Lesson 1: Turning Data Into Information

Introduction

???

Learning objectives for this lesson

Upon completion of this lesson, you should be able to understand:

- the importance of graphing your data
- how to interpret the shape of a distribution
- what is a five-number summary and its interpretation
- the meaning of descriptive statistics
- what "average" means in statistics-speak
- the relationship between mean and median of a distribution
- some basic Minitab statistics and graphing methods

Four features to consider for quantitative variables are:

1. Shape
2. Center or Location
3. Spread (variability)
4. Outliers

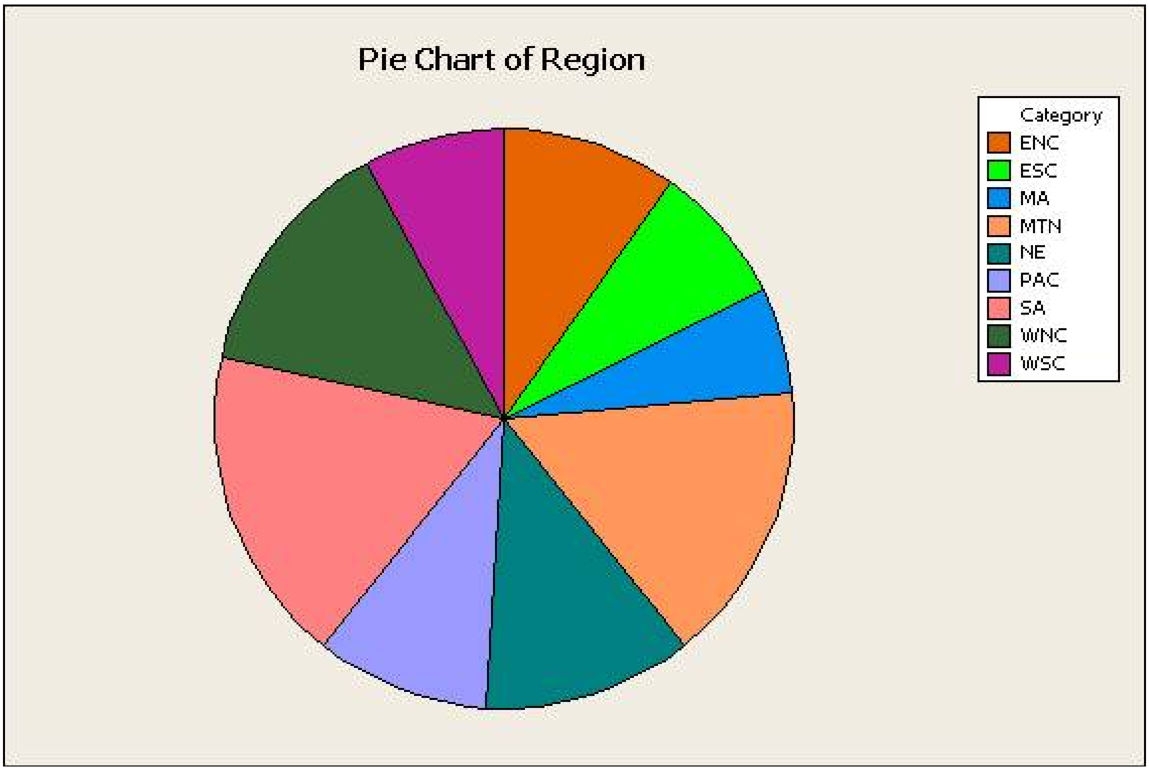
Displaying Distributions of Data with Graphs

The **distribution** of a variable shows its pattern of variation, as given by the values of the variables and their frequencies. The following Minitab data set, SAT_DATA.MTW, (data from College Board) contains the mean SAT scores for each of the 50 US states and Washington D.C., as well the participation rates and geographic region of each state. The data patterns however are not yet clear. To get an idea of the pattern of variation of a categorical variable such as region, we can display the information with a **bar graph** or **pie chart**.

To do this:

