Unit 2: What is the first thing you do with data? MAKE A PICTURE!!!!

The first strand of statistics we will be concentrating on is the data analysis strand. We need ways to show the data so that we can see patterns, relationships, trends and exceptions. So, what should we do first with data? There are three things you always do first with data:

1. **Make a picture.** A display of your data will reveal things you are not likely to see in a table of numbers and will help you to *Think* clearly about the patterns and relationships that may be hiding in your data.

2. **Make a picture.** A well-designed display will *Show* the important features and patterns in your data. A picture will also show you the things you did not expect to see: the extraordinary (possibly wrong) data values or unexpected patterns.

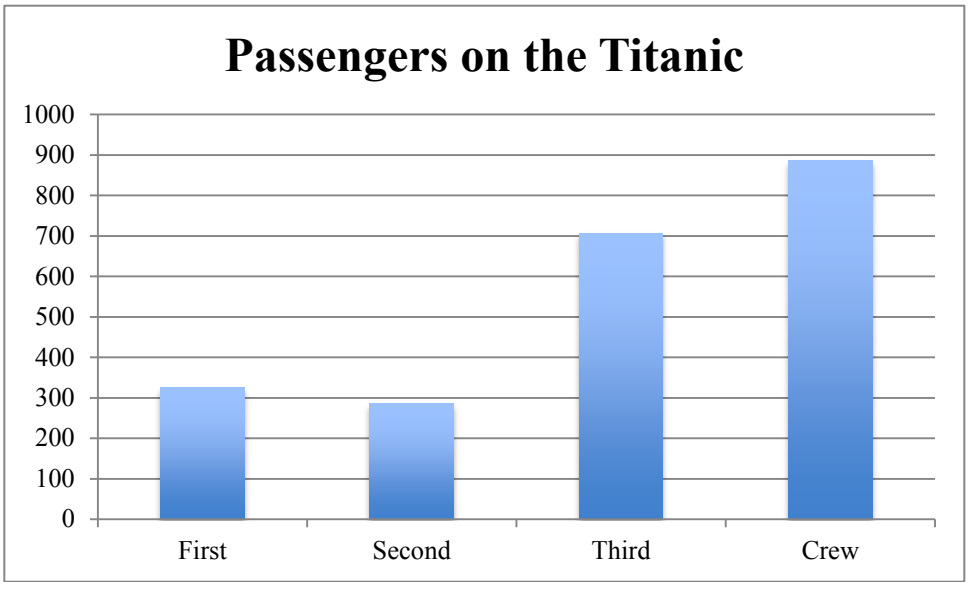
3. **Make a picture.** The best way to *Tell* others about your data is with a well- chosen picture.

Categorical Data

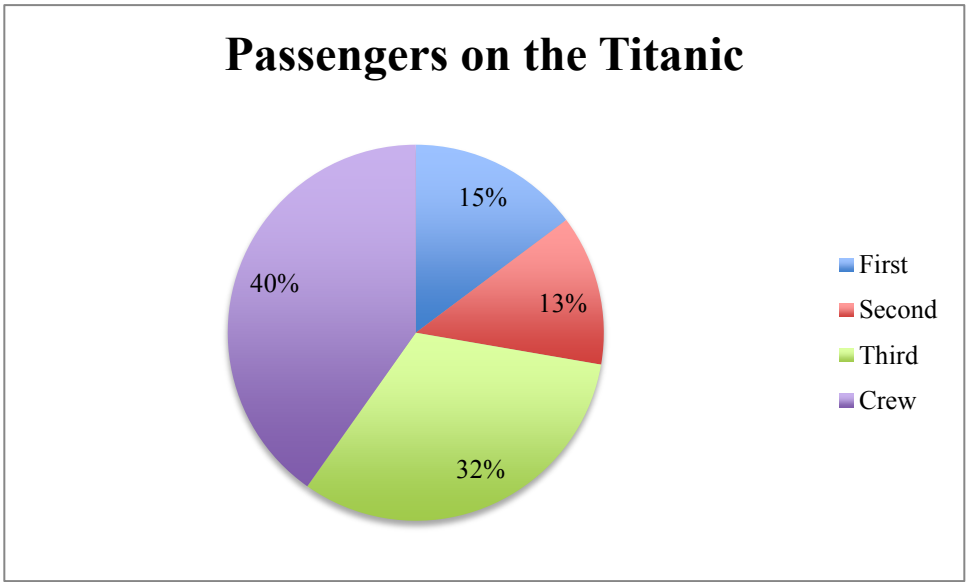
To make a picture, the first thing we need to do is organize our data. One way to organize data is to make a **frequency table**. A frequency table will record totals in various categories. Below you will see frequency table about the passengers on the Titanic.

