Exploratory Data Analysis

Variables

Definition

The big idea of statistics is that we have a question about some large group (the population) that can be answered through measurement. That characteristic which we measure is called the **variable**. Perhaps we want to know the average mass of a kumquat—*mass* is the variable. Perhaps we want to know the proportion of pink VW's in the U.S.—*color* is the variable. The things that we measure (kumquats, VW's) are called **individuals**. The collection of all individuals is called the **population**.

Quantitative vs. Qualitative

Variables come in two basic categories—**quantitative** and **qualitative** (this isn't the only way to classify variables—just the only distinction that's important to us).

Quantitative variables measure *quantities*—mass, time, charge, number, length, etc.

Qualitative variables measure *qualities*—color, flavor, opinions, etc.

Discrete vs. Continuous

Quantitative variables can be broken down into two further categories—**discrete** and **continuous**.

Discrete variables have gaps in their possible values—they can only take on discrete (certain) values. The set of Integers ($\mathbb{Z}$) is an example of a discrete set. Discrete variables will almost always measure the *number* of some thing—the number of houses; the number of people; the number of cars, etc.

Continuous variables have no gaps in their possible values. The set of Real numbers ($\mathbb{R}$) is an example of a continuous set. Continuous variables will typically measure physical phenomena—mass, length, volume, ratio, etc.

The Distribution of a Variable

Definition

The **Distribution of a Variable** is a list (chart, picture—something) that shows what values the variable can take, and how often it takes each value. It turns out that most of the calculations that we'll make this year depend on knowing some things about the distributions of various variables.

Main Points

There are three main features of a distribution that we want to know—**center**, **spread** and **shape**.

The center can be described as the typical value of the variable; or the most common value; or…well, there are lots of ways to say this. More on this later.

The spread can be described as the range of possible values; how wide is the distribution? Again, there are many ways to say this. Again, more on this later.

The shape is a feature that can only be seen. There are two categories of shape that are important to us: **symmetric** and **skew**. Symmetric is self-explanatory. Skew means that one end is larger (taller) than the other. The side that is smaller is the direction of the skew. For example, the distribution in Figure 1 is Skew Right, while Figure 2 shows a fairly symmetric distribution.