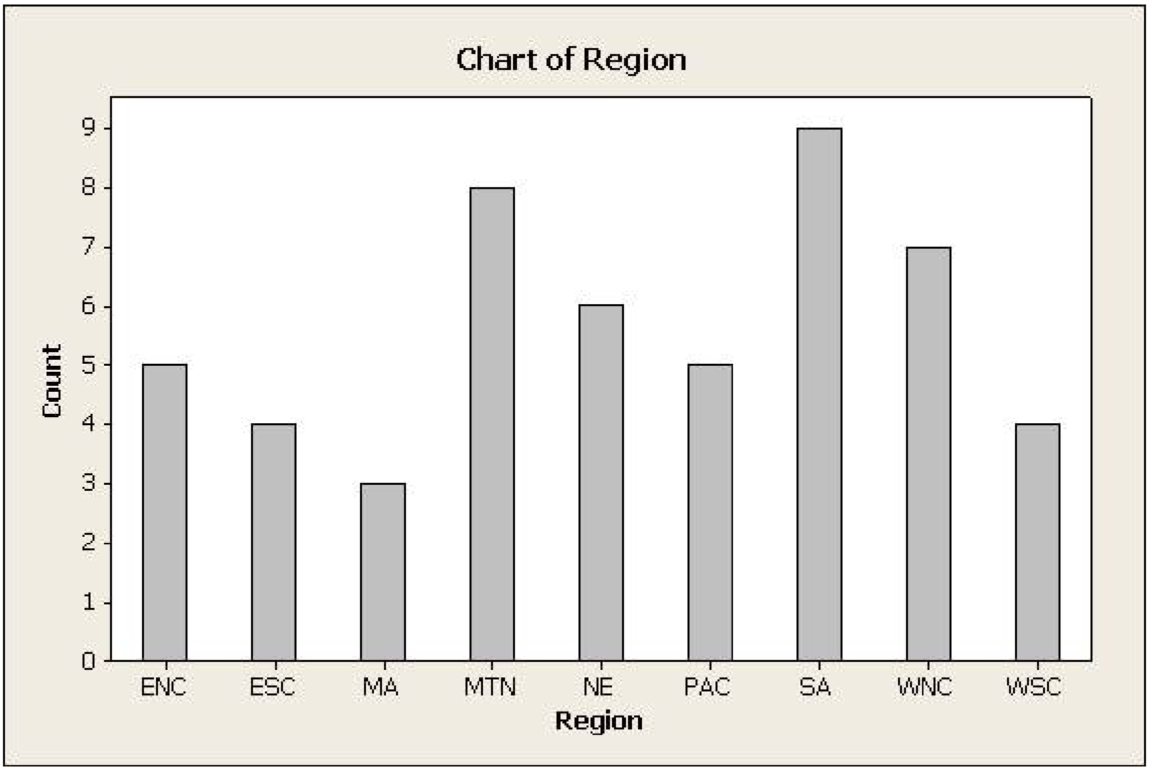
1. Open the data set
2. From the menu bar select Graph > Pie Chart
3. Click inside the window under Categorical Variables. This will bring up the list of categorical variables in the data set.
4. From the list of variables click on Region and then click the Select button. This should place the variable Region in the Categorical Variables window.
5. Click OK

This should result in the following pie chart:



In Minitab, if you place your mouse over any slice of the pie you will get the value of the overall percentage of the pie that region covers. For example, place your mouse over the blue colored slice (again this has to be done in Minitab not on the notes!) and you will see that for the Region MA (Mid Atlantic) 5.9% of the 50 states plus Washington D.C. fall into this category.

To produce a bar graph or bar chart, return to the menu bar in Minitab and from the Graph options select Bar Chart then Simple. The steps will proceed similar from Step 3 above. In the Minitab Bar Chart, however, placing your mouse over a bar produces the number within that category. For example, if you place your mouse over the region labeled MA (again this has to be done in Minitab not on the notes!) you will see that three (3) of the 50 states plus Washington D.C. are classified as Mid Atlantic. Note that 3/51 equals the 5.9% from the pie chart:



