A critical reader's brief guide to statistics

Gregory S. Gilbert
Environmental Studies, UC Santa Cruz

Statistics are simply formal, mathematical ways to (1) **DESCRIBE** observations about groups of things (e.g., to describe the size of a population of animals), and to (2) **TEST** for trends, patterns, or differences among groups. (e.g., to see if plants grow larger as soils become more fertile (a trend) or to see if more rats died after eating a particular chemical than after eating a placebo (a difference)).

Table 1. Weights of frogs in Walden pond.

Frog #	Weight (g)
1	2.1
2	3.0
3	1.8
4	2.5
5	3.2
6	1.4
7	1.8
.	.
.	.
.	.
57	2.5
58	2.2
59	1.9
60	3.1

Table 2. Frequency of weights of frogs.

Weight class (g)	Num. of frogs
0-0.4	2
0.4-0.8	3
0.8-1.2	7
1.2-1.6	11
1.6-2.0	15
2.0-2.4	10
2.4-2.8	8
2.8-3.2	4

Let's say we want to describe the frog population in Walden Pond. We go out and catch 60 frogs and weigh them (Table 1). Some frogs are bigger, some smaller. We can illustrate the variability in the population looking at the frequency **DISTRIBUTION** of the frog's weights (Table 2) and making a graph (Figure 1). This is usually a frequency graph, or histogram, showing the number of frogs in each weight class (a weight class might be all the frogs from 0-0.4 g, or from 2.0-2.4 g, and so on) distribution, that is, the population will tend to have most of its members clustered around some particular weight, with some a little heavier and some a little lighter, and even fewer much heavier or much lighter. We can describe this central tendency in several ways. One common statistic is the **MEDIAN** - this is the value where one half (50%) of the frogs are heavier and 50% of the frogs are lighter (median =1.8 g). Another common statistic is the **MEAN** (in common usage often called the average). The mean is calculated by adding up the weights of all the frogs and dividing by the number of frogs (mean = 1.75 g). This can be thought of as the weight of the "typical" frog. When the distribution of the population approximates a bell-shaped ("normal") curve, the median and the mean are very similar. Sometimes the distribution has a longer tail in one direction than the other - in these cases the median and mean can be very different. But the weight of the "typical" frog is only part of the useful information. We also want to know how variable the population is. One measure of the spread of the distribution is the **VARIANCE**, which is a measure of how far away the members of the population are from the mean. The square

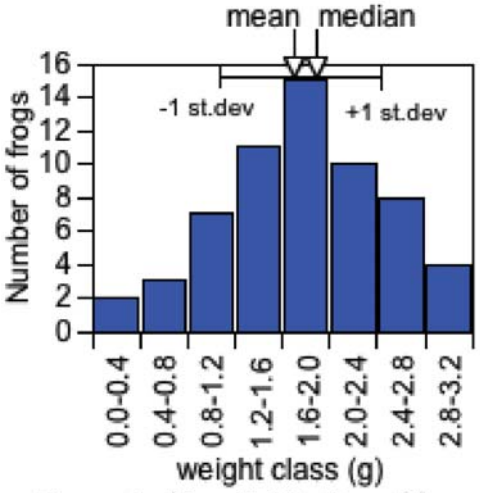

Figure 1. Size distribution of frogs from Walden Pond (N = 60).

root of the variance is called the **STANDARD DEVIATION**, and is very often presented along with the mean (and the sample size, *n*, (number of frogs caught)) to describe the population (mean = 1.75 ± 0.7 g, *n*=60. Translation: Of the 60 frogs measured, the mean frog weighed 1.75 grams, with a variation (standard deviation) around that mean of 0.7 g). The standard deviation tells us something concrete about the distribution: 2/3 of all the frogs will be within one standard deviation of the mean (between 1.05 and 2.45 g), and 95% of all the frogs will be within two standard deviations (within 0.7*2=1.4 g of 1.75g).

Census vs. survey

In the above example, if we had caught all the frogs in the pond and weighed them - it would have been a CENSUS of the pond. However, usually a total census is not possible or desirable. This may be because the study would be too costly, because it would destroy the pond to catch all the frogs, or simply because we don't need to know exactly what the mean frog size is: we just need a good, solid estimate of the mean and the variance. Usually we **SURVEY** a population by taking a **RANDOM SAMPLE** - we measure only part of the population to **ESTIMATE** what the whole looks like. The larger the sample size the closer the approximation of the true mean of the population. It is important the sample be random - otherwise we might **BIAS** the sample toward big frogs (easier to catch) or small frogs (easier to hang onto), or towards calling male frogs.

There are several sources of **ERROR** in the frog study. First, is **measurement error** - we never know exactly how much a frog weighs, and how precise we are depends on our measuring instrument - a bathroom scale is not very precise and leaves a huge measurement error, whereas a $10,000 analytical balance tells us within 0.1 mg how much the frog weighs. The instrument should be chosen to provide us with the minimum necessary precision, without overkill. Measurement error exists in both the census and the survey. In the survey there is an additional source of error, the **sampling error** - this is how far our estimate of the mean frog size is from the true mean of the population (the difference between the mean of all 394 frogs and the 60 randomly sampled frogs). We must design studies to reduce the sampling error to a level adequate to address hypotheses with an acceptable level of certainty. There is a balance between lower sampling error (through larger sample sizes) and the costs of sampling more individuals. Sampling and measurement error should not be confused with natural variability in the population - frogs come in a range of sizes.

Description versus hypothesis testing

Sometimes we just want to describe something, but usually we want to compare two or more groups, or look for a trend in the data. One way is to just eyeball the data - it looks like the frogs in Crystal Pond are bigger than the frogs in Walden Pond (Figure 2) - but eyeballing is fraught with problems of personal bias. Statistics allow us to do formal hypothesis testing through an objective (although somewhat arbitrary) set of rules.