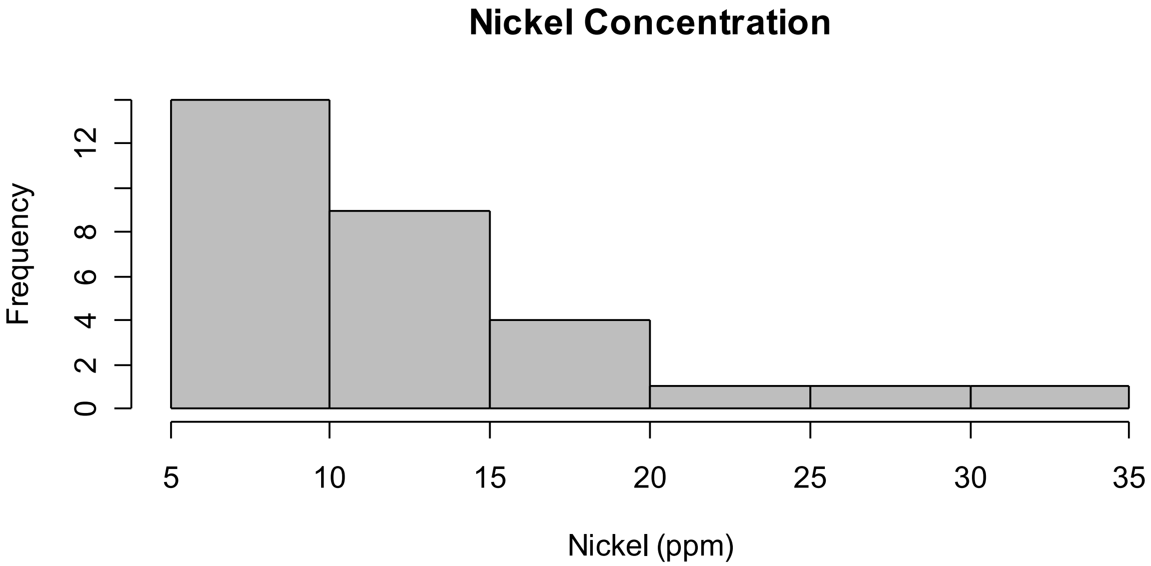

Figure 1 - A Skew Distribution

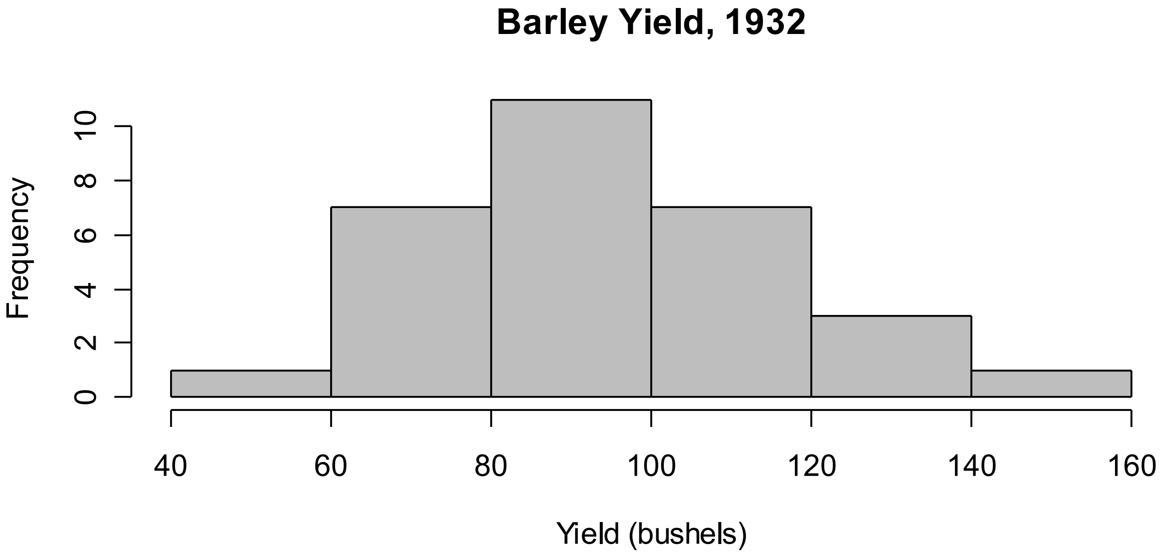

Figure 2 - A (fairly) Symmetric Distribution

Graphic Displays

For Qualitative Variables

Bar Charts

Label the *x*-axis with the values of the (qualitative) variable; label the *y*-axis as frequency (count). Make a bar of the appropriate height for each category. This is a bar chart. Figure 3 gives an example:

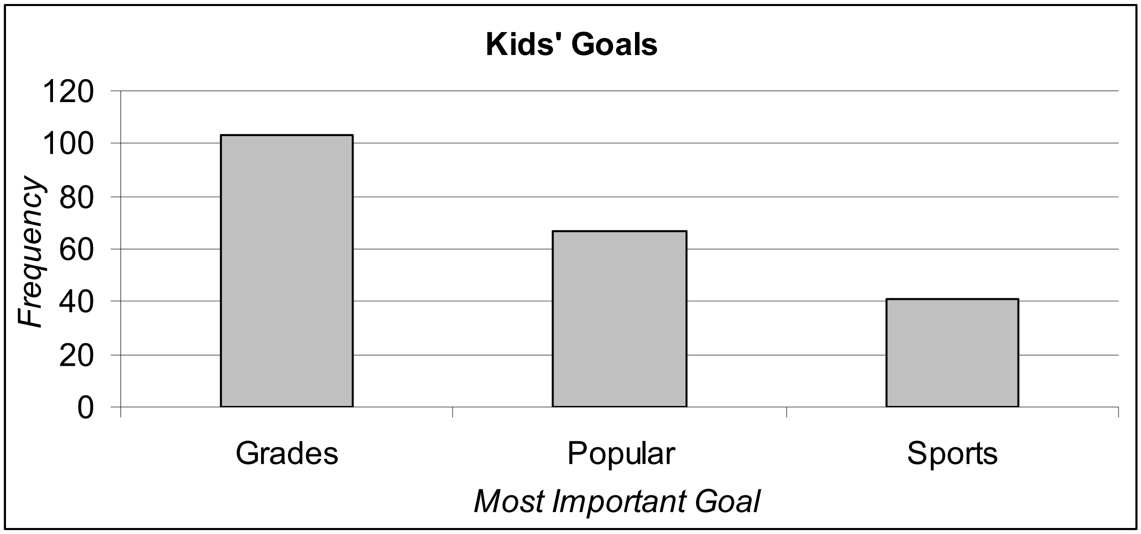

Figure 3 - A Bar Chart

Pie Charts

A circle is divided into sectors (pie slices)—one for each observed value of the variable. The area of each sector is proportional to the percentage of the data that have that value. Figure 4 gives an example:

