Statistics

Introduction:

The simplest way to describe statistics is that statistics is the use of several mathematical methods used to interpret large amounts of data. Statistics can tell us how data is distributed. It can tell us if two sets of data are related. It can tell how probable certain outcomes are. It can predict the future from the past, and much more. Statistics is a vital form of math used by several professions every day.

History:

The first development of statistics was in 1520, when Geronimo Cardano published a Book on Games of Chance. It was the first book ever written with an organized theory of probability. After this several mathematicians began working on statistics. It was and still is a growing field.

In 1805 Legendre discovered and published the least squares method. This is a fairly complicated way of using previous data to estimate future data. It is still very commonly used and taught today. In 1888 Francis Galton (cousin of Charles Darwin) discovered correlation. Correlation is how we determine how well one set of data (example: height) effects another set of data (example: inseam). In 1893 Karl Pearson discovered standard deviations. These are very important, and constantly being used. For example, the IQ scale is based on standard deviations. Anyone with an IQ of 145 or more is considered a "genius," that's because they are at least three standard deviations above the mean (they have a higher IQ then 99.87% of the population).

In 1896 Peirce comes up with the idea of random sampling. Random samples are very important. For example, if we wanted to know how much money the average Food Lion shopper spends, and we go to the Food Lion parking lot at 2:00 on a Saturday, and then stand at the shopping cart return corrals, and ask the first 50 people who bring back their cart, our data is going to be worthless. What we would have collected wouldn't be data on all Food Lion shoppers, but only on Food Lion shoppers who shop on Saturday afternoons, and who use shopping carts, and who return their shopping carts to the corrals. So, if you want your data to actually tell you what you want to know, you have to be sure that you have eliminated all biases (sampling errors).

In 1899 Lord Guinness, of the Guinness Brewing Company hires William Gosset (a statistician) to help the company. In 1910 Florence Nightingale devised several ways to view statistics graphically, including the histogram, bar graph, and time plot.
More recently, in 1970 John Tukey devised two new ways of graphically viewing data. They are the Stem and leaf plot, and the box and whiskers plot. These are two simple, and yet very effective ways of viewing data. And, they are the two methods you will learn to use in the rest of this worksheet.

The Mathematics:

Let's pretend that we wanted to know what is the typical speed of cars on 421 South going down the mountain. So we go, and sit on the side of the road with a radar gun, and record the following speeds for 25 cars in m.p.h.:
44, 48, 51, 53, 56, 56, 57, 58, 58, 58, 60, 60, 60, 60, 60, 61, 62, 62, 63, 63, 65, 65, 67, 70, 79
Statistics gives us several different ways of summarizing large amounts of data. One of the simplest is the Stem and Leaf Plot. A stem and leaf plot organizes all the numbers according to the first digit (the stem). So, our stem and leaf plot, with only the stems would look like this:

4|
5|
6|
7|

We then add the leaves, in order:

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

However, one rule that we must follow is that every stem and leaf plot must have between 8 and 15 stems. Ours only has 4. We can solve this problem by splitting each stem in two, and labeling them 4 low, then 4 high, then 5 low. . . 7 high. That would look like such:

4_L| 4
4_H| 8
5_L| 1 3
5_H| 6 6 7 8 8 8
6_L| 0 0 0 0 0 1 2 2 3 3
6_H| 5 5 7
7_L| 0
7_H| 9

Statistics likes to describe the shape of the data. Normally shaped data will have most of the data in the middle, with less data at each tail. Our data is normally shaped. If most of the data was towards the minimum, with a long tail at the maximum, then the data would be skewed right. If most of the data was near the maximum, with a long tail at the minimum, then the data would be skewed left.

Made up example of data skewed left: (on next page)

1| 1
2| 4

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

From the last real stem and leaf plot we can tell that most cars are traveling between 55 and 64 m.p.h. We can also use certain calculations to give us descriptive numbers. The mean is an average of all numbers. So, to find the mean we add all the numbers, and then divide by the total number:

44+48+51+53+56+56+57+58+58+58+60+60+60+60+60+61+62+62+63+63+65+65+67+
70+79 = 1433
1433/25 = 57.32
So, we can say that the average speed was 57.32 m.p.h.

We can also use the median to describe the data. The median is the middle number, when the numbers are in order. So, you could count one number from the front, then one from the back, until you reach the middle, or you could use the formula n/2 + 1/2, where n is the total number of recorded data.

25/2 = 12.5
12.5+.5 = 13

So we go to the 13th number, it is 60.

So, we can say that the median speed was 60 m.p.h.

Sometimes the median is much more useful then the average. For example, lets pretend that you're about to move into a new neighborhood. And the realtor tells you that the average income in the neighborhood is $10,008,000. However, upon closer inspection you see that there are only 5 houses in the neighborhood. 4 of them are little tiny shacks, and one is Bill Gates mansion. So, the incomes are: $10,000 , 10,000 , 10,000 , 10,000 , and 50,000,000.

The average is: 10,000+10,000+10,000+10,000+50,000,000 = 50,040,000
50,040,000/5 = 10,008,000

So, the realtor wasn't lying. However, a median will probably be a much better estimator of the data.

The median is: 5/2 = 2.5
2.5+.5 = 3
The 3rd number is 10,000
The median is $10,000

The median gives you a much better idea of the income of the normal neighbor then the average did. (extra: What shape is this data? (normal, skewed left, skewed right))

A different way of visualizing the data is a box and whisker plot, also called a 5 number summary. This is because the diagram is made using 5 numbers. These numbers are called:
Minimum
Q_1
Median
Q_3
Maximum

The minimum is the smallest number in your data. The maximum is the largest number in your data. You already know how to find the median. Q_1 is the same concept as the median. But, it is the median of the first half of the data. Q_3 is the median of the second half of the data.

Let's use the 25 cars again. The smallest number is 44, so that is the minimum. The largest number is 79 (maximum). And we have already calculated the median as 60. If you remember the median was the 13th number, so to find Q_1, we need to find the median of the first 13 numbers.

13/2 = 6.5
6.5 + .5 = 7
Q_1 is the 7th number
$Q_1 = 57$

To find Q_3 we count to the 7th number, starting at the median.
$Q_3 = 63$

So, now we have our 5 number summary:

Min.	44
Q_1	57
Med.	60
Q_3	63
Max.	79

From these we can draw out a box plot:

```
              _____
    |----------|__|__|-------------|
40       50       60       70       80
```

The 5 vertical lines are the 5 number summary. This box and whiskers plot can tell us a whole lot. For example, we now know that the middle 50% of cars were traveling between 57 and 63 mph. Only 25% of the cars were traveling between 63 and 79 mph.

We can also tell the shape of the data from a box and whisker plot. If one tail is excessively long, then the data is probably skewed in that direction.

Made up example of data skewed right:

```
   _____________
|--|___|_________|----------------|
```

Why don’t you try doing your own stem and leaf plot and box plot for the following data:

25 people are asked, as they are leaving Wal-Mart how much money they just spent. The answers were as follows (rounded to the nearest dollar):

2, 3, 5, 5, 7, 11, 12, 17, 19, 21, 21, 25, 27, 29, 30, 31, 31, 34, 39, 50, 55, 75, 79, 83, 96

(hint: use a 0 as your first stem)

How is the data distributed?