 24 - 10.5 = 13.5(lb)$.

If you look at the data set with all eight observations, there actually are many numbers that split the data set in half, so the median is not uniquely defined, although "everybody" agrees to use the average of the two middle values. With quartiles there is the same ambiguity but no such universal agreement on what to do about it, however, so Minitab will give slightly different values for Q_1 and Q_3 than we just calculated, and other packages will report even different values. This has no practical implication (all the values are "correct") but it can appear confusing.

Example

The data given below are the head breadths in mm for a sample of 18 modern Englishmen, with numerical summaries generated by Minitab.

COMMENTS:

1. The data were entered into column 1 (C1) of the worksheet and labelled **English HW**.
2. The data are displayed via: Data > Display Data.
3. Summary statistics are obtained via: Stat > Basic Statistics > Display Descriptive Statistics

```
Data Display

English HW
   141   148   132   138   154   142   150   146   155   158   150   140   147
   148   144   150   149   145

Descriptive Statistics: English HW

Variable     N  N*    Mean  SE Mean  StDev  Minimum      Q1  Median      Q3
English HW  18   0  146.50     1.50   6.38   132.00  141.75  147.50  150.00

Variable    Maximum
English HW   158.00
```

Minitab does not give IQR but it is easily obtained: $IQR = 150 - 141.75 = 8.25mm$. $N*$ is the number of missing values (none here), and SE Mean is the standard error of the sample mean, $SE_{\overline{Y}} = s/\sqrt{n}$. The standard error is a measure of the precision of the sample mean $\overline{Y}$.

Graphical Summaries

There are four graphical summaries of primary interest: the **dotplot**, the **histogram**, the **stem and leaf** display, and the **boxplot**. There are more possible, including the individual values plots we looked at in the *Meet Minitab* tutorial, but these four generally are the most useful.

Minitab used to generate both **character** plots, which are viewed in the **session window**, and **high quality** plots, which are displayed in pop-up graphics windows, for just about everything.

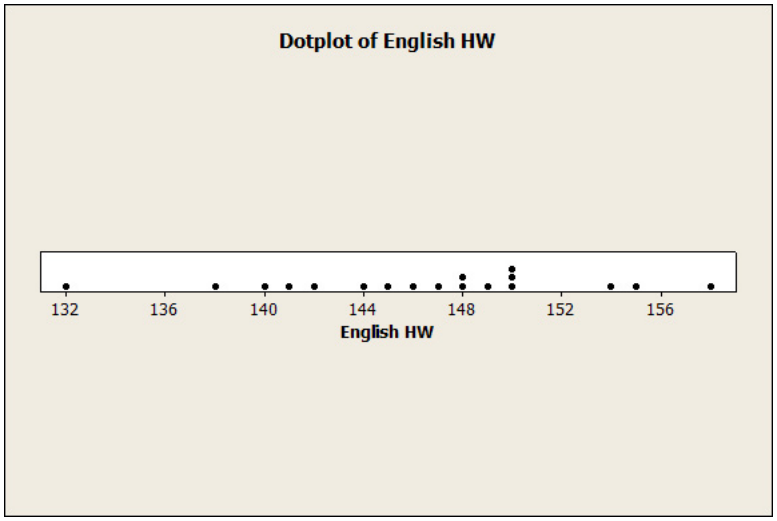
The only remaining character plot obtained by default is the stem and leaf display. Character plots use typewriter characters to make graphs, and can be convenient for some simple displays, but require use of fixed fonts (like Courier) when copied to a word processing program or they get distorted. About the only reason for using character plots any more is if you are using a really old printer such as a dot matrix printer that cannot easily print high resolution graphics. Minitab makes you jump some hoops to get those old-fashioned plots now – if you need them I can help you configure Minitab to allow them.

Most of the plots we will use are created directly from the Graph menu. The plots can be customized. Make liberal use of the on-line help for learning how to customize them. Plots can also be generated along with many statistical analyses, a point that we will return to repeatedly.

Dotplots

The **dotplot** breaks the range of data into many small equal width intervals, and counts the number of observations in each interval. The interval count is superimposed on the number line at the interval midpoint as a series of dots, usually one for each observation. In the head breadth data, the intervals are centered at integer values, so the display gives the number of observations at each distinct observed head breadth.

A dotplot of the head breadth data is given below.



Histogram and Stem and Leaf

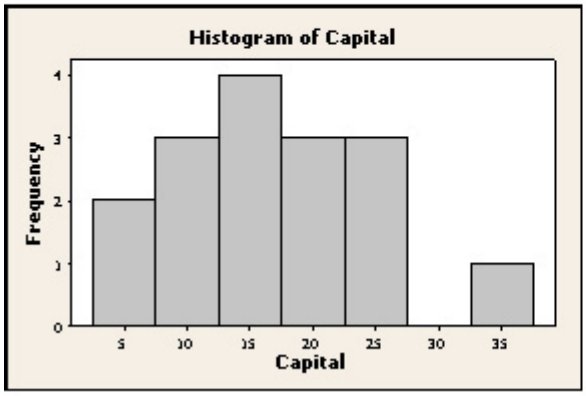
The **histogram** and **stem and leaf** displays are similar to the dotplot, but break the range of data into a smaller number of equal width intervals. This produces better graphical information about the observed distribution, i.e. it better highlights where data values cluster. The histogram can use arbitrary intervals, whereas the intervals for the stem and leaf display use the base 10 number system. There is more arbitrariness to histograms than to stem and leaf displays, so histograms can sometimes be regarded a bit suspiciously.

Minitab's "help" entry for Histograms follows. It is useful to examine these help entries in order to be clear on the terminology the software package uses. Note the term *bin* below.