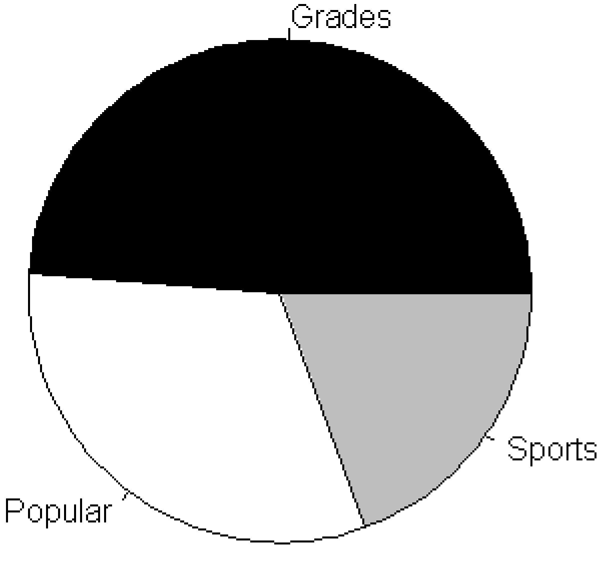

Figure 4 - A Pie Chart

For Quantitative Variables

Histogram

Some people (mistakenly) use the terms bar chart and histogram synonymously—they are not the same thing. For a histogram, the *x*-axis is the quantitative variable, and the *y*-axis is the frequency (count). The *y*-axis can also be labeled with Relative Frequency (percent), Cumulative Frequency (total count for this, and all previous groups), or Cumulative Relative Frequency (total percent for this, and all previous groups). This last type is also called an **Ogive**.

The variable's values are divided into groups (bins, intervals, classes, buckets—there are many different names), and a bar is drawn (of the appropriate height) showing the amount (count, percent, etc.) of the data that fall in that group. Figure 5 gives an example:

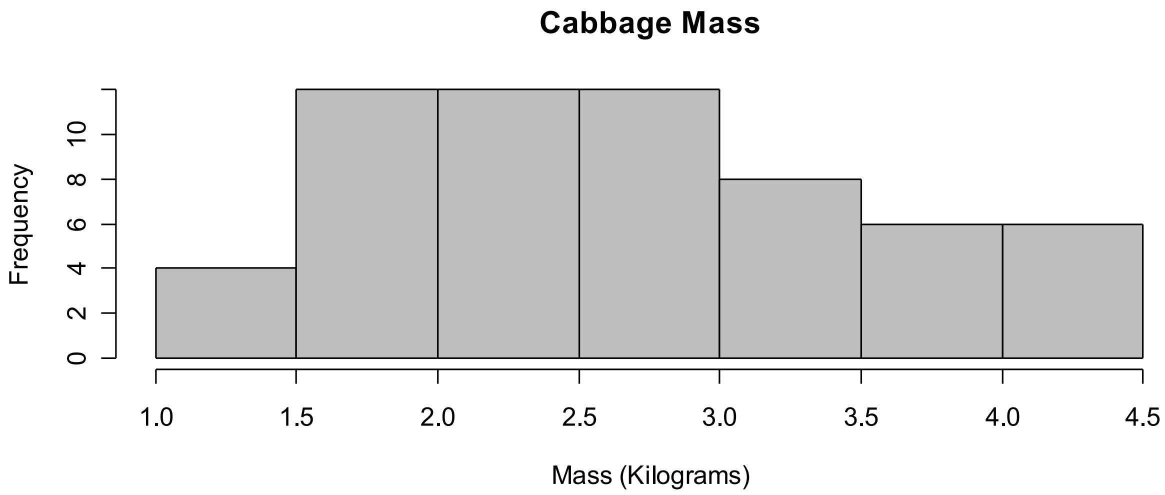

Figure 5 - A Frequency Histogram

Making a good histogram is an art. They key to this is in the choice of groups. To that end, here are some hints for choosing groups:

[1] Ideally, there should be between 5 and 10 groups. A few more than 10 is OK; fewer than 5 groups is a bad thing.