Class	Count
First	325
Second	285
Third	706
Crew	885

From this frequency table we can make a variety of pictures. Below you will see a **Bar Chart** of our frequency table data.



Bar charts are easily done by hand or with technology. Make sure you understand that this is NOT a histogram. Histograms might look similar but they are different. We will deal with those shortly. Another type of picture we could make from the above data is a pie chart. Even though it is possible to do these by hand, it can be tedious, we will only be analyzing pie charts or creating them with the help of technology. Below you will see a pie chart of our Titanic data.



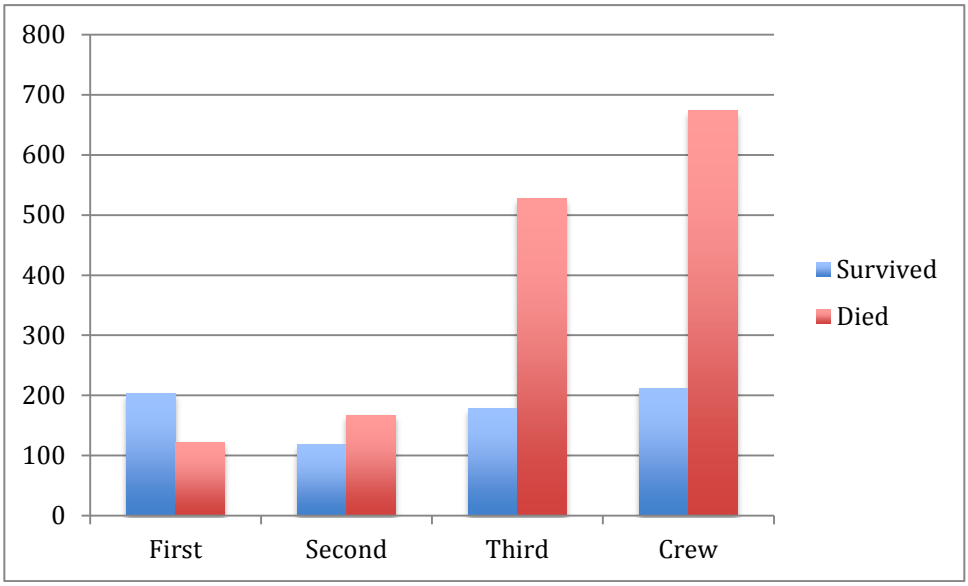
Notice in the above pie chart we are seeing the percents of passengers in each class and not the counts. Our original frequency table allows us to look at the counts or change to the corresponding percentages. If we were to change our counts into percents this table would now be called a **relative frequency table**.

From our frequency table we know how many tickets of each class were sold on the Titanic, and we also know that only about 32% of all those on the ship survived. Was there a relationship between the type of ticket a passenger held and if he/she survived? We will need a different type of table to analyze that relationship. It is called a **2-way table, 2x2 table**. This type of table shows how individuals are distributed along more than one variable. Below you will see a table where the individuals are shown distributed by class and whether they survived.

