Introduction to Data Analysis

When reading primary literature, you have probably found graphs and descriptions of data and statistical tests that you might not be familiar with or really understand. Some of the numbers were followed by a plus/minus sign and sometimes they talked about p-values and ANOVA's and t-tests. The idea of all this odd language is just to tell you something about how much variability there is in the data that was collected, and how much uncertainty there is in the estimates that are reported. Typically we are interested in examining one summary value from the data we collect. Such summary variables are called **statistics**. Examples of statistics include the mean, the maximum, the minim, etc.

Sometimes the statistics we compute measure the variability in the data. For example, say we found out that the average height of people in our class was 65 inches. What would that tell someone who never saw us about what our class looks like? Would they think everyone was just about 65 inches tall? Would they think that half of us were 60" and the other half were 70"? Or maybe most people were 72" and a few were 50"? How could they tell the difference? What else would they need to know? Reporting measures about the spread in the data collected can help give a better understanding of the variability in the data.

Other numbers (and symbols) you see in the papers are statistical ways of telling you a little more about the make-up of the **whole population** rather than just telling you about the "sample" of data you happened to collect. For example, we might really be interested in knowing about the average height of people in all classes, rather than just the average height of people in the one specific class we happened to take. Statistical inference is the procedure whereby we try to understand properties about an entire population on the basis of statistics obtained from a small sample drawn that is collected from this larger population.

Measures (statistics) of Central Tendency

The first thing to think about is what kind of "average" is this? When we say "average," a statistician says "measure of central tendency." There are several different ways to measure (i.e., statistics) that central tendency is often quantified including:

1. **MEAN** is the one most of us mean when we say "average." It's the sum of all the values in the set divided by the number of values.

2. **MEDIAN** is the middle value of all the numbers in the set (half are bigger, half are smaller)

3. **MODE** is the value that occurs most often.

Measures (statistics) of variability

RANGE tells what the largest and smallest values in that data that was collected.

The **STANDARD DEVIATION** gives a more robust indication of how broadly scattered all the values are--not just the largest and smallest. It is a statistic that tells you how tightly all the various examples are clustered around the mean in a set of data. When the examples are pretty tightly bunched together the standard deviation is small. When the examples are spread apart that tells you that you have a relatively large standard deviation.

When the measurements reported have a bell-shaped distribution (which is a fairly common shape for statistics reported in papers), one standard deviation away from the mean in either direction from the mean accounts for somewhere around 68 percent of the data points. Two standard deviations away from the mean account for roughly 95 percent of data points. And three standard deviations account for about 99 percent of the data points.

The **standard deviation** is calculated by the following formula:

$$sd = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1}}$$

- sd is standard deviation. It's also sometimes represented by s
- The $\bar{x}$ with a bar over it is the mean
- The x_i indicates individual values of the data that were collected
- N is the number of data points that were collected
- Σ signifies taking the sum over the data that was collected (i.e., over the x_i's)

Statistical tests

Statistical tests used to answer questions about an entire populations of individuals (rather than just the samples of data we have collected). For instance, we might be interested in knowing whether two or more means, as calculated over all individuals in these populations, are different from one another. As a concrete example, say you think that people who eat vegetarian diets are likely to be shorter than people who include meat in their diets. So you decide to compare the heights of a group of vegetarians at Hampshire and a group of carnivores, and you calculate the mean height of each group. (Of course, you would want to be careful about making claims about causes here – for example,

people who are shorter might prefer to be vegetarian rather than the fact that being vegetarian made them shorter – but we will ignore that issue here).

If the mean heights of the two groups of the data you collected were exactly the same, then you might feel comfortable saying that there isn't any evidence that diet affects people's height. But what if there was a little difference between the means in the groups you measured, like say mean height of members in the vegetarian group was 1/4" shorter than the mean height of the individuals in the group that eats meat? Would you say "Wow, I proved that diets containing meat make people grow taller!!" Maybe, but that would probably be rash. You'd want to know more about the data:

- How many subjects were in each group? Maybe you had so few subjects that you would not feel confident that if you had measured the heights of different groups of vegetarians and meat eaters that you would see a difference in the mean heights between the groups.

- What's the standard deviation? Maybe the range of heights in both groups was so broad that many that was a substantial overlap between the heights of most people in these groups, and again, if different groups were measured you would not see a difference in the mean heights.

- You might be more convinced if the difference between the means was 1/2" or 1" or 2". If the difference between the means was 12", or if the height of every meat eaters was taller than every vegetarians, you'd really start thinking you've got some results that support your hypothesis. But when is the difference big enough to consider it a "real" difference?