[2] The groups should have very natural boundaries. For example, groups of 2 – 3.1, 3.1 – 4.2, 4.2 – 5.3, etc. are not very natural. A better choice would be 2 – 3, 3 – 4, 4 – 5, etc.

Stem and Leaf

The Stem and Leaf plot uses place value to create groups. Here's an example:

```
 - Brain Mass -
0 | 000000001111122222444477
1 | 3
2 |
3 |
4 | 6
5 | 7
```

```
*where 1|3 means 1300g.
```

The last (rightmost) digit becomes the leaf, and is plotted to the right of the vertical bar. The other digits become the stem, which are plotted to the left of the vertical bar. The stems are ordered, and the leaves follow with their associated stem. We also like for the leaves to be listed in order. Plus, notice the legend that tells us how to read/decode the plot.

So the different values in the leftmost digits determine the number of stems. Just like histograms, we want 5 to 10 stems (ideally). What if there isn't very much variation in those digits? What if there is too much variation in those digits?

If there is too much variation (too many stems), then round all of the data by one place value.

If there is too little variation (too few stems), then split the stems.

```
 - Cabbage Mass -
1 | 0344
1 | 556666777789
2 | 000122222234
2 | 556666788888
3 | 00111223
3 | 567888
4 | 022333

*where 1|0 means 1.0 kg.
```

Notice that there are two '1' stems, two '2' stems, etc. The first '1' only has leaves of zero through four; the second has five through nine (5 possible leaves and 2 stems makes 10 digits).

The beauty of the stemplot is that you get all of the benefits of a histogram, but the data are all still intact (usually), and you don't have to work very hard in creating groups.

Dotplot

A dotplot is a lot like a histogram. It is best used with discrete data. For each value in the data set, make a stack of dots to represent the number of times that value occurs. If the data set is large, you can make each dot represent more than one datum. It is more likely that you'll need to read and understand one of these, than it is to actually make one.

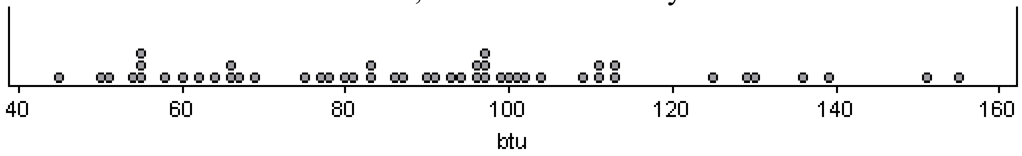

Figure 6 - A Dotplot

Comparing Two Data Sets

If you want to compare two data sets, then make sure that the two graphic displays are as alike as possible—the scale of the axes, the groups for the histogram, etc. If you are using stemplots, use the same set of stems for both sets (thus creating a back-to-back stemplot). See Figures 7 and 8.

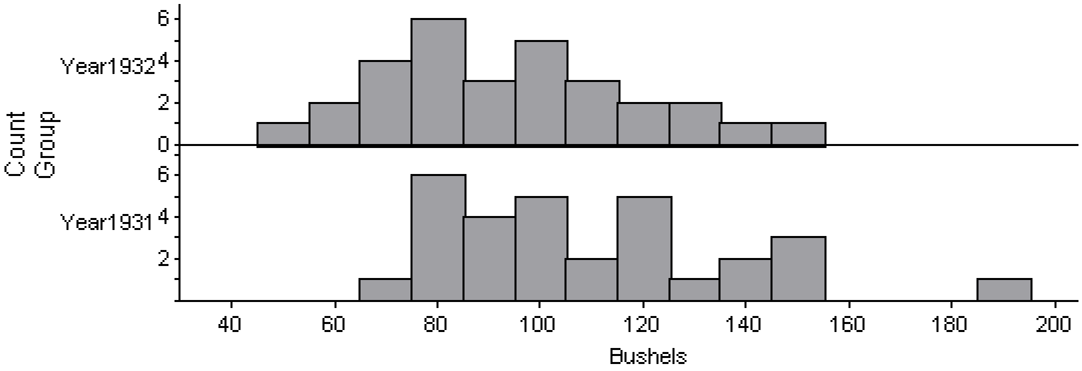

Figure 7 - Histograms with a Common Scale

```
 - 1931 Yield and 1932 Yield -
                  |  4  |
                  |  5  |  0
            99877 |  6  |  226778
                  |  7  |  6
        127990689 |  8  |  001247
                  |  9  |  24679
             2450 | 10  |  0358
                  | 11  |  267
            12671 | 12  |  6
                  | 13  |  0
                  | 14  |  08
                  | 15  |
                  | 16  |
                  | 17  |
                2 | 18  |
**where 9|8|0 means 89 bushels in 1931 and 80 bushels in 1932
```

Figure 8 - Back to Back Stemplots

Examples

[1.] A sample of 13 batches of Portland cement were measured for the heat emitted (calories per gram). The data are shown in Table 1. Let's construct a histogram of the data.