	First	**Second**	**Third**	**Crew**	**Total**
Alive	203	118	178	212	711
Dead	122	167	528	673	1490
Total	325	285	796	885	2201

The totals at the right and bottom of the table are called the **marginal distributions** and each entry is called a **cell**. We can see that 178 third class passengers survived while 118 second class passengers survived. Does this mean that you would have had a better chance of surviving if you held a third class ticket? Not really, if you look at the percentages, 118 out of 285 (41.4%) second class passengers survived while 178 out of 796 (25.2%) third class passengers survived. Because of the different number of passengers in each class, percentages will be a better way to analyze the data. But what percentages

should we be looking at? Should we look at the percentages out of the total number of passengers or the percentages out of each individual class (**Column percents)** or the percentages out of survival status(**Row percents**)? All of these are possibilities and all are potentially useful or interesting. But before we do that lets look at a picture.

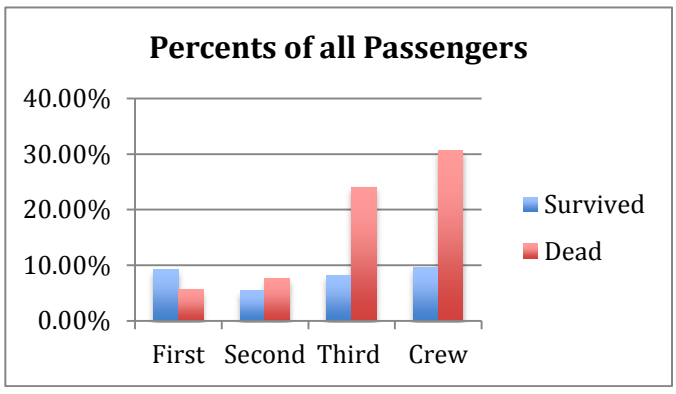


What conclusions can you come up with from this picture? ……..
Clearly it shows that a bigger proportion of first class passengers survived than any other class.

Lets look at the 2-way tables we get when we look at our different percentages.

Percentages of all the passengers

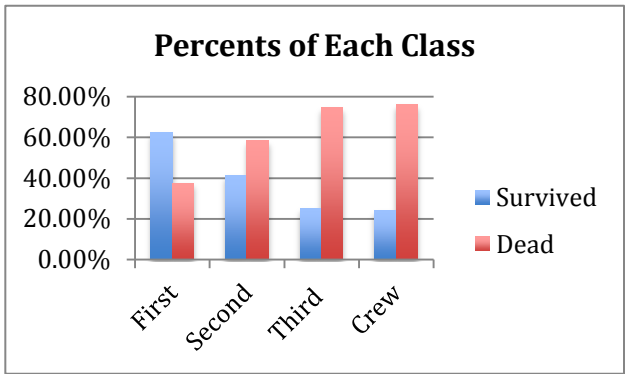
	First	Second	Third	Crew	Total
Alive	9.2%	5.4%	8.1%	9.6%	32.3%
Dead	5.6%	7.6%	24.0%	30.6%	67.7%
Total	14.8%	12.9%	32.1%	49.2%	100%



What conclusions can you come up with from the above table and bar graph?

Percentages by class (Column percents)

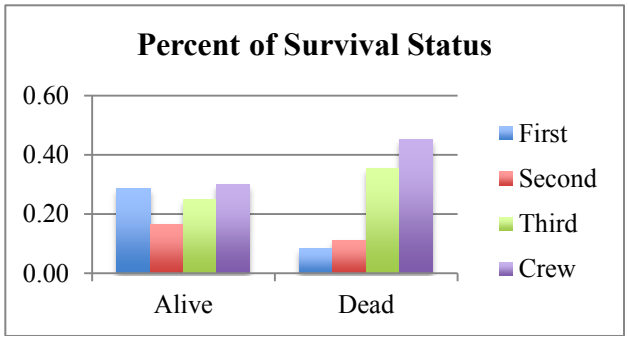
	First	Second	Third	Crew	Total
Alive	62.5%	41.4%	25.2%	24.0%	32.3%
Dead	37.5%	58.6%	74.8%	76.0%	67.7%
Total	100%	100%	100%	100%	100%



What conclusions can you come up with from the above table and bar graph? Are they different from before?

