BASIC STATISTICS GUIDE – BIOL 320 Spring 2016

Mario Muscedere

Imagine you want to test whether men have better eyesight than women. You administer a vision test to all the students in BIOL 300 by showing each student the same sequence of 100 small letters from 20 feet away, asking the student to identify each letter, and recording how many correct responses each student gives. When you have tested all the students in the class and calculated the average number of correct responses for men and women, you get the following result: on average, female students got 80 letters correct, while male students got 75 letters correct.

Are you confident that this difference is evidence that women have better visual acuity than men? How certain are you that this difference is not due to chance? To what degree can you be confident that this difference generalizes to all men and women, not just those in BIOL 300? These questions are faced by all research scientists and have led to the generation of statistical techniques to evaluate scientific results, and a set of accepted conventions on the correct way to interpret these results. This guide will introduce you to some basic statistical concepts that will allow you to interpret your data from BIOL 300 and have a better understanding of the statistical details you read in the primary biological literature.

Statistical hypothesis testing

Scientists first generate **hypotheses** to describe the possible outcomes of their study. In modern science, hypotheses must be evaluated with respect to a **null hypothesis** (**H_0**), which describes the default outcome if there is no experimental effect. For example, a null hypothesis might state that there is no relationship between two variables, or that groups being evaluated have the same average value of some variable. In our hypothetical experiment, the null hypothesis is that men and women have the same visual acuity; that is, men and women on average will get the same number of letters correct on our vision test. The **alternative hypothesis** (**H_a**) asserts that there is a relationship between variables, or a difference between groups. The possibility that the alternative hypothesis is true may be suggested by other observations or previous research and is generally what motivates a researcher to undertake an experiment. In our hypothetical experiment, the alternative hypothesis is that men and women have different visual acuity; that is, men and women will differ in the number of letters they get correct on our vision test.

Factors affecting hypothesis tests

Once data have been gathered from a study, statistical **tests of significance** can be performed to assess the level of support for the alternative hypothesis. There are many such tests, but they are all affected mainly by three factors:

1) As the **magnitude** of the effect predicted by the alternative hypothesis gets *larger*, our confidence in the alternative hypothesis increases. For example, our confidence that women have better visual acuity than men would be greater if we observed a 20-point average difference in their performance (e.g., 80 correct responses on average for women vs. 60 correct responses on average for men) than if we observed a 5-point average difference in their performance (80 correct responses on average for women vs. 75 correct responses on average for men).

2) For any given effect magnitude, as the **variance** of the data gets *smaller*, our confidence

in the alternative hypothesis increases. Variance refers to the variability of the data, such as how spread out the individual data points are around the average value. For example, if all the visual acuity test results fell within a narrow range with no overlap between sexes (e.g., all the women scored between 78 and 82, while all the men scored between 73 and 77), our confidence in an observed 5-point average difference between women and men would be much greater than if there was the same 5-point difference between the averages, but the results were more variable and broadly overlapping between the two sexes (e.g., all the women scored between 68 and 92, and all the men scored between 63 and 87).

3) As the **number of observations** (often called **sample size** and abbreviated as **N**) gets *larger*, our confidence in the alternative hypothesis increases. For example, if we tested 100 men and 100 women, our confidence in an observed 5-point average difference between women and men would be greater than if we observed the same 5-point average difference in a sample of only 10 men and 10 women.

Intuitively, we often focus only on the magnitude of a difference when assessing a pattern (i.e., women on average scored 5 points higher on the visual acuity test than men). However, statistically, we need to know something about the spread of the data and the number of observations to conduct a formal test of significance. The ***p* values** assigned to the outcome of statistical tests depend both on the value of some specific **test statistic** (different for each test) and on the **degrees of freedom** (**d.f.**) of the test, which is closely related to the number of observations. The *p* value represents our level of confidence in the null hypothesis: the lower the *p* value, the more confident we are that the null hypothesis is wrong and, thus, the more evidence we have in favor of the alternative hypothesis.

Statistical significance

Below a certain *p* value, a test is deemed **statistically significant.** This means we are confident that our results are not due to chance, and therefore will **reject the null hypothesis** and treat our results as evidence in favor of the alternative hypothesis. In the life sciences, this threshold value of *p* (called **alpha**) is typically set to 0.05 (5%) by convention. Assuming that all of the assumptions of the test used are met, we should have a false positive (rejecting the null hypothesis when it is really true) only 5% of the time. In other words, we have at least 95% confidence in our rejection of the null hypothesis. If a *p* value is higher than 0.05 the test is not considered statistically significant, meaning we **fail to reject the null hypothesis**. In these cases we do not consider our test as evidence in favor of the alternative hypothesis, because the chance of a false positive is too high (>5%).

Performing statistical tests

There are several free, online statistical packages that can perform many commonly used statistical tests (such as a *t* test). Vassar College provides a good, free online package called VassarStats at http://faculty.vassar.edu/lowry/VassarStats.html. For example, you can perform a *t* test on VassarStats by clicking on "t-Tests &Procedures", then "Two-Sample t-Test for Independent or Correlated Samples". Click the "Independent Samples" button at the top of the page, follow the instructions to import your data, and click the "Calculate" button. The table labeled "*Results*" will now display the *t* statistic (under "t"), the test degrees of freedom (under "df"), and the *p* value of the test (next to "P", in the "two-tailed" box).

Test reporting

t-tests

t statistic, degrees of freedom as a subscript, and p-value. Treatment means ± a measure of variability like SEM or 95% CI are often reported as well, if there are only a few treatments.