Table 1 - Heat Emitted by Portland Cement

78.5	74.3	104.3	87.6	95.9	109.2	102.7
72.5	93.1	115.9	83.8	113.3	109.4	

I notice that the data range from the 70's to the 110's—so setting group widths at 10 should create enough groups. Now I just need to count how many data are in each group.

Value	Frequency
70 – 80	3
80 – 90	2
90 – 100	2

100 – 110	4
110 – 120	2

Now, scale and label the axes; make bars for each group.

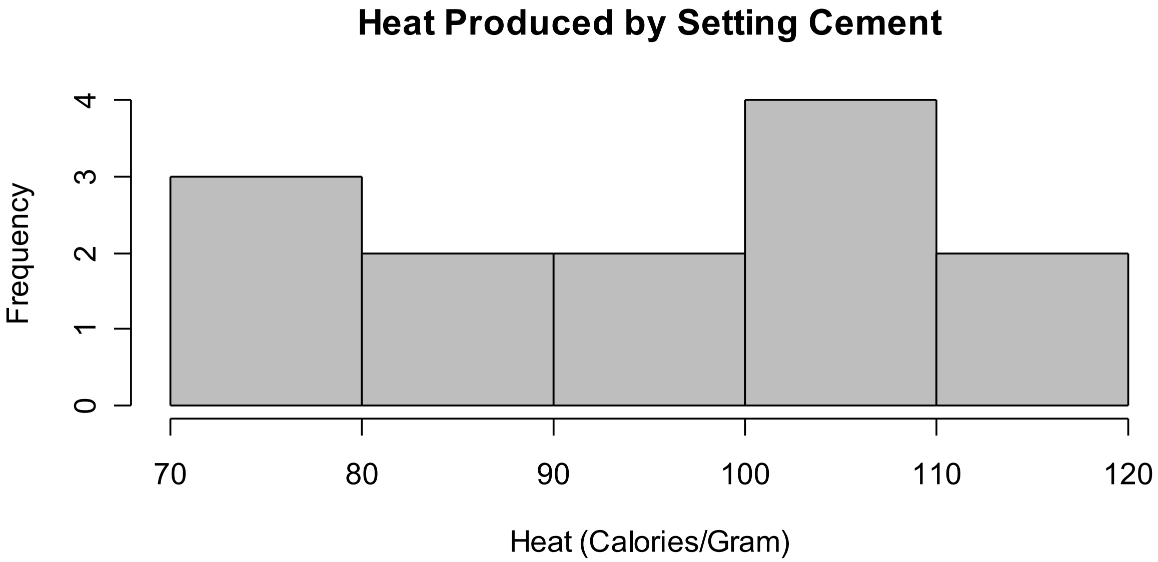

Figure 9 - Histogram for Example 1

[2.] A new material was tested for durability in the soles of boys' shoes. The durability was measured (units are not given), and the results are shown in Table 2. Let's make a stemplot of the data.

Table 2 - Material Durability

13.2	8.2	10.9	14.3	10.7	6.6	9.5	10.8	8.8	13.3

The data are all given to one decimal place, so the stems will be the integer part, and the leaves will be the tenths digits. The integer parts range from 6 to 14, which is good—make stems for the numbers six to fourteen, and start plugging in the leaves!

```
 6|6
 7|
 8|28
 9|5
10|978
11|
12|
13|23
14|3
```

After the first pass, the leaves aren't in order—let's fix that. Plus, let's add a legend.

```
 - Shoe Wear -
 6|6
 7|
 8|28
 9|5
10|789
11|
12|
13|23
14|3
** where 13|3 means 13.3 (units). **
```

[3.] the data below show the number of people living in 40 randomly selected households. Let's make a dotplot.