Percentages by survival status (Row percents)

	First	Second	Third	Crew	Total
Alive	28.6%	16.6%	25.0%	29.8%	100%
Dead	8.2%	11.2%	35.4%	45.2%	100%
Total	14.8%	12.9%	31.1%	40.2%	100%



Are your conclusions from this 2x2 table and bar graph the same or different from above? Which 2-way table did you think shows you the best relationship between the type of ticket a passenger held and their survival status?

Check for Understanding

A statistics class reports the following data on Sex and Eye color for students in the class:

Eye Color

Sex	Blue	Brown	Green/Hazel/Other	Total
Males	6	20	6	32
Females	4	16	12	32
Total	10	36	18	64

1. What percent of females are brown-eyed?

2. What percent of brown-eyed students are female?

3. What percent of students are brown-eyed females?

4. What's the distribution of Eye Color?

5. What's the distribution of Eye Color for the males?

6. Compare the percent that are female among the blue-eyed students to the percent of all students who are female?

7. Create a bar graph of Eye Color distribution within each Sex. What conclusions can you make?

Quantitative Data

Quantitative data can also be organized into a frequency table but since the data is numeric that leads us to creating more interesting pictures. We will be discussing **dotplots**, **stem and leaf plots**, **histograms**, and later **boxplots**. When presented with each of the types of pictures we will be looking at 4 things **shape**, **center**, **spread** and **outliers**.

Shape: We will describe our shape as symmetric, skewed right, skewed left or having no shape.

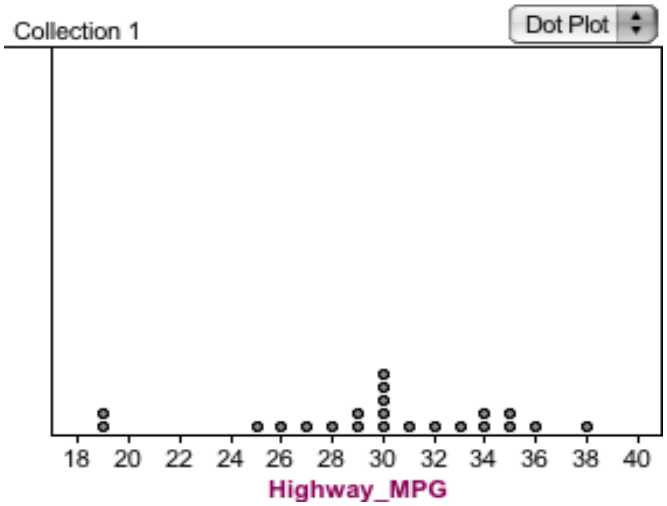
Center: For now we will approximate where we think the center of the picture is. Think balancing point. Later we will calculate a numerical center.

Spread: For now we will use smallest to highest or our Range (highest – smallest). Later we will calculate two different numbers to use when describing spread.

Outliers: Any numbers that fall outside our overall pattern. For now, this is a matter of judgment.

Dotplots

A dotplot is a simple display. It just places a dot along an axis for each data point. Dotplots are a great way to display a small data set. Below, you will see a dotplot of EPA highway gas mileage ratings for 22 model year 2013 midsize cars.



What can we see in our dotplot? The shape looks somewhat symmetrical, especially if we ignore the two dots over 19 mpg. The center appears to be around 30, the spread goes from 25 to 38 (ignoring the outliers) or we can say our range is 13 (38 – 25) mpg. The two dots at 19 mpg are considered outliers.

Notice with the dotplot we don't actually see the numerical data, just a quick picture. Another quick graphing technique that allows us to actually see the individual numerical values is a stemplot, or stem and leaf plot.