Histograms

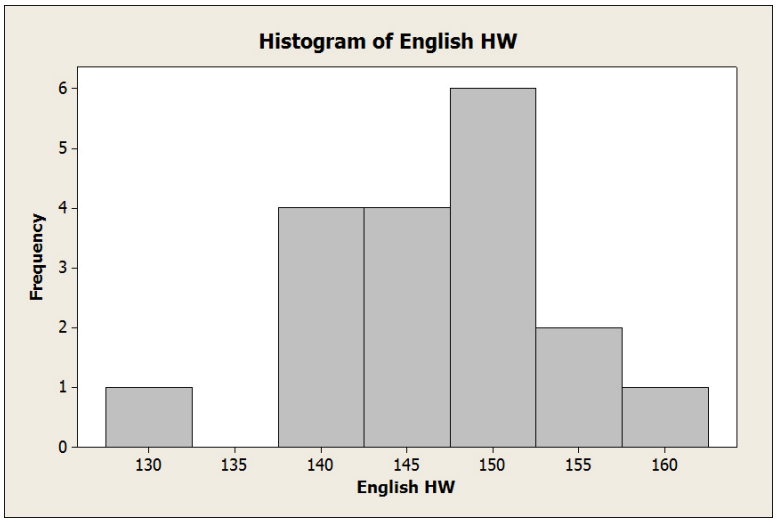
Graph > Histogram

Use to examine the shape and spread of sample data. Histograms divide sample values into many intervals called bins. Bars represent the number of observations falling within each bin (its frequency). In the histogram below, for example, there are two observations with values between 2.5 and 7.5, three observations with values between 7.5 and 12.5, and so on.



Observations that fall exactly on an interval boundary are included in the interval to the right (or left, if the last bin).

The Minitab histogram of head breadth data is below. The bins have width 5 and are centered at the values 130, 135, etc. The actual bins are 127.5-132.5, 132.5-137.5, 137.5-142.5, etc.



Most Minitab graphical commands allow you to modify the graphical display. For example, with the histogram you might wish to use different midpoints or interval widths. I will let you explore the possibilities.

A **stem and leaf** display defines intervals for a grouped frequency distribution using the base 10 number system. Intervals are generated by selecting an appropriate number of lead digits for

the data values to be the stem. The remaining digits comprise the leaf. Minitab's description (from "help") follows:

Stem-and-Leaf

Graph > Stem-and-Leaf
Stat > EDA > Stem-and-Leaf
***Character Graphs* > Stem-and-Leaf**

Use to examine the shape and spread of sample data. Minitab displays a stem-and-leaf plot in the Session window. The plot is similar to a histogram on its side, however, instead of bars, digits from the actual data values indicate the frequency of each bin (row).

Below is a stem-and-leaf plot for a data set with the following five values: 3, 4, 8, 8, and 10.

```
Stem-and-leaf of C1   N   = 5
Leaf Unit = 1.0

 1    0  3
 2    0  4
 2    0
(2)   0  88
 1    1  0
```

The display has three columns:

- The leaves (right) – Each value in the leaf column represents a digit from one observation. The "leaf unit" (declared above the plot) specifies which digit is used. In the example, the leaf unit is 1.0. Thus, the leaf value for an observation of 8 is 8 while the leaf value for an observation of 10 is 0.
- The stem (middle) – The stem value represents the digit immediately to the left of the leaf digit. In the example, the stem value of 0 indicates that the leaves in that row are from observations with values greater than or equal to zero, but less than 10. The stem value of 1 indicates observations greater than or equal to 10, but less than 20.
- Counts (left) – If the median value for the sample is included in a row, the count for that row is enclosed in parentheses. The values for rows above and below the median are cumulative. The count for a row above the median represents the total count for that row and the rows above it. The value for a row below the median represents the total count for that row and the rows below it.

  In the example, the median for the sample is 8, so the count for the fourth row is enclosed in parentheses. The count for the second row represents the total number of observations in the first two rows.

Look carefully at the description above. If I used numbers 30, 40, 80, 80, and 100 instead of those in the example, the plot would look just the same with one exception – what is it? Make sure you can tell what the number is from its entry in the display. The display almost looks upside down, since larger numbers are on the bottom rather than the top. It is done this way so that if rotated 90 degrees counter-clockwise it *is* a histogram.

The default stem and leaf display for the head breadth data is given below. The second and third columns of the Minitab display give the stems and leaves, respectively. The data have three digits. The first two comprise the stem. The last digit is the leaf. Thus, a head breadth of 154 has a stem of 15 and leaf of 4. The possible stems are 13, 14, and 15, whereas the possible leaves are the integers from 0 to 9. Each stem occurs twice here. The first (top) occurrence of a stem value only holds leaves 0 through 4. The second occurrence holds leaves 5 through 9. The display is generated by placing the leaf value for each observation on the appropriate stem line. For example, the top 14 stem holds data values between 140 and 144.99. The stems on this line in the display tell us that four observations fall in this range: 140, 141, 142 and 144. Note that this stem and leaf display is an elaborate histogram with intervals of width 5. An advantage of the stem and leaf display over the histogram is that the original data values can essentially be recovered from the display.

```
Stem-and-Leaf Display: English HW

Stem-and-leaf of English HW  N  = 18
Leaf Unit = 1.0

 1   13  2
 2   13  8
 6   14  0124
(6)  14  567889
 6   15  0004
 2   15  58
```

The description for the stem-and-leaf display given above is not entirely complete for Minitab displays. In Minitab, the data values are always *truncated* so that a leaf has one digit. The leaf unit (1 for the head breadth data) tells us the degree of round-off. This will become clearer in the next example. One question that I have not yet answered is: What does the leftmost column in the display tell us? (read the help entry).