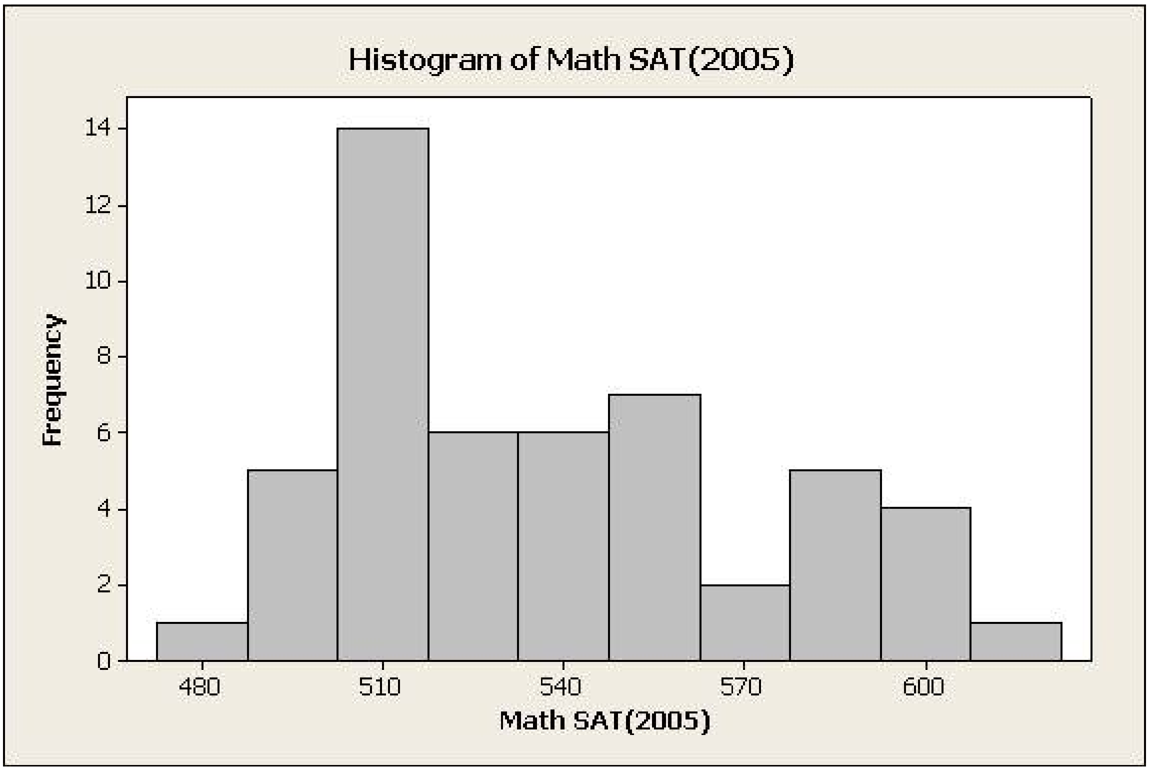
But what of variables that are **quantitative** such as math SAT or percentage taking the SAT? For these variables we should use **histograms**, **boxplots**, or **stem-and-leaf plots**. Stem-and-leaf plots are sometimes referred to as stemplots. Histograms differ from bar graphs in that the represent frequencies by area and not height. A good display will help to summarize a distribution by reporting the **center**, **spread**, and **shape** for that variable.

For now the goal is to summarize the distribution or pattern of variation of a *single* quantitative variable. To draw a histogram by hand we would:

1. Divide the range of data (range is from the smallest to largest value within the data for the variable of interest) into classes of equal width. For the math SAT scores the range is from 478 to 608. This range of 130 when divided by 10 comes to 13. For ease, Minitab goes to widths of 15.
2. Count the number of observations in each class. That is, count the number from 473 to 488, 489 to 504, etc.
3. Draw the histogram using the horizontal axis as the range of the data values and the vertical axis for the counts within class.

Minitab can also produce a histogram by:

1. Open the data set (Use this link data set if you do not have the data set open)
2. From the menu bar select Graph > Histogram
3. Select Simple and click OK
4. Click inside the window under Graph Variables. This will bring up the list of quantitative variables in the data set.
5. From the list of variables click on Math SAT(2005) and then click the Select button. This should place the variable Math SAT(2005) in the Graph Variables window.
6. Click OK



Again, when in Minitab, you can place your mouse over one of the bars and the number of observations (value) for that class and the length of that class (bin) will be displayed. For example, if you place your mouse over the bar for 510 Minitab will display a value of 14 and bin of 502.5, 517.5 meaning that 14 of the reported average SAT math scores for the 50 states plus Washington D.C. were between 502.5 and 517.5. The heights of the bars in the histogram also give you the count (frequency) for each class. Notice that the height for the class centered at 510 is also 14.

The distribution appears to be **centered** near 530, meaning that of the 51 observations half would be at or below 530 and the other half would be at or above 530. The values are **spread** from 478 to 608 giving a **range** of 130. The **shape** appears to be somewhat **skewed right** or **positively skewed** meaning that a bulk of the data gathers on the left of the graph with the remainder of the data filling in, or trailing, to the right.

The advantage of a histogram is that the construction is easy for a large data set. A disadvantage is that individual observations are not specified. For example, from the histogram you cannot tell what the mean Math SAT(2005) score is for Alabama. For a relatively small or medium size data sets, a quick graphical display that includes each observation we can construct a **stem-and-leaf plot**. The stem-and-leaf plot consists of a vertical list of stems after which are recorded a horizontal list of one-digit leafs.

Example: We can construct a stem-and-leaf plot for the mean Math SAT(2005) scores.

1. First we would need to rank the data from minimum observation (478) to the maximum (608).
2. Next, we would create vertical stems based on the first two digits;

47|
48|
49|
50|
. . .
60|

3. Note that even though some stems may not have an observation, for example there are not observations of scores from 480 to 489, we still need to create a stem for this group if we already created a stem prior to this level (where we did with the stem 47).

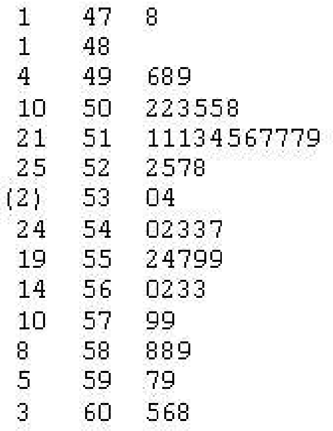
   Now we fill in the leafs by entering the single last digit horizontally for its respective stem:

   47| 8
   48|
   49|689
   50|223558
   . . .
   60|568

From this plot we can now see each individual observation such as 478, 502 (of which there are two such observations since there are two leafs of "2") and 608.