Table 3 - Number of People living in a Household

2	6	2	2	1	7	3	1	1	1
5	5	1	2	3	1	4	4	4	2
2	2	3	3	2	2	2	3	4	3
1	5	3	3	5	2	2	3	2	4

To make the dotplot, start with a frequency distribution.

Value	Frequency
1	7
2	13
3	9
4	5
5	4
6	1
7	1

Now—plot some dots!

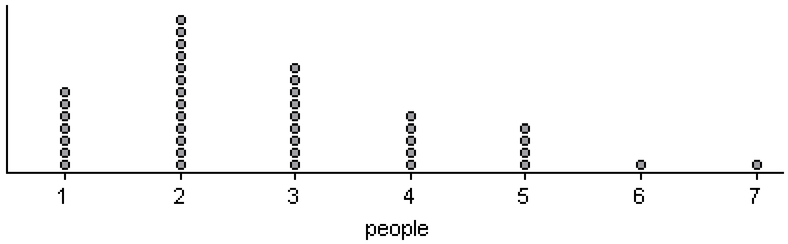

Figure 10 - Dotplot for Example 3

Numeric Summaries

Measures of Center

Mean

The most popular measure of center is formally known as the **Arithmetic Mean** (though really geeky statisticians will call it the *First Moment*), but is often referred to simply as **the mean**. Some will even refer to it as the **average**, but that's a bad idea for us.

Equation 1 - The Formula for the Sample Mean

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Equation 1 reads "*x* bar equals the sum of all *x* sub i from i equal one to *n*, divided by *n*." In plain English—add the numbers, and divide by the number of numbers (*n* typically represents the number of data collected). We could do this by hand, but we'll typically let the calculator do it for us.

The mean is a good measure of center if the data aren't too skew—if there aren't any outliers.

An **outlier** is any datum that does not seem to be in accord with the rest. For example, if you are measuring heights of students at your school, you'd probably get data like 1.85m, 1.58m, 1.6m, etc. If you happened to get one that was 2.5, that would be unusual—it's an outlier.

Median

The median is (probably) the second best measure of center. It is defined as a value where 50% of the data have smaller values (notice that this definition actually allows many different values for the median). In practice, we often refer to it as "the middle number." The median is also said to measure the "typical" value of the distribution.

It doesn't have a formula, and it doesn't have a standard symbol—M and $\tilde{x}$ are most common.

The median is resistant to outliers, so it is a better measure of center when the distribution is skew / outliers are present.

In practice, we find the median by listing the data in order, and picking the middle datum (if there are an odd number of data), or the mean of the two middle data (if there are an even number of data). Often, we'll simply let the calculator do this for us.

I should point out that different software will calculate the median in different ways—so sometimes the median that is listed in a textbook (or computer output) will differ (slightly) from what you think it ought to be. They should, at least, be close (unless the data are highly varied…).

Other

There are other measures of center—mode, midrange, trimmed mean…

The **Mode** is the most frequent datum in the data set. If several data are equally frequent, then the distribution is multimodal; if all data are equally frequent then there is no mode. Not very useful…

The **Midrange** is the value halfway between the maximum and minimum.

A **Trimmed Mean** is the mean after some percentage of the data are trimmed from the ends (high and low).

These are merely curiosities to us—mean and median are the real players in our game.

An Example

[4.] Back to Example 1—what is the mean heat produced? What is the standard deviation of these data?

The mean is 95.423, and the standard deviation is 15.044.

Measures of Spread

Standard Deviation and Variance

A deviation for a datum is the difference between the datum and the mean of the data. So the standard deviation is sort of like a "mean deviation."

The problem comes from the fact that deviations can be positive or negative, and we want the standard deviation to measure the spread of the data. If we just added all the deviations and divided by the number of data, we may (will!) get zero (no spread) even though there is some spread to the data. Eliminating negative numbers can be done two ways: absolute value, and squaring (actually, the square root of something squared is a definition of absolute value). For reasons that deal with calculus, we choose to square.