The various stem and leaf dialog boxes allow user specified increments. The increments can only be numbers evenly divisible into powers of 10, for example, 1, 2, 5, 10, 20, 50, 100, and so on (Why?). The displays given below were generated using interval widths (increments) of 2 and 10, respectively.

In the second display, each of the three stems *may* occur five times. The first copy holds leaves 0 and 1. The second copy holds leaves 2 and 3, and so on. Of the three displays, the default stem and leaf gives the most informative summary. The displays below have too few and too many stems, respectively. Our goal is to see clustering and shape of distributions of numbers.

```
Stem-and-Leaf Display: English HW

Stem-and-leaf of English HW  N  = 18     <- used increment of 2
Leaf Unit = 1.0

    1   13 2
    1   13
    1   13
    2   13 8
    4   14 01
    5   14 2
    7   14 45
    9   14 67
    9   14 889
    6   15 000              <- holds leaves 0 & 1
    3   15                                2 & 3
    3   15 45                             4 & 5
    1   15                                6 & 7
    1   15 8                              8 & 9



Stem-and-leaf of English HW  N  = 18     <- used increment of 10
Leaf Unit = 1.0

    2   13 28
  (10)  14 0124567889
    6   15 000458
```

Boxplots

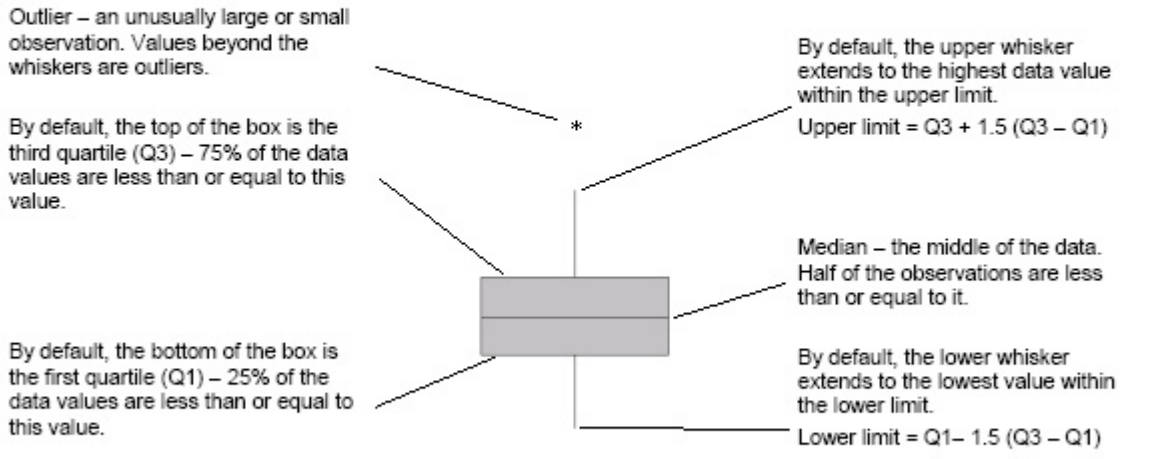
The **boxplot** breaks up the range of data values into regions about the center of the data, measured by the median. The boxplot highlights **outliers** and provides a visual means to assess **"normality"**. The following help entry outlines the construction of the boxplot, given the placement of data values on the axis.

Boxplots

Graph > Boxplot

Stat > EDA > Boxplot

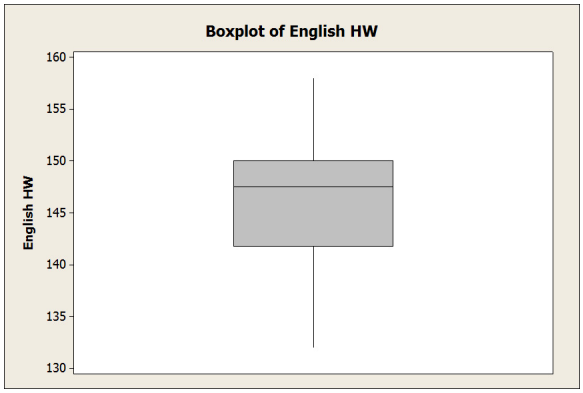
Use boxplots (also called box-and-whisker plots) to assess and compare sample distributions. The figure below illustrates the components of a default boxplot.



Note By default, Minitab uses the quartile method for calculating box endpoints. To change the method for a specific graph to hinge or percentile, use Editor > Edit Interquartile Range Box > Options. To change the method for all future boxplots, use Tools > Options > Individual Graphs > Boxplots.

The endpoints of the box are placed at the locations of the first and third quartiles. The location of the median is identified by the line in the box. The whiskers extend to the data points closest to but not on or outside the outlier fences, which are $1.5IQR$ from the quartiles. Outliers are any values on or outside the outlier fences.

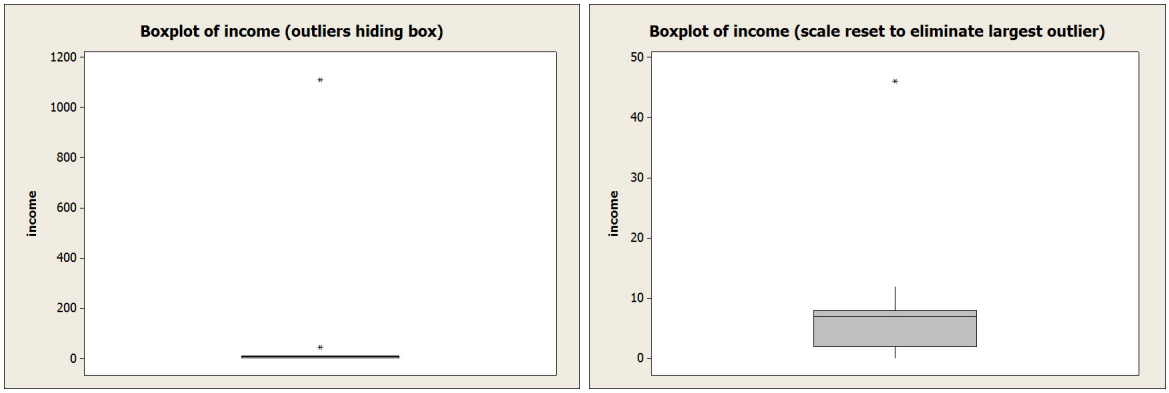
The boxplot for the head breadth data is given below. There are a lot of options that allow you to clutter the boxplot with additional information. Just use the default settings. We want to see relative location of data (the median line), an idea of spread of data (IQR, the length of the box), shape of the data (relative distances of components from each other – to be covered later), and identify outliers (if present). The default boxplot has all these components.