We can also create this stem-and-leaf plot using Minitab:

1. Open the data set (see link at beginning of notes if you do not have the set open)
2. From the menu bar select Graph > Stem-and-Leaf
3. Click inside the window under Graph Variables. This will bring up the list of quantitative variables in the data set.
4. From the list of variables click on Math SAT(2005) and then click the Select button. This should place the variable Math SAT(2005) in the Graph Variables window.
5. Click OK



```
 1    47  8
 1    48
 4    49  689
 10   50  223558
 21   51  11134567779
 25   52  2578
(2)   53  04
 24   54  02337
 19   55  24799
 14   56  0233
 10   57  99
 8    58  889
 5    59  79
 3    60  568
```

As you can see, Minitab produces a bit more information; the left-hand column. When we construct these graphs by hand this column is not required, but Minitab has the machinery capability to do this quickly. The interpretation of this left-hand column is quite simple. The row with the parentheses indicates the row that contains the median observation for this data set and the number in the parentheses is the number of leafs (i.e. observations) in this row. Preceding this point, the numbers indicate how many observations are in that row and prior. For instance, the first "1" indicates that there is one observation in that row (478). The second "1" indicates that there is a total of one observation in that row plus any preceding rows (no observations in this row plus the 478 observation in the first row). The next number "4" says that there are a total of four observations in this row plus any preceding rows (the three observations in this row: 496, 498, and 499 plus the observation of 478). This continues to the row containing the median observation. After the median, the interpretation continues except the number indicates how many observations are in that row and any rows following that row.

Again from this graph we can get the range (478 to 608 = 130), the center (the median is around 530 to 534), and the shape (which shows that many of the observations are gathering near the top and tailing off to the bottom. If you could flip this graph so the stems are place horizontally you would see that the tail goes off to the right with the bulk of the data based to the left symbolizing a distribution shape that is **skewed to the right** or **positively skewed**).

Stem-and-leaf plots can be constructed using what is called **split stems**. In split stems, the stems are either divided into groups of two or groups of five. For instance, if your data consisted of 50 observations ranging from 10 to 30 you could split the stems in two (creating two stems each of “1”, “2”, and 3”) and then for the leafs place any observations of 0 to 4 on the first stem and 5 to 9 on the second stem. The stems could also be split by creating 5 of each stem and then placing in leafs and observations of 0,1: 2,3: 4,5; 6,7; 8,9. Minitab will automatically create split stems if the data warrant a splitting.

Shape

The shape of a dataset is usually described as either **symmetric**, meaning that it is similar on both sides of the center, or **skewed**, meaning that the values are more spread out on one side of the center than on the other. If it is **skewed to the right**, the higher values (toward the right on a number line) are more spread out than the lower values. If it is **skewed to the left**, the lower values (toward the left on a number line) are more spread out than the higher values. A symmetric dataset may be **bell-shaped** or another symmetric shape. The shape is called **unimodal** if there is one prominent peak in the distribution, and is called **bimodal** if there are two prominent peaks. Figures 1, 2 and 3 show examples of three different shapes.

Figure 1: A symmetric shape.

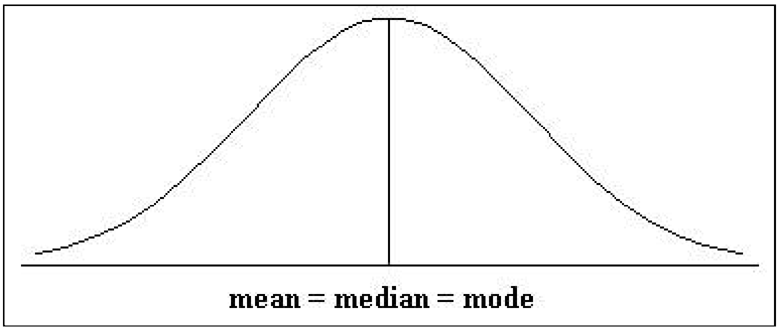


Figure 2: Data skewed to the right.