Stemplots (Stem and Leaf Plots)

Stem and leaf plots work like dotplots but they show more information. They use part of the number, called the stem and use the next digit to make the leaf. For example, if we had a test score of 83, we would write it 8|3, where the 8 serves as the stem and 3 as the leaf. If we wanted to display scores 83, 76, and 88 together, we would write:

8| 3 8
7| 6

Below is a stemplot of the pulse rates of 24 women taken by a researcher at a health clinic:

5| 6
6| 0 4 4 4 8 8 8 8
7| 2 2 2 2 6 6 6 6
8| 0 0 0 0 4 4 8

This display is OK, but a little crowded. We can split the stems by putting leaves 0 – 4 on one line and 5 – 9 on another. Below is a split stem, stem and leaf plot.

5| 6
6| 0 4 4 4
*| 8 8 8 8
7| 2 2 2 2
*| 8 8 8 8
8| 0 0 0 0 4 4
*| 8

For numbers with three or more digits, you'll often decide to truncate (or round) the number to two places, using the first digit as the stem and the second as the leaf. So, if you had 432, 540, 571, and 638, you might display them as shown below with an indication that 6|3 means 630 – 639.

4| 3
5| 4 7
6| 3

Just like with the dotplot we will examine the shape, center, spread and possible outliers of our stemplots. Using our split stem plot of pulse rates we can conclude that our shape is roughly symmetric, the center is about 72, the spread goes from 56 to 88 and there do not appear to be any outliers.

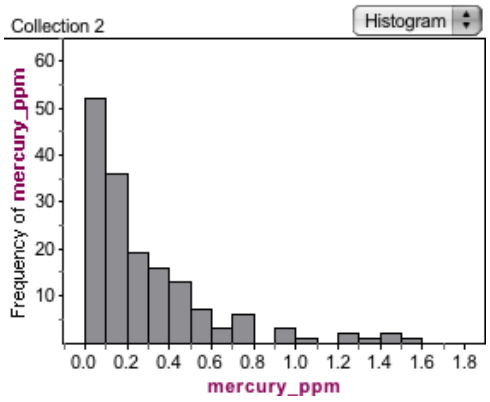
As I mentioned earlier, dotplots and stem plots are convenient for small data sets, but what happens when we have a large data set, what type of picture should we use. We will focus on two particular type graphs, a histogram and a boxplot (we will look at these later).

Histograms

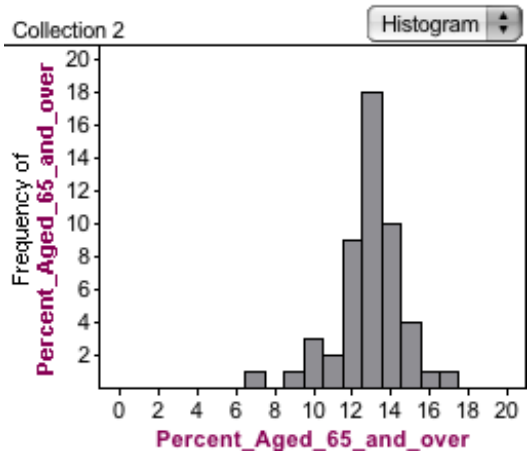
Many people confuse a bar graph with a histogram, but they are different. Bar graphs are used to display categorical data and histograms can only be used to display numerical (quantitative) data.

Many books will go through a list of steps on how to do a histogram by hand, stating that you first need to determine class width or bin width (how wide each bar is), second find out how many data members fall into each class and lastly draw the histogram with the scale of your variable you are displaying on the horizontal axis and frequency or percent on the vertical axis. We will not be doing this. We will be using technology to make our histograms and then ask what are the histograms showing us (shape, center, spread, outliers) about our data. Below you will see a variety of histograms.

Mercury content (ppm) of canned light tuna



Percent of Residents Aged 65 and over



Amount of sodium in randomly selected cereals