Equation 2 – The Formula for the Sample Standard Deviation

$$s_x = \sqrt{\frac{\sum_{\mathrm{i}=1}^{n}\left(x_{\mathrm{i}} - \bar{x}\right)^2}{n-1}}$$

Notice in Equation 2, however, that we divide by *n* – 1 instead of *n*. The reason for this comes later, when we want to use values from the sample to estimate values in the population. The **sample variance** $s_x^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}$ is an unbiased estimator of the **population variance** $\sigma_x^2 = \frac{\sum_{i=1}^{n}\left(x_i - \mu\right)^2}{n}$ (trust me!); thus, we must divide by *n* – 1 to get a "good" (we *can't* say unbiased; again, trust me!) value for the sample standard deviation. There are those who believe that this is incorrect, and that the sample standard deviation should be calculated using *n* (The IB Programme is among them). Not us—not for the AP program.

Since the sample mean is greatly influenced by outliers, and the sample standard deviation uses the sample mean in its calculation, it is also greatly influenced by outliers.

We will probably not ever do this by hand—we'll let technology do it for us.

Quartiles and Interquartile Range

For measures of center, there was an alternative—something that wasn't greatly affected by outliers. So it is also with measures of spread. But first, we must discuss **measures of position**.

Measures of Position indicate a datum's relative position in the data set—how many data are above/below. We've already seen one—the median. But there are more.

If you divide the data into quarters, then the dividing lines are the **Quartiles**. The **First Quartile** (Q_1) marks the lowest quarter (25%) of the data. The **Second Quartile** marks the lowest two quarters (50%) of the data (so $Q_2 = \tilde{x}$). The **Third Quartile** marks the lowest three quarters (75%) of the data. We typically don't talk about the Fourth Quartile (which probably ought to be the maximum value of the data).

In practice, we find the quartiles as the medians of the upper and lower halves of the data. If there is a middle datum, it is not included in either half. Again, different technologies will do this differently, so your answers may not always be the same as the book (or test, or…).

The minimum, first quartile, median, third quartile, and maximum make the **Five Number Summary** for a data set. These will come in handy (graphically) a little later.

Of course, you don't have to divide into quarters—you could divide into tenths (**Deciles**), or hundredths (**Percentiles**)…

The difference $Q_3 - Q_1$ is called the **Interquartile Range**. This is an alternate measure of spread that is not greatly influenced by outliers.

Note that Q_1 and Q_3 mark the boundaries of the middle 50% of the data…a potentially useful fact.

Other

There are other measures of spread—**range** being the most interesting of them. The range is simply the difference in the maximum and minimum. This is the easiest measure of spread to calculate from a graph.

An Example

[5.] Back to Example 2—what is the five number summary?

Minimum: 6.6; Q_1: 8.975; Median: 10.75; Q_3: 12.630; Maximum: 14.300

Hints on Shape

In general, if the mean is less than the median, then the shape is probably skew left. If the mean is larger than the median, then the shape is probably skew right. If the mean and median are approximately equal, then the shape is probably symmetric.

Remember, this is only a rule of thumb to be used if you don't have any other way to look at the shape!

The Box Plot

Standard

So we now have another graphic display to introduce—the Box Plot (or Box and Whiskers plot). The box plot connects the values in the **Five Number Summary**—Minimum, Q_1, Median, Q_3, and Maximum.

Draw an axis horizontally (for the variable). Make small vertical marks at each value of the five number summary. Connect the lines for Q_1 and Q_3, making a divided rectangle. Now draw lines from the edges of the box to the remaining vertical lines. Figure 11 shows an example:

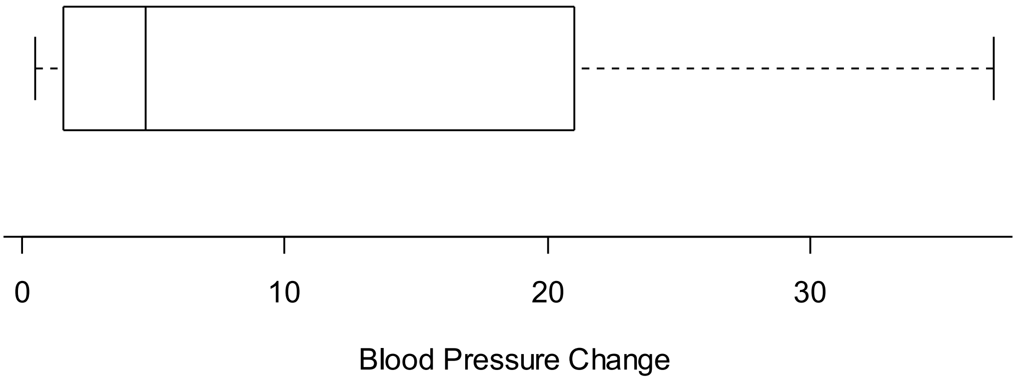

Figure 11 - A Standard Boxplot

Modified

There is another way to draw a boxplot—a way that identifies and marks outliers.

