Handy Sheet of Introductory Statistical Knowledge

****Note, I am presenting this information as you would encounter it by speaking to other biologists. If you were to talk stats like this with a statistician they would probably get steam coming out of their ears!!**

This document is organized as follows:
1.1 Types of data (continuous vs categorical)
1.2 Types of variables (independent vs dependent)
1.3 The mean and standard deviation of a sample
1.4 Graphing data
1.5 Hypothesis testing and comparative stats (t-test, chi-squared, linear regression)

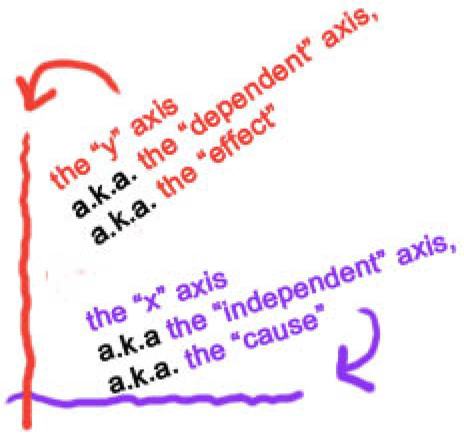
1.1 Types of data

You will probably be dealing with two main types of data - **continuous** and **categorical** data. If your data could literally be any real number (i.e. weight → 23.6, 76.8, 94.6, 33, etc.) then it is a continuous variable. However, if your data consist of discrete categories (i.e. small/medium/large; sick and well, etc.) then it is a categorical variable. Categorical variables can either be **nominal** or **ordinal**. Nominal variables can literally be any non-ordered category whereas ordinal variables have some sort of order (small/medium/large; first/second/third). When looking at the relationship between one variable and another, you may have continuous data split between categories (weights segregated into male vs female).

1.2 Types of variables

When you are looking for the relationship (or dependency) between two variables, one variable will be the **independent** variable and one variable will be the **dependent** variable. An **independent** variable is exactly what it sounds like. It is a variable that stands alone and isn't changed by the other variables you are trying to measure. For example, someone's age might be an independent variable. Other factors (such as what they eat, how much they go to school, how much television they watch) aren't going to change a person's age. In fact, when you are looking for some kind of relationship between variables you are trying to see if the independent variable causes some kind of change in the other variables, or dependent variables.

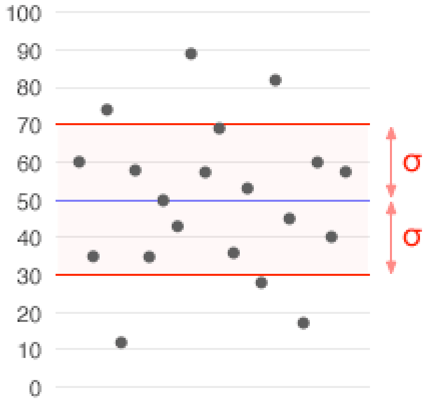
Just like an independent variable, a **dependent** variable is exactly what it sounds like. It is something that depends on other factors. For example, a test score could be a dependent variable because it could change depending on several factors such as how much you studied, how much sleep you got the night before you took the test, or even how hungry you were when you took it. Usually when you are looking for a relationship between two things you are trying to find out what makes the dependent variable change the way it does.



If an experiment compares an experimental **treatment** with a control treatment, then the independent variable (type of treatment) has two levels: experimental and control. If an experiment were comparing five types of diets, then the independent variable (type of diet) would have 5 levels. In general, the number of levels of an independent variable is the number of experimental conditions.

1.3 The Mean and Standard Deviation of a Sample

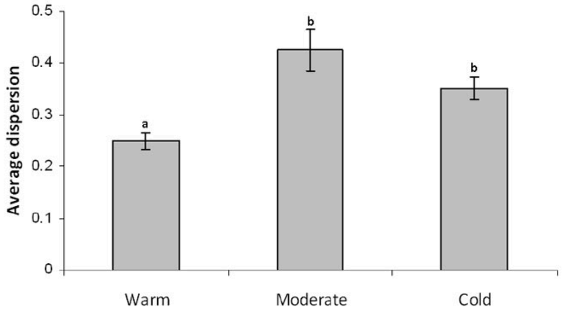
In scientific research we often are trying to get a handle on the descriptive statistics of an entire **population**. However, usually it is not practical to measure the entire population (too many individuals, too large an area, some are hiding, etc.) so we settle for a subset of the total number of individuals (a **sample**). It is handy to be able to summarize our data in a couple values. Two very important summary statistics are **mean** and **standard deviation**. The plot below shows some data that range from 1 to 100.

