

A Quick Statistics Primer for FSBio201

While reading research articles, you may have noticed phrases such as "statistically significant", "ANOVA", "t-test", and " $p < 0.05$ ". This handout offers a quick summary of why we do statistics and how to interpret statistical analyses.

Why do we need statistics?

Populations usually show variation in whatever trait or response you might be interested in measuring, and you are typically measuring this trait or response not in the entire population, but rather in a **sample** from this population. Thus, when comparing data from two different groups, researchers are confronted with the problem of determining whether a difference is real or merely due to sampling error because the populations have variation.

For example, compare the following data:

sample 1: 5 2 3 4 5 5

sample 2: 6 5 4 7 5 4

The mean value for the first sample is 4.0, and 5.2 for the second sample. Are the two populations (from which these samples came) different with regard to this trait or response?

One way to think about this data is to set up a **null hypothesis**, or "hypothesis of no difference". Do we have evidence to **reject** this hypothesis and then **posit that there is a difference** between the two groups? Or must we **not reject** this hypothesis, and claim that the evidence does not support a difference between the groups?

Thus, we have a decision to make: to **reject** or **not reject** the null hypothesis. Here's what can happen:

	null hypothesis true	null hypothesis false
not reject null hypothesis	correct decision!!	OOPS! (Type II-error)
reject null hypothesis	OOPS! (Type I-error)	correct decision!!

Statistical tests such as t-test, ANOVA, Chi-square and others are performed on the data to determine how likely it is we are to commit the (Type I) error of (falsely) rejecting a true null hypothesis - that is, we would be claiming a difference between groups, when actually there is no difference (only an apparent difference produced by sampling error). This probability is the infamous **p-value**. We desire a low p-value, because that means that we can (with low risk of being wrong) reject the null hypothesis and claim a difference. Typically, researchers will claim a difference when $p < 0.05$ (although researchers in some fields would want $p < 0.01$ or even $p < 0.001$ - suppose you are researching the effectiveness of a drug that had very nasty side effects?)

A t-test on the data above gives $p = 0.13$, so we would **not** want to claim a difference between the populations based on the samples given.

How to perform a t-test using Excel:

- 1) You will need to have the values for your two variables arranged in two columns, one for each variable.
- 2) Find and select **Tools>Data Analysis...** from the menu. In case you do not see 'Data Analysis' as an option, you will need to first select **Tools>Add-Ins...** and select 'Analysis ToolPak' on the list of Add-Ins and hit **OK**.
- 3) Choose one of the t-tests from the list – generally you will want to choose "Assuming Unequal Variances".
- 4) Select the columns for your two variables, enter '0' as hypothesized difference, and click **OK**.
- 5) The results of the test will be given in a new sheet. Of particular interest is the **P(T<=t) two-tail** value, which is the p-value of the test.

Practice using the data above – you should get $n = 0.13$