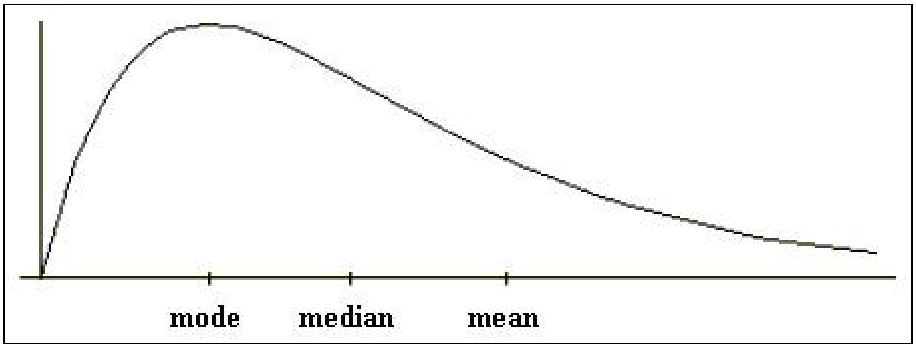
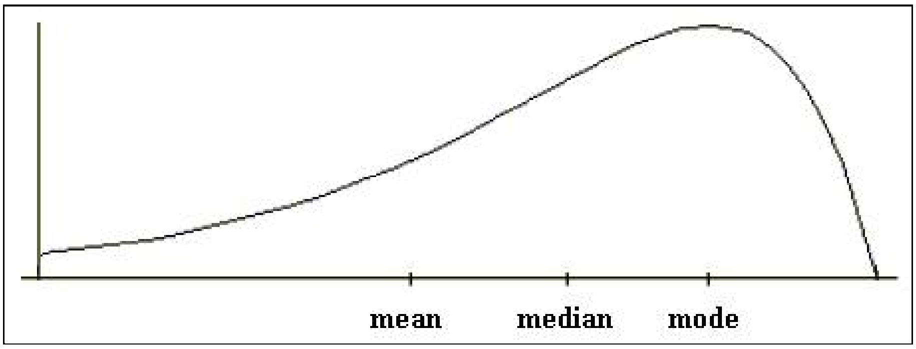


Figure 3: Data skewed to the left.



Describing Distributions with Numbers

Location

The word location is used as a synonym for the “middle” or “center” of a dataset. There are two common ways to describe this feature.

1. The **mean** is the usual numerical average, calculated as the sum of the data values divided by the number of values. It is nearly universal to represent the mean of a sample with the symbol $\overline{X}$, read as “x-bar.”
2. The **median** of a sample is the middle data value for an odd number of observations, after the sample has been ordered from smallest to largest. It is the average of the middle two values, in an ordered sample, for an even number of observations.

So far we have mentioned **center** in a vague manner. **Spread** is inadequately described by range which only provides information based on the minimum and maximum values of a set of data. With center and spread being the two most important features of a data distribution they should be carefully defined.

One measure of center is the **median** or middle value. When the total number of observations is an odd number, then the median is described by a single middle value. If the total number of observations is even, then the median is described by the average of the two middle values.

A second measure of the center is the **mean** or arithmetic average. To find the mean we simply sum the values of all the observations and then divide this sum by the total number of observations that were summed.

Example: From the SAT data set we can show that the participation rates for the nine South Atlantic states (Region is SA) are as follows: 74, 79, 65, 75, 71, 74, 64, 73, and 20. In order to find the median we must first rank the data from smallest to largest: 20, 64, 65, 71, 73, 74, 74, 75, and 79. To find the middle point we take the number of observations plus one and divide by two. Mathematically this looks like this where n is the number of total observations:

$$\frac{n+1}{2} = \frac{9+1}{2} = 5$$

Returning to the ordered string of data, the fifth observation is 73. Thus the median of this distribution is 73. The **interpretation of the median** is that 50% of the observations fall at or below this value and 50% fall at or above this value. In this example, this would mean that 50% of the observations are at or below 73 and 50% are at or above 73. If another value was observed, say 88, this would bring the number of observations to ten. Using the formula above to find the middle point the middle point would be at 5.5 (10 plus 1 divided by 2). Here we would find the median by taking the average of the fifth and sixth observations which would be the average of 73 and 74. The new median for these ten observations would be 73.5. As you can see, the median value is not always an observed value of the data set.

To find the **mean**, we simply add all of the numbers and then divide this total by total numbers summed. Mathematically this looks like this where again n is the number of observations:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{74+79+65+75+71+74+64+73+20}{10} = 66.11$$

APPLET

Let's play around with what these concepts related to location involve using the following applet from Rice University:

Spread (Variability)

The word spread is used as a synonym for variability. Three simple measure of variability are:

Example of Calculating Range and Interquartile Range (IQR)

1. The **range** is found by subtracting the minimum value from the largest value. From the example used above to calculate the mean and median the range for PAC states would be: Range = 79 – 20 = 59

2. To find the IQR we must first find the **quartiles**. The first quartile (**Q1**) is middle of the values **below** the median and the third quartile (**Q3**) is the middle of the values **above** the median. Using the PAC example, we have 9 observations with the median being the fifth observation. Q1 would be the middle of the four values of below the median and Q3 would be the middle of the four values above the median:

$$Q1 = \frac{64+65}{2} = 64.5 \qquad Q3 = \frac{74+75}{2} = 74.5$$

The IQR is found by taking Q3 minus Q1. In this example the IQR = 74.5 – 64.5 = 10.