Example

The data printed below in Minitab are the incomes in 1000 dollar units for a sample of 12 retired couples. Numerical and graphical summaries are given. There are two stem and leaf displays provided. The first is the default display. For the second, I specified in the dialog box that the outliers be trimmed. Similarly, the first boxplot is the default display, while the second was obtained by specifying in the dialog box that rows with incomes larger than 50 be excluded.

```
income

   7   1110   7   5   8  12   0   5   2   2  46   7


Descriptive Statistics: income

Variable   N  N*   Mean  SE Mean  StDev  Minimum   Q1  Median    Q3  Maximum
income    12   0  100.9     91.8  318.0      0.0  2.8     7.0  11.0   1110.0

Stem-and-Leaf Display: income

Stem-and-leaf of income    N  = 12
Leaf Unit = 100                          <- rounds off to leaf unit precision

  (11)   0 00000000000                     leaf unit = 100 so stem = 1 leaf = 1
    1    0                                 is 1100 to nearest 100
    1    0
    1    0
    1    0
    1    1 1


Stem-and-Leaf Display: income (Trimming off outliers)

Stem-and-leaf of income    N  = 12
Leaf Unit = 1.0

    1    0 0
    3    0 22
    5    0 55
   (3)   0 777
    4    0 8
    3    1
    3    1 2

        HI    46, 1110,                  <- high outliers
```



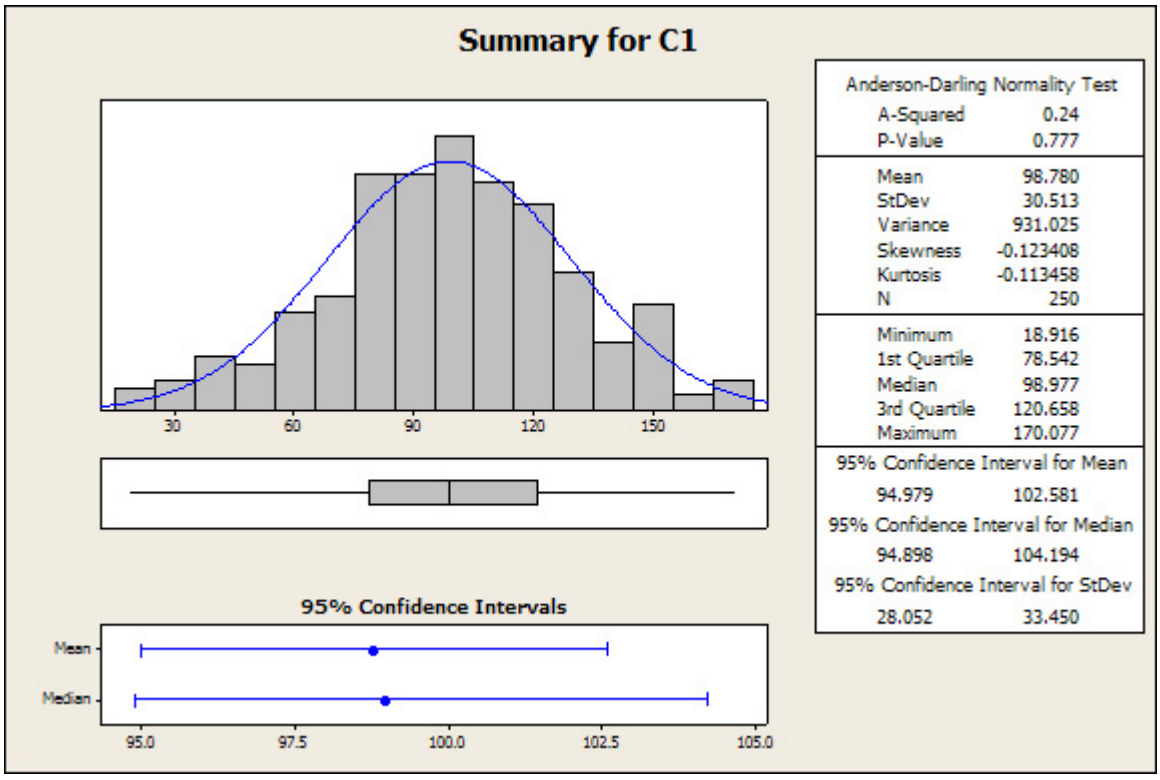
Interpretation of Graphical Displays for Numerical Data

In many studies, the data are viewed as a subset or **sample** from a larger collection of observations or individuals under study, called the **population**. A primary goal of many statistical analyses is to generalize the information in the sample to **infer** something about the population. For this generalization to be possible, the sample must reflect the basic patterns of the population. There are several ways to collect data to ensure that the sample reflects the basic properties of the population, but the simplest approach, by far, is to take a random or "representative" sample from the population. A **random sample** has the property that every possible sample of a given size has the same chance of being the sample eventually selected. Random sampling eliminates any systematic biases associated with the selected observations, so the information in the sample should accurately reflect features of the population. The process of sampling introduces random variation or random errors associated with summaries. Statistical tools are used to calibrate the size of the errors.

