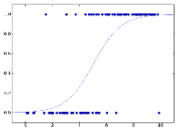
STATISTICS AT-A-GLANCE Rachel Novella

		Independent Variable **(AKA manipulated/predictor variable)**	
		Continuous (Quantitative)	Categorical (Qualitative)
Dependent Variable (AKA response variable)	Continuous (Quantitative)	Statistical Analysis: Regression Graphical Representation: Scatterplot with trendline or connected dots (“line graph”)	Statistical Analysis: ● 2 groups compared: T-test ● 3+ groups compared: ANOVA Graphical Representation: Bar graph of means ± standard error or possibly box plot
	Categorical (Qualitative)	Statistical Analysis: Logistic regression Graphical Representation: A graph of a logistic regression is accurate but not friendly: Consider displaying percentages in a bar graph 	Statistical Analysis: If data are countable, chi square (X^2) test on counts Graphical Representation: If data are countable, graph the counts *If data are not countable, consider revising data plan.*

Adapted from Gotelli, N. J. (2004). *A Primer of Ecological Statistics*. Sunderland, MA: Sinauer Associates.

OVERVIEW OF STATISTICAL TESTS:

General terminology:

- P value – This is the probability that a pattern you conclude is wrong (Type I Error). Scientists usually accept a 5% chance or less of being wrong, so significant difference is accepted as $P<0.05$. The *smaller* the P value, the stronger the pattern/differences!
- Critical value – This is the threshold for a given statistical test to get the P value you’re interested in (usually 0.05 as stated above). Your calculation must exceed this threshold to get a P value smaller than that. Example: X^2 > critical value to get significant $P < .05$

Regression:

The most common is a linear regression (though nonlinear regressions are possible, too, and may be more appropriate for your data). A trendline, or line of best fit, is calculated, and the regression tests how close the data are to the trendline. R^2 calculation describes how closely fitted the data are to the line (or curve if nonlinear). P values still indicate statistically significant relationships between independent and dependent variables, but a P value is given for each term in the line/curve equation that was fitted. (Remember $y=mx+b$? That means you’ll get a P value for m and for b.)

Example: Temperature during egg development correlating to hatchling weight
“Linear regression showed a positive correlation ($R^2 = 0.81$) between egg temperature and hatchling weight; hatchlings weighed more when grown in warmer conditions.”

Logistic Regression:

This is a special nonlinear regression where the curve looks like a wide “S” and is used when your data are binary (male/female, survived/died, purple/white). This regression is based on determining how steep the rise is of the “S” and how closely it matches the data points. A P value below 0.05 will indicate a statistically significant effect of your independent variable on predicting the outcome between the two categories.

Example: Temperature during egg development predicting gender (male/female) of hatchling
“Logistic regression showed that eggs grown at warmer temperatures were significantly more likely to develop into females ($P = 0.02$).”

T-Test / ANOVA:
This test compares means (averages) of the different groups. To compare whether means are *significantly* different, we use standard error as a measure of "spread" of the data for each mean. If $P < 0.05$, we conclude the means are statistically significantly different.

NOTE: An easy way to estimate this (for decent sample sizes, 10+ samples per group) is to graph the means and make error bars to show the standard error (NOT std deviation). If 2x SE bars don't overlap each other, P is definitely less than 0.05 and the means are definitely significantly different. If 1x SE bars do overlap each other, P is definitely greater than 0.05 and the means are definitely NOT significantly different. This leaves a gray area, however, where the 2x bars could overlap but the 1x don't overlap—that's what computers are for. In any case, it's always stronger to report an actual P value instead of just estimating >0.05 or <0.05.

When only two means are being compared, it's called a t-test. The t-statistic is calculated to measure how many standard errors away the two means are and compared against a known table of critical values (developed from the t distribution) based on the number of degrees of freedom in your experiment (number of samples in that mean minus one, for t-test). If your t-statistic is greater than the critical value, then $P<0.05$, and you reject the null hypothesis of no difference and conclude that one mean is significantly higher than the other. If your t-statistic is less than the critical value (did not pass the threshold), then $P>0.05$, and you fail to reject the null hypothesis; there is no significant difference between the two groups' means.

Example: Gender of hatchling predicting hatchling weight
"A t-test showed that male hatchlings weighed significantly more than female hatchlings (P=0.014)."

But what if you're comparing three or more means? Well the problem is that statistical tests assume your samples and comparisons are all independent. If you use a mean more than once for comparison, it isn't really independent anymore, so you have to be more conservative in drawing conclusions. ANOVA (Analysis of Variance) is a test that runs all of the t-tests simultaneously and adjusts the critical value to make it harder to show significant difference (more conservative). Example: If you're comparing means of groups A, B, and C, the ANOVA will simultaneously run t-tests to compare A/B, A/C, and B/C. Since three tests were run, we use a different critical value (called a Bonferroni Correction). ANOVA applies this correction automatically, so if your ANOVA says the P value is less than 0.05, there is a significant difference in the means.

Example: Color of plastic mouse model determines number of predator attacks on a beach (white, brown, black)
"ANOVA showed a significant effect of mouse model color on the rate of predator attacks (P=0.041)."

If you wanted to know *which* groups were significantly different from each other, you'd need to follow up your ANOVA with a Tukey test. This gives you groupings that would let you say whether the mean for group A was significantly different than B and C, but B and C were not significantly different from each other (or any possible combination of differences/non-differences).

Example: Color of plastic mouse model determines number of predator attacks on a beach (white, brown, black)
"A Tukey test showed that at the P=0.05 level, white mouse models were attacked significantly less on the beach than brown or black mouse models, but there was no significant difference in attack rates between brown models and black models."

Chi Square (X^2) Test:
Chi square test compares the observed results against an expected distribution. This is often used by genetics researchers, for example, in comparing allele frequencies in a population against what would be expected at genetic equilibrium. It could also be used to test the offspring of a genetic cross, where a Punnett square gave the expected values for comparison against what is observed. For most experiments, though, the null hypothesis is that there will be no difference between the groups, so the expected values are basically an average of the observed values. You would perform the chi square calculation for each group, as a measure of how different observed was from expected, and sum them to get the chi square statistic (X^2), then compare the experiment's X^2 against a table of critical values. The appropriate critical value is chosen based on the number of degrees of freedom (usually* the number of groups minus one, for X^2 test) and P-value of 0.05. If your X^2 exceeds the critical value, then the data are different enough to be significantly different, $P<0.05$. As always, it's better to report an actual P value instead of just estimating >0.05 or <0.05.

Example: Soil pH (acidic/basic [not tested along numberic range]) causes different flower colors in hydrangea plants
"A chi square test showed that hydrangea plants grown in acidic soil had significantly more blue flowers, and those grown in basic soil pH had significantly more pink flowers (P = 0.003)."

NOTE: The chi square test does NOT work on percentages, only on actual counts.
*NOTE: Special Case: In genetic equilibrium (Hardy-Weinberg) studies, the comparisons are based on genotype, but the degrees of freedom is equal to number of *alleles* minus one. So for most cases, studying one gene that has a dominant allele and a recessive allele, you'll make three comparisons (homo. dom., hetero.s, homo. rec.) but only have *one* DF (not two like normal X^2 rules say).