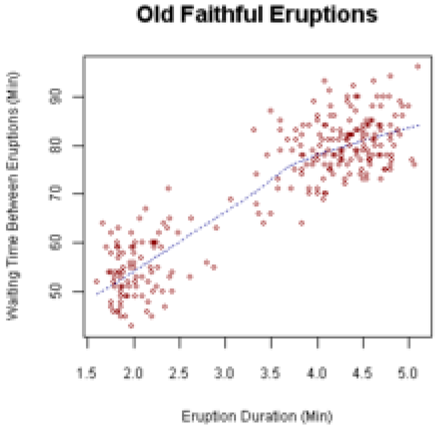


If I want to calculate the mean of these data, I need to take the value of each point, add them all up, and divide the sum by the total number of points. The mean here is represented by the blue line. You notice that the blue line is approximately in the middle of the points. The mean represents the "central" value of your data. Now, the standard deviation (represented here by red lines on either side of the mean) represents the spread in the data. If all the data are close to the mean, then the standard deviation is small. The more data points that are far from the mean, the larger the standard deviation. *Most statistical tests will compute the standard error - learn more about the difference between the standard error and standard deviation.*

1.4 Graphing data

A common way to represent data is with either a **bar plot** or a **scatter plot**. Some data do not require plots at all and are best displayed in data tables. If you find you need a plot, it is very simple to decide between a bar plot and a scatter plot. If you have two continuous variables, a scatter plot is best. If you have one categorical variable and one continuous variable then a bar plot is best. The height of the bars in a bar plot represent the mean of the data in each categories while the error bars represent the standard deviations. Scientists like to say that the "independent" variable goes on the x-axis (the bottom, horizontal one) and the "dependent" variable goes of the y-axis (the left side, vertical one) → you will be expected to also use this convention. This convention applies for both scatter plots and bar plots. Note - barplots are very common but **boxplots** are actually a more appropriate way to display your data in R.





1.5 Hypothesis Testing and Comparative Statistical Tests

When you start out with an experiment you may have an initial research question (Why is the sky blue? How do snakes move? How do trees support their weight?). If you want to use science to investigate this question, then you must turn this question into a **hypothesis**. A hypothesis must be testable and therefore must be more specific than an initial research question. For example, if my initial question was: “Why do plants need fertilizer?”, then I need to learn a little bit about plants and fertilizer to narrow this down to a testable hypothesis that is manageable. I know that fertilizer contains nitrogen and phosphorus and I know that plants need nitrogen to build chloroplasts so that they can photosynthesize. There are many possible hypotheses to spin off of this single question...one might be, “A 20% reduction in nitrogen supply will result in significantly decreased photosynthesis rates compared to the control”. See how this hypothesis is built off of information that I already know?

Now, to determine if my results are “significant” I have to run a comparative statistical test to make sure any perceived differences did not occur just by chance alone. Statistical tests are used to determine if there is REALLY a difference and/or pattern amongst your data. Statistics are the agreed-upon standard for all fields where hypothesis testing is employed. Typically the 3 most basic tests are chi squared, student’s t test, and linear regression). Don’t worry about understanding the specifics of these tests right now. You will be completing your statistics in R - there are MANY tutorials online and on YouTube.

The type of statistical test you will use depends on your specific hypothesis and the types of data you are collecting. The type of test you use will also determine the form of your formal null and alternative hypothesis. Please see this flowchart on how to determine which test to use, what your null and alternative hypotheses should look like, and how to use the p-value to determine if your results are significant.

What is a p-value?

*A p-value means **only** one thing (although it can be phrased in a few different ways), it is: The probability of getting the results you did (or more extreme results) given that the null hypothesis is true. In biology the standard is (AKA we have decided) that $p \leq 0.05$ is a small enough probability to reject the null...*

*****We can NEVER prove or accept a hypothesis...we can support them. It is REALLY important that you understand the difference of this language and know how to use it correctly.**

So, if **$p > 0.05$, we fail to reject the null hypothesis**. (In the language of logic, this is different than stating you accept the hypothesis--*which is incorrect*.) **This means that you have no evidence that they are different.** What you cannot say is that they are the same, similar, or equal--only that you cannot say they are different.