Whether we are looking at a histogram (or stem and leaf, or dotplot) from a sample, or are conceptualizing the histogram generated by the population data, we can imagine approximating the "envelope" around the display with a smooth curve. The smooth curve that approximates the population histogram is called the **population frequency curve**. Statistical methods for inference about a population usually make assumptions about the shape of the population frequency curve. A common assumption is that the population has a normal frequency curve. In practice, the observed data are used to assess the reasonableness of this assumption. In particular, a sample display should resemble a population display, provided the collected data are a random or representative sample from the population. Several common shapes for frequency distributions are given below, along with the statistical terms used to describe them.

The first display is **unimodal** (one peak), **symmetric** and **bell-shaped**. This is the prototypical normal curve. The boxplot (laid on its side for this display) shows strong evidence of symmetry: the median is about halfway between the first and third quartiles, and the tail lengths are roughly equal. The boxplot is calibrated in such a way that 7 of every 1000 observations are outliers (more than $1.5(Q_3 - Q_1)$ from the quartiles) in samples from a population with a normal frequency curve. Only 2 out of every 1 million observations are extreme outliers (more than $3(Q_3 - Q_1)$ from the quartiles). We do not have any outliers here out of 250 observations, but we certainly could have some without indicating nonnormality. If a sample of 30 observations contains 4 outliers, two of

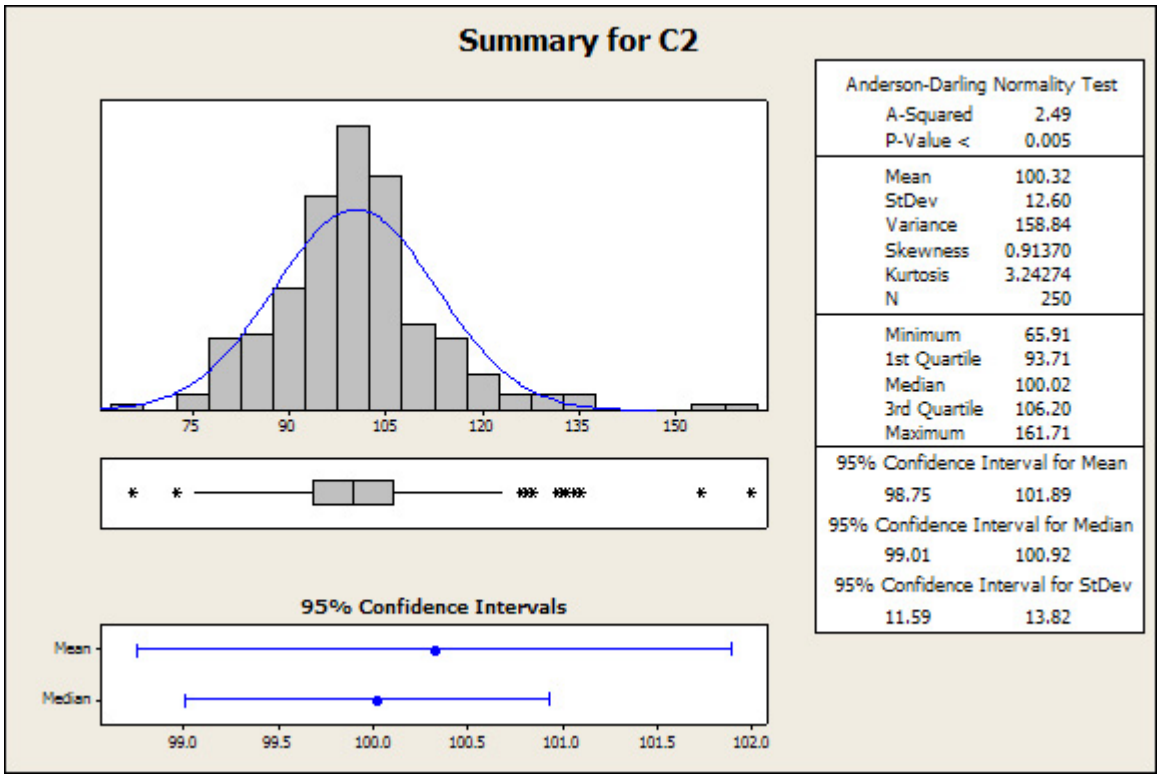
which are extreme, would it be reasonable to assume the population from which the data were collected has a normal frequency curve? Probably not.




```
Stem-and-Leaf Display: C1

Stem-and-leaf of C1  N  = 250
Leaf Unit = 1.0

 1     1   8
 5     2   1378
 9     3   3379
 17    4   11223567
 25    5   13455789
 38    6   2222444458899
 65    7   1122222334555556677778888889
 98    8   000011112233445555666666678888889
(32)   9   11112223344455555555667888899999
 120   10  000123334444444455566667788889
 90    11  001111122233344455566668899
 64    12  00011111122234444555555689
 39    13  001112344466779
 24    14  011366677778
 12    15  001133
 6     16  04669
 1     17  0
```