> Larval survivorship over 24 hours did not differ between subcolony pairs of young and old workers (paired *t* test: $t_{11} = 0.321$, $P = 0.754$). Larval survivorship was consistently high, with all 10 larvae surviving the duration of most trials and no fewer than 8 larvae surviving in all cases. Larvae tended by old workers, however, gained significantly more mass than larvae tended by young workers (Fig. 1a; old > young: 11/12 subcolony pairs), whether measured on an aggregate ($t_{11} = 3.624$, $P = 0.004$) or per capita basis ($t_{11} = 3.313$, $P = 0.007$). Over 24 hours, the mass gain of larvae tended by old workers (X ± SE) was 20.1 ± 6.34%, (733 ± 215 μg) and larvae lost mass in only 1 of 12 trials. The average mass gain of larvae tended by young workers was only 2.58 ± 2.49%, (41.7 ± 102 μg) and larvae lost mass in 6 of 12 trials.

One-way (aka one-factor) ANOVA

F statistic, degrees of freedom as a subscript[1], and p-value. Treatment means are usually not reported in the text for ANOVA's, because there are always ≥3 treatments and trying to list them all out in paragraph form would be difficult. In fact ANOVA's are a good example of a test that is really much more easily described in a figure than in words.

Among mature workers, which we assume have completed most or all neural development, brain volume relative to body size (estimated as 2×half central brain volume/head width) differed significantly among worker groups (Figure 6; ANOVA: $F_{5,54} = 39.3$, $p<0.0001$).

linear regression

t or F statistic (depends on the program used to calculate the regression – VassarStat uses the t statistic), degrees of freedom as a subscript (one number for t, two for F[1]), p-value, and R^2 value. The t or F statistic and associated p-value assess the *significance* of the relationship; the R^2 value assesses the *strength* of the relationship (how variable are the individual points around the regression line).

> Slow MC fibers were estimated to attain maximum thickness in minors (as indicated by the mean thickness of these fibers in AC4s) on approximately day 11 of adult life, based on a linear model of growth over the first 3 time points measured (regression: $F_{1,22} = 165$, $P < 0.0001$, $R^2 = 0.89$).

ANCOVA

ANCOVA's are similar to ANOVA's with the inclusion of at least one continuous independent variable (often called the "covariate"). They can become quite complicated (like two-way and factorial ANOVAs). But a simple ANCOVA that includes one continuous dependent variable, one categorical independent variable (the "treatment" variable), and one continuous independent variable (covariate) can generally be described by reporting the F statistic, degrees of freedom as a subscript[1], and p-value for each of the two independent variables. Alternately, an ANCOVA can be reported quite economically with a table that presents all the test statistics (see Example 2 below).

[1] F statistics have TWO degrees of freedom associated with them (a "treatment" or "between groups" df which is usually small, and an "error" df which is generally larger) – the reason is not important for our purposes, but they are both listed (in order, separated with a comma) when reporting F.

Example 1[2]
Plasma corticosterone levels consistently were at least threefold higher in female lizards than in males (two-factor ANCOVA with treatment and sex as the factors, SVL as the covariate, and plasma corticosterone levels 1 h post stimulus as the dependent variable: sex F[1,188]=139.46, P<0.0001, all interactions non-significant; Fig. 2A). In contrast, a lizard's body size did not influence its plasma corticosterone levels (two-factor ANCOVA with treatment and sex as the factors, SVL as the covariate, and plasma corticosterone levels 1 h post stimulus as the dependent variable: SVL F[1,188]=0.99, P=0.32, all interactions non-significant).

Example 2[3]

[From the Methods] Among mesocosms with ants, we conducted an ANCOVA to determine if the relative effect of ant biomass on litter mass loss varied with functional category, using generalist foragers and dacetines which were adequately represented for this analysis.

[From the Results] There was no significant difference in mass lost from mesocosms among ant functional categories, in an ANCOVA model with log ant biomass as a covariate (Table 1).

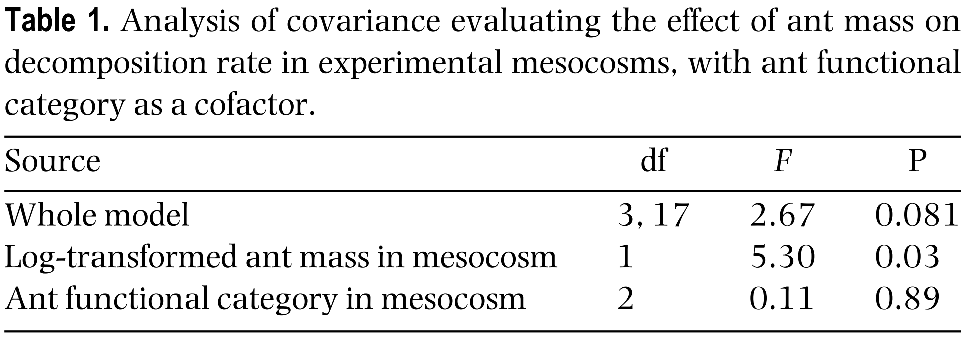

Table 1. Analysis of covariance evaluating the effect of ant mass on decomposition rate in experimental mesocosms, with ant functional category as a cofactor.

Source	df	*F*	P
Whole model	3, 17	2.67	0.081
Log-transformed ant mass in mesocosm	1	5.30	0.03
Ant functional category in mesocosm	2	0.11	0.89

[2] Langlikde & Shine (2006) How much stress do researchers inflict on their study animals? A case study using a scincid lizard, *Eulamprus heatwolei*. Journal of Experimental Biology 209: 1035-1043.
[3] McGlynn & Piorson (2012) Ants accelerate litter decomposition in a Costa Rican lowland tropical rain forest. Journal of Tropical Ecology 28: 437-443.