Courtesy of Dr. Dan Blumstein,
UCLA Center for Behavior, Evolution, and Culture

Dan's Quick-and-Dirty Guide to Statistics

AKA: how to analyze your data without getting intimidated

The bottom line is this: statistics are a powerful tool that you will need to properly apply to ask research questions. Knowing something about them before designing experiments is essential because it will help you design clean and focused hypotheses and collect data suitable for analysis. During our labs we'd like you to begin to develop statistical skills that you'll hone throughout the quarter. The best way to learn statistics is by analyzing your own data. By the end of the quarter you should feel very comfortable with some of these analyses.

Broadly, there are four types of common statistical analyses.

1) Comparing the mean values of two or more groups.
2) Searching for associations or correlations between two continuous variables.
3) Trying to explain variation in a dichotomous dependent variable.
4) Searching for associations between categorical variables.

Some important terms

Before you select a statistic you have to know a few terms. The *dependent variable* is the variable that has a value that depends on something else. For instance, body mass may depend on sex. *Independent variables* explain variation in dependent variables and are often experimentally manipulated or assigned. In this case, sex would be the independent variable. Variables can be *continuous*—that is they can be described, by numbers, with values varying along a continuum. Body mass, measured in grams, is a continuous variable. Variables can be *discrete*—that is they are described by categories and thus, there is a limited range of values that they can have. Sex is a discrete variable. Variables can also be *ordinal*—that is they can be ranked in some way. The results of a race are ordinal; you can't come in 23.124th, but you can come in 23rd place. *Dichotomous* variables are those that take on two mutually-exclusive values. For instance, you can be alive or dead following some treatment. An animal can sing or not sing. It can breed or not breed.

Asking statistical questions

The goal of many statistical analyses is to understand how some categorical treatment influences the expression of some continuous trait. For instance, does taking multi-vitamins lead to an increase in body mass? This deceptively simple question gets a bit more complex when you realize that it may or may not be influenced by sex. That is, multi-vitamins may influence body mass in males differently than in females. In this case, there might be what is called an *interaction* (more on this in class). Thus, your experiment should be designed to include both males and females and your statistics will need to incorporate these two *factors* (i.e., sex and vitamin use). We'll help you with more complex ANOVA designs to make sure you're collecting your data properly and to make sure you analyze your data properly.

Other analyses seek to understand associations between continuous variables. For instance, is brain size *correlated* with body size? If there is a hypothesized 'causal' relationship between your independent and dependent variables, we use *regression* analyses, or more generally *general linear models*. If our dependent variable is dichotomous, then we use *logistic regressions*.

Finally, you might have a series of counts and want to know if say the number of blue and red individuals is independent of sex. In this case you'd apply some sort of *contingency table analysis*.

Normality and non-parametric tests

The common statistics we'll be using often assume normal distributions and homogenous variance. A normal distribution means that when you plot a frequency distribution of your values, the variable should be roughly bell-shaped. Homogenous variance means that the variation for each of the compared groups should be roughly the same. There are formal tests to evaluate these, but some are seriously biased. Plotting and visually examining your data is a very important step in data analysis.

What if your data are not normally distributed? Don't panic yet. First, you might try to transform them. The logic behind transformation is that it's somewhat irrelevant on what scale you measure something. For instance, you could measure mass in grams or kilos. Or, you could measure mass in grams or the natural logarithm of grams. Typical transformations include: log transformations—to eliminate outliers, and angular transformations—to normalize proportional data. If transformations do not 'normalize' or 'homogenize' your data, you might have to use non-parametric tests. These tests do not assume normality and are thus referred to as 'distribution free'.

What is significance?

Finally, how do you know if you've got a significant effect? By tradition, p-values < 0.05 are deemed significant. What this means is that you're comfortable making a mistake with your conclusion 1 in 20 times. Formally, you're comfortable making a 'type-I error' and concluding significance when in fact there is none. However, what happens when you conduct 10 analyses with the same set of data? When you conduct many correlations or conduct many t-tests with the same set of data to look for relationships between variables, you are more likely to find significance simply by chance alone. Thus, you should make it more difficult to achieve significance. A common method is the *Bonferroni correction*—which simply divides your p-critical value of 0.05 by the number of analyses you've run. So, if you conducted 10 correlations, 0.05/10 = 0.005. And, your new critical p-value for concluding significance should be values less than 0.005 are significant.

Selecting a test

All of these analyses can be conducted in StatView. Ask yourself these questions to help select the appropriate test.

<u>1) Is your dependent variable continuous and your independent variable categorical?</u> Are your data normally distributed? Are the variances of your groups similar? Do you have repeated measures on the same individual across treatments?

Based on the answers to the above questions, you'll need to select one of the following tests.

Parametric: t-test, paired t-test, ANOVA, repeated-measures ANOVA

Non-parametric: Mann-Whitney U test; Wilcoxon matched-pairs signed rank test, Kruskal-Wallis non-parametric ANOVA, Friedman non-parametric ANOVA (if repeated measures)

2) Is your dependent variable continuous and your independent variable continuous? Are your data normally distributed? Do you have one or more than one independent variable?

Based on the answers to the above questions, you'll need to select one of the following tests.

Parametric: correlation, regression, general linear modeling (can include multiple independent variables which are both categorical and continouous)

Non-parametric: Spearman rank correlation

3) Do you have a dichotomous dependent variable and a set of continuous and categorical independent variables?

Parametric: logistic regression (there is no non-parametric analog for this test)

4) Do you have counts that allow you to generate a contingency table (e.g., N red males, N red females, N blue males, N blue females)? Are there non-zero values in all the cells of your contingency table?

Chi square, G-test, Fisher exact test (2 x 2 tables only)

Don't forget, "statistics are your friends!"