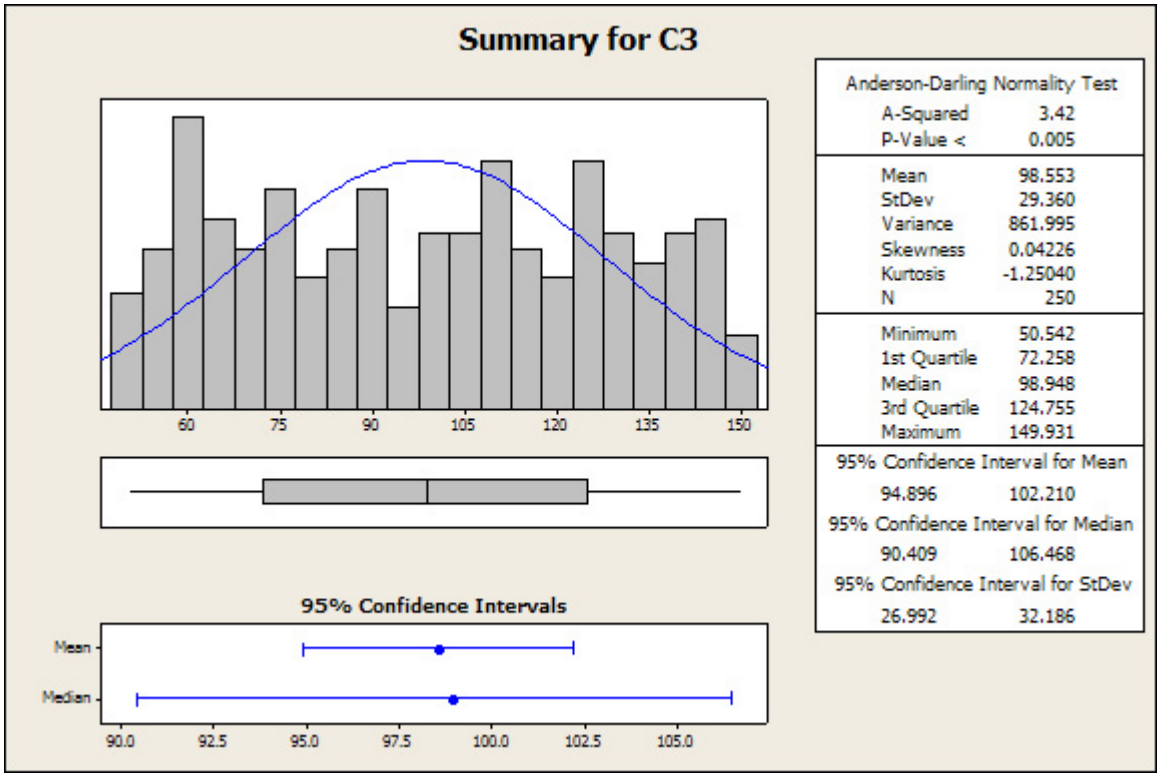
The boxplot is better at highlighting outliers than are other displays. The histogram and stem and leaf displays below appear to have the same basic shape as a normal curve (unimodal, symmetric). However, the boxplot shows that we have a dozen outliers in a sample of 250 observations. We would only expect about two outliers in 250 observations when sampling from a population

with a normal frequency curve. The frequency curve is best described as unimodal, symmetric, and **heavy-tailed**.




```
Stem-and-Leaf Display: C2

Stem-and-leaf of C2  N  = 250
Leaf Unit = 1.0

 1    6   5
 11   7   2578899999
 45   8   000112223333333445677777888889999
 124  9   00000011122222233333334444445555555566666666666677777777888888889+
(84)  10  000000000000000000111111111122222233333333344444444445555556666+
 42   11  00000011122222333345556689
 16   12  000113567
 7    13  12345
 2    14
 2    15  3
 1    16  1
```

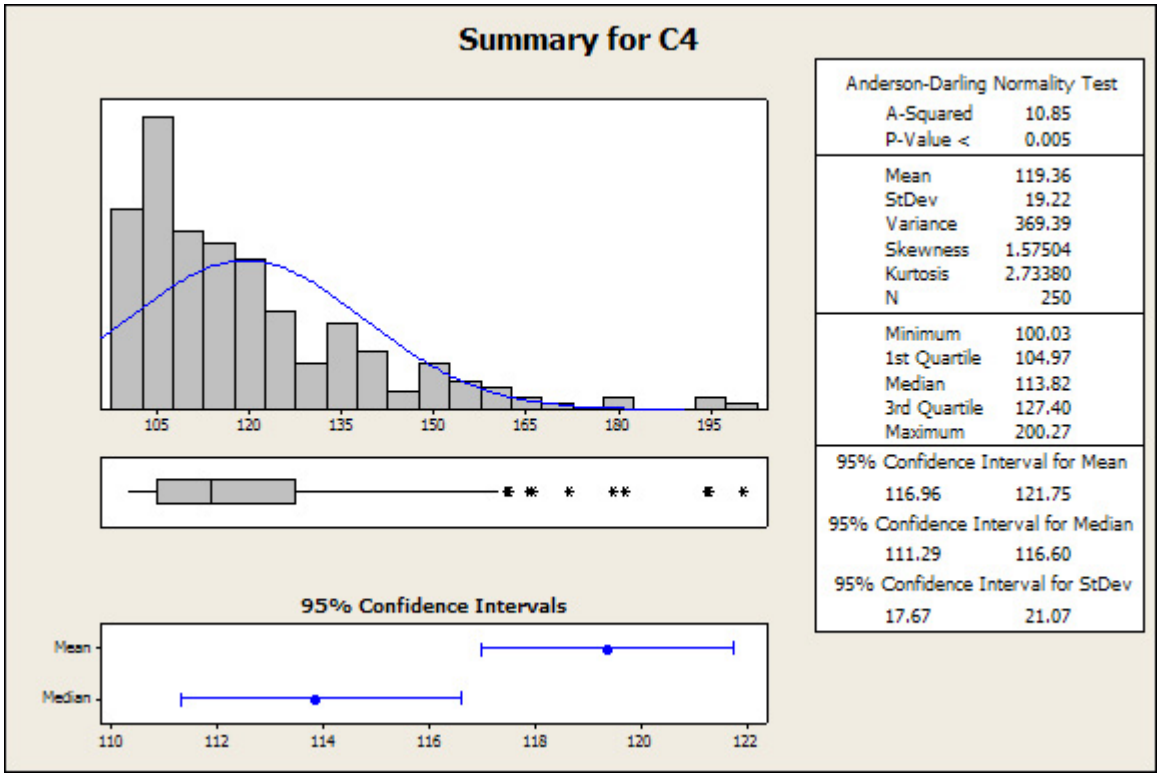
Not all symmetric distributions are mound-shaped, as the display below suggests. The boxplot shows symmetry, but the tails of the distribution are shorter (lighter) than in the normal distribution. Note that the distance between quartiles is roughly constant here.




```
Stem-and-Leaf Display: C3

Stem-and-leaf of C3  N  = 250
Leaf Unit = 1.0

 29    5   00111122334555666677777899999
 56    6   00000111122333456666677788 9
 82    7   00011123334445555666788 9
 108   8   11122222344556667778888889
(18)   9   001113334466788 89
 124   10  00000122345556666778889999 9
 97    11  00000111223334566668899 9
 73    12  00113334444445555666788 99
 48    13  00000111233344456777888999
 22    14  00012444555666667779 99
```