This **five number summary**, consisting of the minimum and maximum values, Q1 and Q3, and the median, is an excellent method to use when describing a quantitative data set.

3. **Standard deviation** = roughly, the average difference between individual data and the mean. This is the

most common measure of variation.

The example given below shows the steps for calculating standard deviation by hand:

Example of Calculating Standard Deviation

Five students are asked how many times they talked on the phone yesterday. Responses are 4, 10, 1, 12, 3.

Step 1: Calculate the sample mean. = (4+10+1+12+3)/ 5 = 30/5 = 6.

Step 2: For each value, find the difference between it and the mean.

Data Value	Deviation from mean
4	-2 (4 – 6)
10	4 (10 - 6)
1	-5 (1- 6)
12	6 (12- 6)
3	-3 (3 - 6)

Step 3: Square each deviation found in step 2

Data Value	Deviation from mean	Squared Deviation
4	-2	4
10	4	16
1	-5	25
12	6	36
3	-3	9

Step 4: Add the squared deviations found in step 3 and divide by (n – 1)

$$(4 + 16 + 25 + 36 + 9) / (5 – 1) = 90 / 4 = 22.5.$$

This value is called the **variance**.

Step 5: Take square root of value found in step 4. This is the standard deviation, and is denoted by s.

$$s = \sqrt{22.5} = 4.74$$

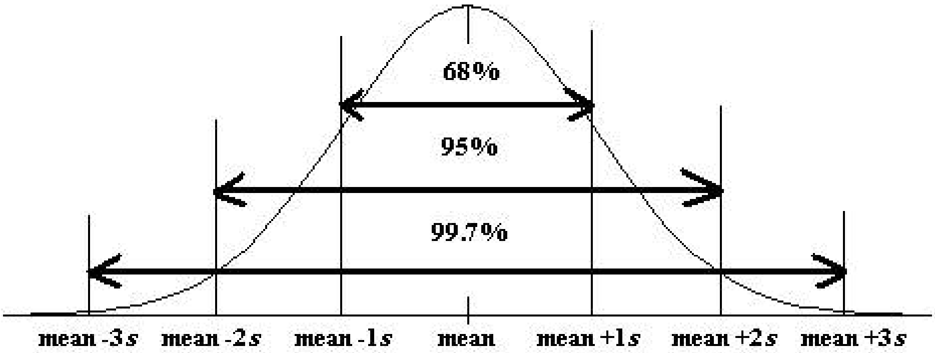
Very roughly, the standard deviation is the average absolute size of the deviations from the mean (numbers found in Step 2).

Standard Deviation and Bell-Shaped Data

For any reasonably large sample of **bell-shaped** data, these facts are approximately true:

- About 68% of the data will be in the interval mean ± s.
- About 95% of the data will be in the interval mean ± (2 × s).
- About 99.7% of the data will be in the interval mean ± (3 × s).

This is called the ***Empirical Rule***.



Example of Empirical Rule

Suppose the pulse rates of n = 200 college men are more or less bell-shaped with a sample mean of = 72 and a standard deviation s = 6.

- About 68% of the men have pulse rates in the interval 72 ± 6, which is 66 to 78.
- About 95% of the men have pulse rates in the interval 72 ± (2 ×6), which is 60 to 84.
- About 99.7% of the men have pulse rates in the interval 72 ± (3 ×6), which is 54 to 90

Finding Outliers Using IQR

Some observations within our data set may fall outside the general scope of the remaining observations. Such observations are called outliers. To aid in determining whether any values in the data set can be considered outliers we can employ the IQR.

Example: From the participation rates of the 9 South Atlantic states given above, we found an IQR of 10. Using this we can determine if any of the 9 observations can be considered outliers. We do this by setting up a “fence” around Q1 and Q3. Any values that fall outside this fence are considered outliers. To build this fence we take 1.5 times the IQR and then subtract this value from Q1 and add this value to Q3. This gives us minimum and maximum fence posts in which to compare the values of the data set.