In a jury trial a verdict of "not guilty" is not the same thing as saying someone is innocent. When the jury returns a verdict of "not guilty," they are saying there is a legal insufficiency to find a defendant guilty.

They cannot say the defendant is innocent--*they may even believe the defendant is not*--what they are saying is they "fail" to find the defendant guilty...just like we "fail to reject" the H_o if $p > 0.05$.

On the other hand, if **$p \leq 0.05$, we reject the null hypothesis**. Statistically, **you have supported the alternative hypothesis, but you cannot say that you "accept" it**. There still is a 5% chance (or less, *depending on the actual p-value*) that your results are due to random effects. This means that it is unlikely that your results are due to random effects, but it is still possible (*even though the chance becomes exceedingly more improbable as the p-value decreases to a smaller and smaller number*). **Therefore, you can say that you support the alternative hypothesis**--you have strong (or very strong) statistical evidence that there is a difference between your "treatments" but, you cannot "accept" the H_a, it still is possible *(although unlikely)* that differences are due to random effects.

Does a smaller p-value mean my results are MORE significant?
No! It is all or nothing...the p-value is NOT a measure of "how" significant your difference is.

Technical vs. Biological Significance
The way statistics work, if you have a p-value of 0.06 you *could just collect a bunch more data and get your p-value down to 0.05 and get a statistically significant result. This is basically because the more data you have, the more sure you are of any difference you may have. Therefore, you can discern smaller and smaller differences with larger data sets. However, you have to start asking yourself…"Is this statistical difference biologically significant?". In other words, is the difference so small that in reality it really doesn't matter?

What is the deal with "degrees of freedom"?
For this course don't worry too much about degrees of freedom. It is a way of taking the size of your data set into account (you can ask more questions of a larger data set).

----------------------------further notes---

What kinds of data do I use a chi-squared test for?
This test is used when you have two categorical variables (i.e. gender and color). Remember, you could turn a continuous variable (such as length) into a categorical variable (big vs small)...just make sure the distinction you make between the categories is meaningful.

Chi-Squared Test of Independence

H_0: the relative proportions of one variable are independent of the second variable
H_a: the relative proportions of one variable are not independent of the second variable

Chi-Squared Goodness of Fit Test

H_0: There is no significant difference between the "observed" and "what is expected if the variable had no effect"

H_a: There is a significant difference between the "observed" and "what is expected if the variable had no effect"

What kinds of data do I use a t-test for?
T-tests are used when you have one continuous variable and one categorical variable (i.e. gender and mass). The T-test is really comparing the means of the two groups (male mass vs. female mass).

H_0: There is no significant difference between the means of the two groups

H_a: There is a significant difference between the means of the two groups

What kinds of data do I use a linear regression for?
Linear regression is good when you have 2 continuous variables (mass vs. length). In addition to the p-value there is an R^2 value. This R^2 value is an indication of how good the linear trend line fits your data. The closer it is to one, the better the fit.

H_0: There is no linear relationship between the two variables

H_a: There is a linear relationship between the two variables

What is a paired t-test?
A paired t-test is when you measure an individual, do something to them, and then measure them again. We are NOT doing this in the bean beetle experiment (all individuals measured only once) so you will never do a paired t-test.

One-tailed vs. two-tailed tests?
ALWAYS choose a two-tailed test. That being said, if you are using a significance level of 0.05, a two-tailed test allots half of your alpha to testing the statistical significance in one direction and half of your alpha to testing statistical significance in the other direction. This means that .025 is in each tail of the distribution of your test statistic. When using a two-tailed test, regardless of the direction of the relationship you hypothesize, you are testing for the possibility of the relationship in both directions. For example, we may wish to compare the mean of a sample to a given value *x* using a t-test. Our null hypothesis is that the mean is equal to *x*. A two-tailed test will test both if the mean is significantly greater

than *x* and if the mean significantly less than *x*. The mean is considered significantly different from *x* if the test statistic is in the top 2.5% or bottom 2.5% of its probability distribution, resulting in a p-value less than 0.05.