A statistical test is a way of helping you understand whether the differences in the statistics from the data you collected actually reflect differences if in the population (if you had measured all individuals) or if the differences you measure are just due to the random fluctuations in the data you happened to measure. If your hypothesis is that the variable you're looking at (diet) probably doesn't really influence the outcome (height), then you'll expect no "real" difference between the means in the data you collected. In statistics talk that's called the "null hypothesis." It means that your hypothesis is "there are no real differences in the populations I am measuring" (at least in terms of the specific measurement you are making).

Results from running a statistical test can tell you that it is very unlikely that there are no differences in the populations that you are measuring – i.e., you can *reject* your null hypothesis that there is not differences in these populations. Such a test would show that the difference you see between the means in the sample of data you collected is probably due to a "real difference" in the means of the larger populations that the data came from (e.g., if you have measured <u>all</u> vegetarians and meat eaters, then there would be a difference in their heights). It might seem a bit strange to show that you see a real difference by saying it is unlikely that there is no difference, but this is the logic that enables these tests work.

So how do you figure out if you've rejected your null hypothesis? There are a large number of difference tests that you can run which will give you an answer, and the test you choose to use will depend on type of question you are asking (e.g., are you interesting in seeing whether there are difference between means, are you interested in in asking questions about proportions, etc.). However all these tests report there results using a p-value, which is the probability you would get a statistical value, as or more extreme than the one you observed, if the null hypothesis was true (i.e., if nothing interesting was happening, how likely would it be that I would get such an extreme statistics?). A small p-value means there is a very small probability that you would get such an extreme statistic if nothing interest was happening – which is evidence we can reject the null hypothesis. For example, suppose we had a p-value of .01. This would mean that if the null hypothesis was true, we would only get a statistic these extreme 1 out of 100 times. This provides us with evidence that the null hypothesis is unlikely to be true.

Traditionally scientists have said that if a p-value is less than .05 we will call the result "statistically significant" and we will reject the null hypothesis and say there is a different between the populations that we got our data from. By setting such a threshold (of say .05) before we do this analysis, this means that only 5% of the results in the literature would incorrectly say there is a difference in the populations when in fact there is not a difference (although in practice, since not all non-significant results are reported, this percentage could be higher – a problem which is known as the "file draw effect").

A statistical test that is often used to see whether there is a difference between the population means of two sets of data is called a **t-test.** T-tests use the difference between the sample means, the sample standard deviations, and the number of items of each group you're measuring to compute a "t" statistic which can be used to determine the p-value.

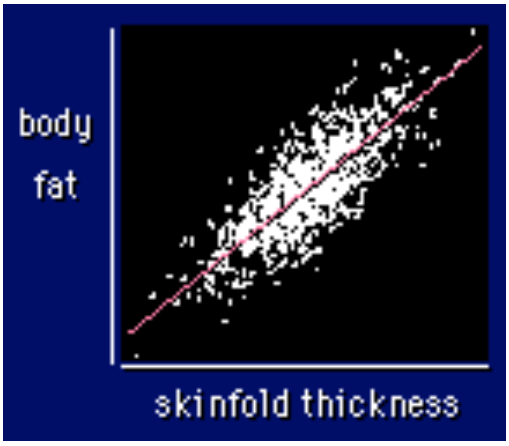
The analysis of variance (ANOVA) is another test that can be used to assess whether the population means of two or more groups are "really" different. ANOVAs give the same results as a t-test when you are comparing only two groups, but you can also use it to analyze whether the population means differ between more than two groups, and for more completed cases when you are interested in simultaneously assessing different types of information (for instance you can compare women vs. men AND veggies vs. meat-eaters both at once). ANOVAs assess whether there are differences in means, by looking at the variance within each group compared to the differences between groups, that's why it's an "analysis of variance." Instead of a "t" value, an "F" statistic is calculated and this is used to determine the p-value.

Correlation

The correlation coefficient ("r") indicates the extent to which the pairs of numbers increase or decrease together. If every data point you collect contains a pair of values (let's call the first value x, and the second value y), we can create a "scatter plot" which plots each data's x and y values as a point on a graph (see figure below). A high correlation value would mean that all these points lie on a straight line, which means that the x and y values increase and decrease together (e.g., if you have a large x value then you would also have a large y value).

The correlation coefficient is a number between -1 and 1 that measures the degree to which two variables are linearly related. If there is perfect linear relationship with positive slope between the two variables, they have a correlation coefficient of "1." If there is positive correlation, whenever one variable has a high value, so does the other; whenever one variable has a low value, so does the other. If there is a perfect linear relationship with negative slope between the two variables, we have a correlation coefficient of "–1." If there is negative correlation, whenever one variable has a high value, the other has a low value (So, for perfect linearity, $r = \pm 1$). A correlation coefficient of "0" means that there is no linear relationship between the variables.

The picture below is of a scatter plot that shows data that have a correlation of +0.9



How big does a correlation have to be before it means something? Statistical tests exist which can give you p-values to determine whether the correlation value measure on the sample data you have reflect an actual difference that you would observe if you had collected all the data from the entire population!