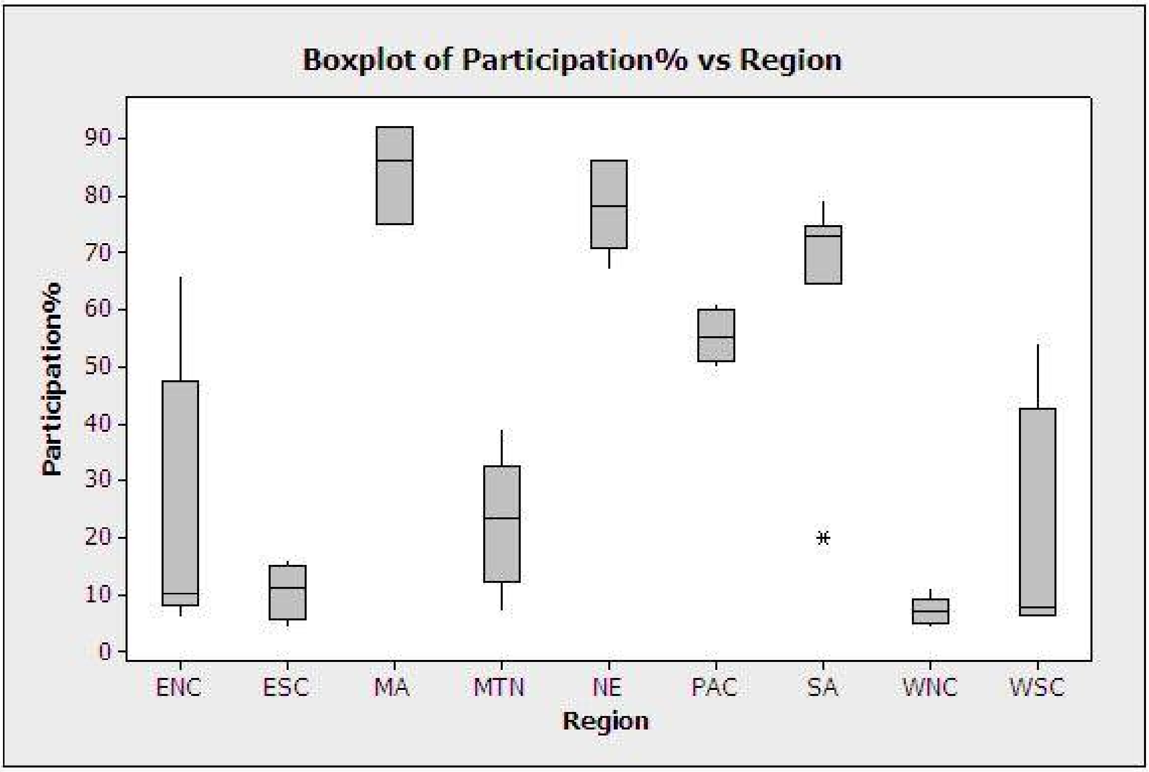
Q1 – 1.5*IQR = 64.5 – 1.5*10 = 64.5 – 15 = 49.5
Q3 + 1.5*IQR = 74.5 + 1.5*10 = 74.5 + 15 = 89.5

Comparing the 9 observations we see that the only data set value outside these fence points is 20 indicating that the observation value of 20 would be considered an outlier for this set of data.

Graphing the Five-Number Summary

The five-number summary (minimum, maximum, median, Q1, and Q3) are used to create a boxplot. A boxplot is very useful for displaying a single quantitative variable or side-by-side boxplots can be used to compare more than one quantitative variable. Using Minitab on the SAT data set we can create the side-by-side boxplots of the Participation Rates by the different Regions:

1. Open the data set (Use this link to the data set, SAT_DATA.MTW, if you do not have the data set open)
2. From the menu bar select Graph > Boxplot
3. Select One Y With Groups
4. Click inside the window under Graph Variables and from the list of variables click on Participation% and then click the Select button. This should place the variable Participation% in the Graph Variables window.
5. Click inside the window under Categorical Variables For Grouping and from the list of variables click on Region and then click the Select button. This should place the variable Region in the Categorical Variables For Grouping window.
6. Click OK



From this boxplot you can compare the distributions of Participation Rates across the nine levels of Region. In Minitab, if you place your mouse cursor over one of the boxplots, for example the boxplot for SA, a pop-up will appear that gives the values of Q1, Median, Q3, IQR, and the sample size. For SA these values are 64.5, 73, 74.5, 10, and 9 respectively. See how these values match those we found when we calculated the five-number summary for SA? If you place your mouse cursor on the “*” for SA the pop-up gives the value of this observation (20) and the row within the data set where this value resides (Row = 49). As we calculated by hand above, Minitab also identified the Participation% of 20 as an outlier for the South Atlantic region.

How is the boxplot created by the five-number summary? The box portion of the plot is made up of Q1 (the bottom of the box); Q3 (the top of the box); and the Median (the line in the box). The whiskers of the boxplot, those lines extending out of the box, are determined by 1.5*IQR. The length of these whiskers depends on the values of the data set. If you return to the boxplot you will notice that for any given boxplot the lengths of these whiskers are not necessarily identical (see the boxplot for the region ENC). These whisker lengths extend to the value of the data set for that region which is closest to the fence posts without extending past them. Using ENC to illustrate this concept, the lower fence post (8 – 1.5*39.5 = – 51.25) is less than 0 (obviously a participation rate of less than zero cannot exist since the lowest possible participation rate would be if no one took the SAT in that state, or 0%). The closest observed participation rate for ENC is 6% so the bottom whisker extends to 6.

