Statistical Data Analysis

A. What are the different types of variables?

We have distinguished between *two different types of variables* found in your dataset:

- **continuous** variables—measurements can take on any value along a continuous range
- **categorical** variables—values are categories that can be used to divide the dataset into groups

The appropriate statistical analysis depends on which types of variables are used for testing your prediction.

B. How well does a sample represent a population?

- Any dataset is collected from a *sample* of all possible individuals that could be measured from an entire *population* of individuals.
- Because we usually cannot measure every individual, we use statistical tests to build *inference* on whether or not there is a strong enough relationship between variables in our sample to draw conclusions about a pattern in the entire population.
- We assume, and try to assure by our sampling method, that our sample is *representative* of the population as a whole.

How is a population defined, and what kind of sample is representative of it? The answers depend on the exact question posed.

FOR EXAMPLE: We might ask, "Are male students taller than female students at the College of Charleston?" The population of interest, as defined by the question, is all students currently at the college. If we could measure every student, we could say definitively whether the average male is taller or shorter than the average female. However, we are more likely to measure only a sample from that population. In that case, we could make only a statement about the *probability* that male and female heights differ based on (1) the sample means and (2) an estimate of how well the sample means are likely to represent the population means.

To choose a *representative sample*, we first try to avoid potential *biases*. For example, it would be best to avoid sampling strictly from areas where we expect a disproportionate number of athletes, who might be taller than average. In most cases, a scheme that samples students at random from the population would help to provide a representative sample. Second, we try to choose a sample as large as practical, to avoid the possibility that our sample will provide, by chance, an unusual collection of values from the population.

Given a representative, unbiased sample that is large enough to test a prediction, we can then use *statistics* and inference to draw conclusions about the entire population of students. In other words, we can *generalize* the results of analyzing a sample to an *appropriate* population.

C. What is a *statistical* hypothesis?

***Statistics* provide a formal way of deciding whether a biological prediction about a population is likely supported using a sample of data.**

For example, your bar graphs may have showed at least some difference between groups in the average value. Were these differences, calculated from a *sample*, large enough to conclude that the *populations* actually differ? To answer this question, we use the data to determine how likely it is that the difference between samples was due to *chance* rather than to a *real difference between populations*.

For any statistical test we define two alternative hypotheses:

- ***the null hypothesis*** **(H_o)**: the result expected if there were no relationship between variables (for quantitative traits). Stated another way, that the differences or relationships observed represent random variation between groups or random associations among variables
- ***the alternative hypothesis*** **(H_a):** the result expected if there were a relationship between variables, or the difference in means is too large to be accounted for by random variation among individuals (either a difference between groups or an association between variables)

We assume by default that there is no relationship *until* we have good enough evidence to reject the hypothesis of no relationship. This process reflects the *conservative* nature of science—we do not accept a new, alternative idea unless the evidence is highly convincing. In fact, a typical criterion for "rejecting the null hypothesis" is that the relationship must be so convincingly strong that it should occur by chance (that is, because of a chance sampling of the population) no more than 5% of the time.

Four important notes

1. A statistical test leads to only one of two conclusions: (a) *failure to reject* the null hypothesis, or (b) *rejection* of the null hypothesis *in favor of* the alternative hypothesis. The test does not lead one to accept the null hypothesis nor to prove the null or alternative hypotheses.
2. Rejection of the null hypothesis does not mean that the biological *mechanism* that you described in your prediction is responsible for the relationship between variables. You as a biologist will use inference to make the link between the statistical result and the biological hypothesis. The effect could always be due to some other mechanism you didn't propose.
3. Rejection of the null hypothesis—a statistical outcome—does not necessarily mean that the effect is an important biological outcome. As a biologist, it is still necessary to consider the magnitude of an effect when judging its biological significance. Some weak relationships may actually be biologically meaningful.
4. The words prove and insignificant are not appropriate when describing the outcome of statistical tests. Instead "provide evidence" and "not statistically significant" are correct.

D. How is an appropriate statistical test chosen?

Scientists have access to a large and growing array of statistical tests. Two tests are very commonly used: the **correlation analysis**, the **t-test**. Which test to use depends on whether the variables are continuous or categorical. See APPENDIX A on the last page for details about these tests.

E. How is a statistical test applied?

Regardless of which test is used, the procedure is similar:

(1) Calculate a **test statistic**,
(2) Compare the test statistic to a **critical value**,
(3) Determine a ***P*-value** (or ***P***) based on comparing the test statistic to other critical values based on degrees of freedom, and
(4) Reach a conclusion to reject the null hypothesis only if ***P*** is less than **alpha**.

Here are the details:

- What is a **test statistic**? A single value computed from the data from your sample.
    - For the two tests mentioned above, the names of the test statistics are *r* (correlation analysis) and *t* (t-test).
- What is a **critical value**? A value that can be looked up in a table (or by Excel).
    - Critical values are calculated by statisticians based on the type of test, the **degrees of freedom**, and **alpha**. If your test statistic is greater than the critical value, then the data from which you calculated the test statistic are "extreme," and any relationship you found between two variables is unlikely to be due to chance.
- What are the **degrees of freedom (d.f.)**? A number based on the sample size of the data used (see APPENDIX B for how to calculate for each test).
- What is a ***P*-value**? ***P*** is the probability that a relationship between variables measured *from your sample* is due to chance rather than to an actual relationship *in the population*.
- What is **alpha**? The *upper limit* on the risk you are willing to take that a relationship between variables measured in your *sample* is due to chance rather than to an actual relationship in the *population*.
    - If ***P*** < **alpha**, then the probability of such an error *with your data* is less than the upper limit you set, and you can reject the null hypothesis with confidence. Alpha is typically set equal to 0.05 (a 5% chance of rejecting the null hypothesis when the relationship you found is actually due to chance).

E. Do these tests assume anything about my data?

Yes, but many tests work even with small violations of these assumptions, so we will not worry here about testing them. The kinds of statistical tests you will use make just a few basic assumptions that are worth considering:

- Data for continuous variables are assumed to have a normal (bell-shaped) distribution, with many more measurements close to the average value and progressively fewer measurements at more extreme values. This kind of distribution is typical of many types of data.
- Data points are assumed to be independent of one another. For example, when measuring heights of students at College of Charleston, we assume that a large number of measurements are not taken from any single family, because family members are likely to be similar in height and therefore do not represent independent measures of height.
- In addition, the **t-test** assumes that measurements for the two groups you are comparing have equal standard deviations. If the standard deviations you calculated are not terribly different, you probably meet this assumption. If they are terribly different, a version of the t-test is available that can account for this difference in standard deviations.

F. How do I report statistics in my research report?

- In the results section of your report you will present your test statistics, degrees of freedom (df) and p value in the format ($r=0.9$, $df=59$, $P<0.001$) or ($t=3.91$, $df=20$, $P=0.001$).
- Use the conclusion to evaluate your prediction(s) and to answer the original question posed. This will go in your discussion section of your report.