First, identifying outliers—**Tukey's Rule**. Dr. John Tukey (who invented stemplots and boxplots) suggested that any value above $Q_3 + 1.5 \cdot IQR$ or below $Q_1 - 1.5 \cdot IQR$ is probably an outlier. So when marking our vertical lines, don't put one at the minimum—put it at the lowest value that isn't an outlier. Same for the maximum—put the line at the highest non-outlier. Then mark the outliers with separate symbols. Figure 12 shows an example:

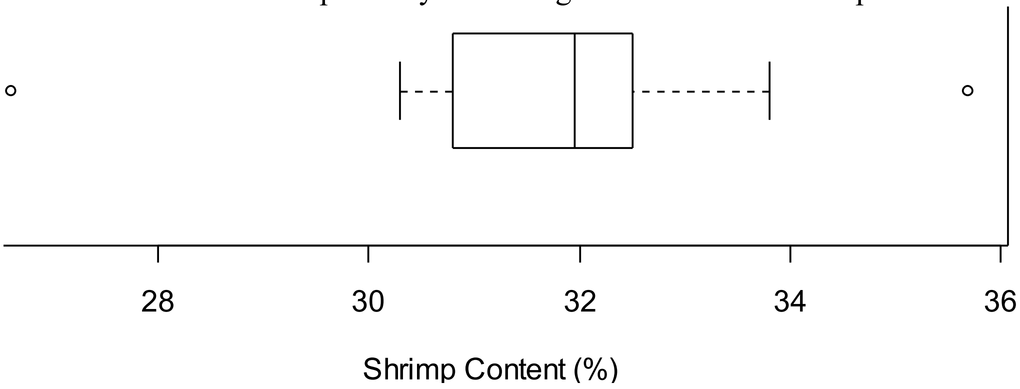

Figure 12 - A Modified Boxplot

Often, we will let the calculator make these plots for us.

An Example

[6.] A sample of kale was sent to 15 different laboratories, and its DDT content was measured (parts per million). The results are shown in Table 4. Let's construct a boxplot of the data.

Table 4 - DDT in kale

2.79	2.93	3.22	3.78	3.22
3.38	3.18	3.33	3.34	3.06
3.07	3.56	3.08	4.64	3.34

So the first thing we need is the five number summary.

{2.790, 3.075, 3.220, 3.360, 4.640}.

So we need an axis scaled from around 2.5 to around 5. Perhaps we should check for outliers…the IQR = 3.360 – 3.075 = 0.285, so $Q_3 + 1.5 \cdot IQR = 3.7875$ and $Q_1 - 1.5 \cdot IQR = 2.6475$. So there is a high outlier (in particular, 4.64)—let's make a modified boxplot. Scale the axis from around 2.5 to about 5; make your marks; connect the lines.

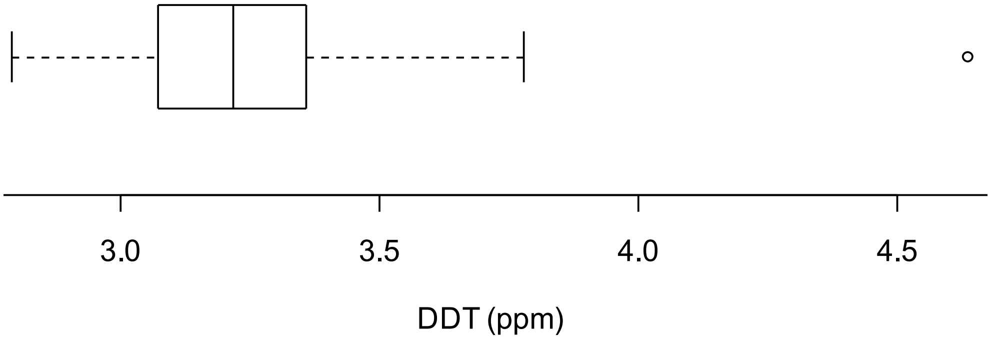

Figure 13 – The DDT Boxplot

Comparing Distributions

When describing a single distribution, you want to comment on its center, spread and shape. When comparing two distributions, you will want to compare center, spread and shape! Don't forget to make the comparisons *in context*.