Using the boxplot to interpret the shape of the data is fairly straightforward. We consider the whiskers and the location of the median compared to Q1 and Q3. If the data were bell-shaped the median would very near the middle of the box and the whiskers would be of equal length. A data set that was **skewed to the right, positively skewed**, would be represented by a boxplot where the median was closer to Q1 and the upper whisker was longer than the lower whisker. The opposite would be true for a data set that was **skewed to the left, negatively skewed**: the median would be closer to Q3 and the lower whisker would be longer than the upper whisker.

The figures 4, 5 and 6 show examples of the three different shapes based on a boxplot.

Figure 6: Skewed to the left.

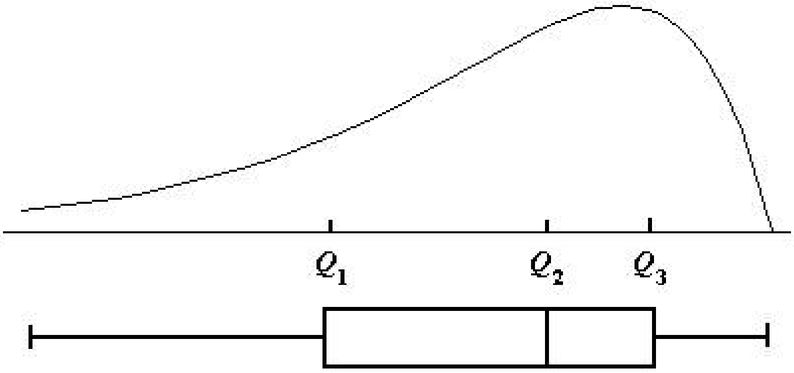


Figure 5: Skewed to the right.

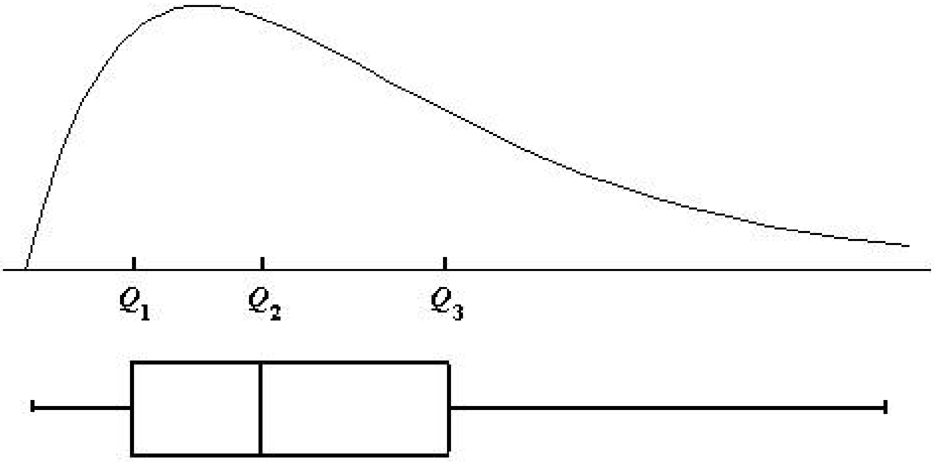
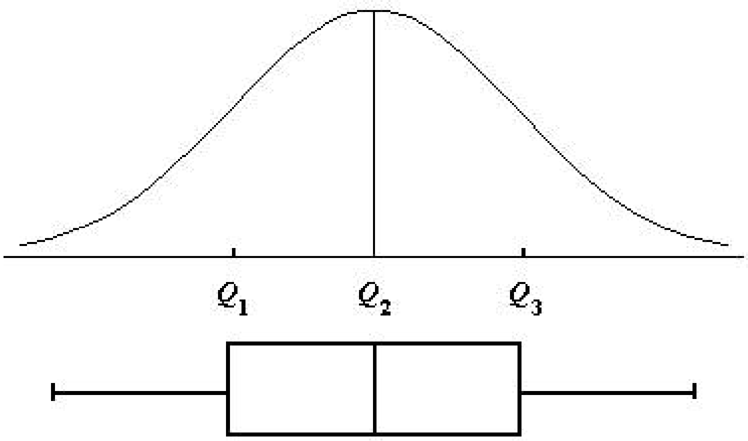


Figure 4: Symmetric.



Summary

In this lesson we learned the following:

- the importance of graphing your data
- how to interpret the shape of a distribution
- what is a five-number summary and its interpretation
- the meaning of descriptive statistics
- what "average" means in statistics-speak
- the relationship between mean and median of a distribution
- some basic Minitab statistics and graphing methods

Next, let's take a look at the homework problems for this lesson. This will give you a chance to put what you have learned to use...