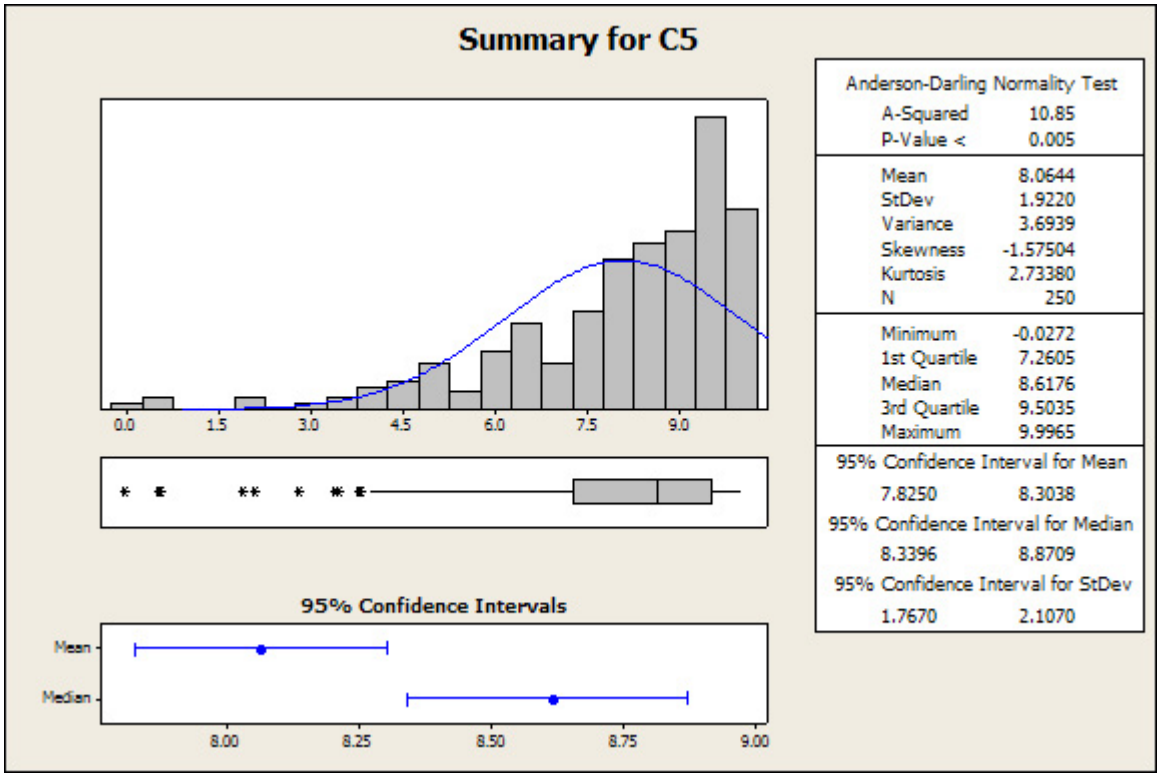
The mean and median are identical in a population with a (exact) symmetric frequency curve. The histogram and stem and leaf displays for a sample selected from a symmetric population will tend to be fairly symmetric. Further, the sample means and medians will likely be close.

The distribution below is unimodal, and asymmetric or **skewed**. The distribution is said to be **skewed to the right**, or upper end, because the right tail is much longer than the left tail. The boxplot also shows the skewness - the region between the minimum observation and the median contains half the data in less than 1/5 the range of values. In addition, the upper tail contains several outliers.



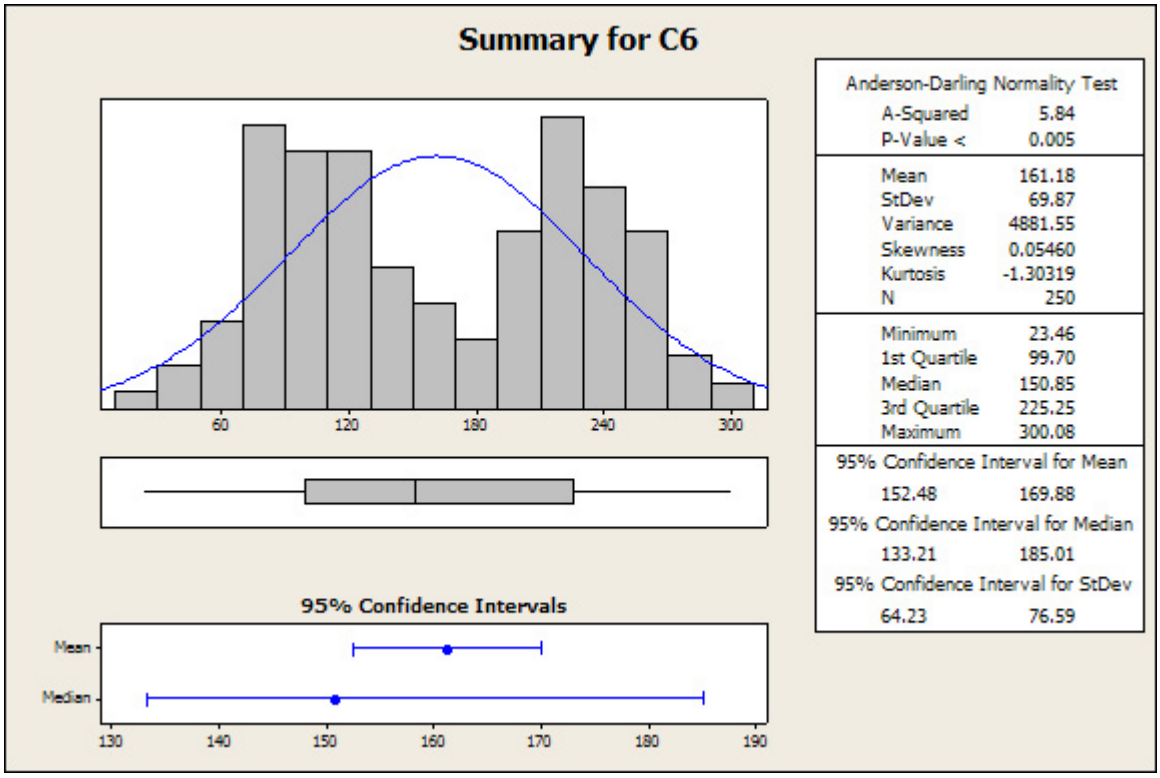

```
Stem-and-Leaf Display: C4

Stem-and-leaf of C4  N  = 250
Leaf Unit = 1.0

 105   10  000000000000000000000001111111111222222222333333333333334444444555+
(57)   11  000111112222222333333344444444455666667777777788888899999
 88    12  000000011223344445556666677788
 57    13  0112333445555666778889
 34    14  00112234489
 23    15  00111235556
 12    16  001256
 6     17  19
 4     18  1
 3     19  44
 1     20  0
```

The distribution below is unimodal and **skewed to the left**. The two examples show that extremely skewed distributions often contain outliers in the longer tail of the distribution.




```
Stem-and-Leaf Display: C5

Stem-and-leaf of C5  N  = 250
Leaf Unit = 0.10

 3    0   055
 4    1   8
 6    2   08
 12   3   347899
 23   4   34446788899
 34   5   01556778899
 57   6   01112233334444556667889
 88   7   11222233333444555566778899999999
(57)  8   00000111111222222223333344555555555666666677777778888899 9
 105  9   000000001111111111222222233333334444444444455555556666666666666677+
 0    10
```

Not all distributions are unimodal. The distribution below has two modes or peaks, and is said to be **bimodal**. Distributions with three or more peaks are called **multi-modal**.




```
Stem-and-Leaf Display: C6

Stem-and-leaf of C6  N  = 250
Leaf Unit = 10

  4     0  2233
 12     0  44455555
 32     0  66666777777777777777
 64     0  88888888888888888999999999999999
 95     1  00000000000001111111111111111
 115    1  22222222222333333333
(15)    1  444444445555555
 120    1  6666677777
 110    1  88899999999
 99     2  000000000001111111111111111
 71     2  22222222222222222333333333333333
 38     2  4444444445555555555555
```

The boxplot and histogram or stem and leaf display (or dotplot) are used **together** to describe the distribution. The boxplot does not provide information about modality - it only tells you about skewness and the presence of outliers.

As noted earlier, many statistical methods assume the population frequency curve is normal. Small deviations from normality usually do not dramatically influence the operating characteristics of these methods. We worry most when the deviations from normality are severe, such as extreme skewness or heavy tails containing multiple outliers.

Interpretations for examples

The head breadth sample is slightly skewed to the left, unimodal, and has no outliers. The distribution does not deviate substantially from normality. The various measures of central location ($\overline{Y} = 146.5,\ M = 147.5$) are close, which is common with fairly symmetric distributions containing no outliers.

The income sample is extremely skewed to the right due to the presence of two extreme outliers at 46 and 1110. A normality assumption here is unrealistic.

It is important to recognize the influence that outliers can have on the values of $\overline{Y}$ and s. The median and interquartile range are more robust (less sensitive) to the presence of outliers. For the income data $\overline{Y} = 100.9$ and $s = 318$, whereas $M = 7$ and $IQR = 8.3$. If we omit the two outliers, then $\overline{Y} = 5.5$ and $s = 3.8$, whereas $M = 6$ and $IQR = 5.25$.

The mean and median often have similar values in data sets without outliers, so it does not matter much which one is used as the "typical value". This issue is important, however, in data sets with extreme outliers. In such instances, the median is often more reasonable. For example, is $\overline{Y} = 100.9$ a reasonable measure for a typical income in this sample, given that the second largest income is only 46?

Minitab Discussion

We used to need to include a lot of pointers on how to use Minitab in these notes. I find the *Meet Minitab* tutorial to cover most of what is needed. The operational details you need to look up include

- Saving Data in Your Current Worksheet
- Importing Data into MINITAB
- Saving Session Window
- Cutting and Pasting Text from Session Window

I will demonstrate most of what you need, and I will be happy to answer questions. You will learn a lot more by using the software than by watching, however.