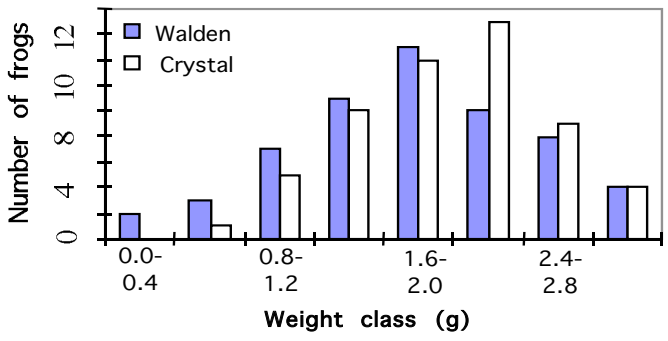

Fig. 2. Number of frogs by weight class in Walden and Crystal Ponds (*N*=60 in each).

A key step in the scientific method is stating a falsifiable **NULL HYPOTHESIS (H_o)**.
H_o: The frogs in Walden Pond are NOT different in size from the frogs in Crystal Pond (their mean weights are equal). Notice

that the null hypothesis is stated in the negative.

The complement of H_o is the **ALTERNATE HYPOTHESIS**, H_a.
H_a: The frogs in Walden Pond are different in size from the frogs in Crystal Pond (their mean weights are different).

To test the null hypothesis, we collect random samples of frogs from each pond, weigh them, and get estimates of the mean and variance for each pond. We might be tempted to say that, well, the frogs in Walden Pond had a mean of 1.75 g, and those in Crystal Pond of 1.83 g, so the sizes are different between the two ponds, and we reject the null hypothesis. HOWEVER, we must think about the sources of error in those estimates of the mean - natural variability in frog size, measurement error, and sampling error - all wrapped up into the standard deviation. The question is really "is the difference between the estimates of the means of the two populations significantly larger than the difference we might expect given the measurable error around our estimates of the means"? In this case we can use a ratio of the difference between the means to the standard deviations of the samples as a test - if the ratio is greater than a particular value (which can be looked up on a table) we reject the null hypothesis of no difference. If the ratio is smaller than that value we can not reject H_o, and we say that there is no difference between the means. This is called a *t*-test, one of many tests in statistics.

By convention we are usually willing to accept a 5% chance (1 in 20 times) that we are wrong in rejecting the null hypothesis. If we reject the null hypothesis (and say that there is a difference between the means) when really there is no difference, this is a **TYPE I ERROR - a false positive**. Depending on the question, it is sometimes wise to accept only a 1% or 0.1% chance of incorrectly rejecting the null hypothesis. Alternatively, perhaps the risk of saying there is no effect when an effect exists is so great that we are willing to accept a false positive 10% of the time. We set the acceptable level of risk of making a Type I error in advance by choosing an "alpha" level - 5% risk is $\alpha = 0.05$, or 1% risk is $\alpha = 0.01$. If the test value (such as the ratio in *t*-test above) is above the critical value for that alpha, we reject the null hypothesis and say that "$P \leq 0.05$", which means that based on the test, there is less than or equal to a 5% chance that we are wrong in rejecting the null hypothesis. In scientific hypothesis testing we say that a result that allows us to reject the null hypothesis is statistically **SIGNIFICANT** at $P \leq 0.05$. Scientists often use the terms "highly significant" for $P \leq 0.01$ and marginally significant for $0.05 \leq P \leq 0.10$, as well. Be cautious of different uses of "significant" - a statistically significant result may not be particularly "significant" in understanding the large picture. Similarly people often say that a discovery is "significant" because of its impact on science or society, without reference to its statistical significance.

On the other hand, if we do not reject the null hypothesis (and say that there is no difference between the means) when there really is a difference (H_o is false), we call this a **TYPE II ERROR - a false negative**.. The **POWER** of a test is our ability to detect a true difference between the two groups, and is represented as Beta , β = probability of a Type II error. If we know a little about the mean and variance of the populations ahead of time we can design a study to give us the Power to detect differences of a desired magnitude, if they exist.

Do <u>not</u> interpret $P \leq 0.05$ as a 5% probability that something will happen - it is a measure of the probability of being wrong in our evaluation of an hypothesis. The probability of an event is something that can be measured in much the same way as frog weights. If we want to know if brand A light bulb is more break resistant that brand B, we could take 100 of each and drop each one to the floor from a height of 1 meter, counting the proportion of bulbs that breaks for each brand. Through this process we might determine that 20/100 =

20% of brand A and 30/100 = 30% of brand B broke. Using probability theory we can then calculate the variance associated with those estimates (a function of sample size and the measured proportion broken) and conduct a test (often a chi-square (χ^2) test) to see if we can reject the null hypothesis that there is no difference in the probability of breakage between the two brands, with 5% chance of being wrong in our assessment. The probability of bulbs breaking, and the probability of us mistakenly evaluating our null hypothesis are two different issues.

Kinds of variables

Sometimes we look at how two variables change with respect to one another, or at how they are **CORRELATED**. Correlation between two variables does NOT mean that one causes the other (e.g., number of crimes in a state is correlated with the number of public schools, but (hopefully) education does not cause crime). When designing studies to look at causal relationships we ask how the potentially causative **INDEPENDENT** variable influences the behavior of the **DEPENDENT** variable (e.g., how does the independent variable "population size" affect the dependent variable "crime rate"?). By tradition, the independent variable is placed on the x (horizontal) axis, and the dependent variable on the y (vertical) axis. How would you figure out if the increase in temperature caused the increase in chemical production, or if it were only correlated with it (Figure 3)?

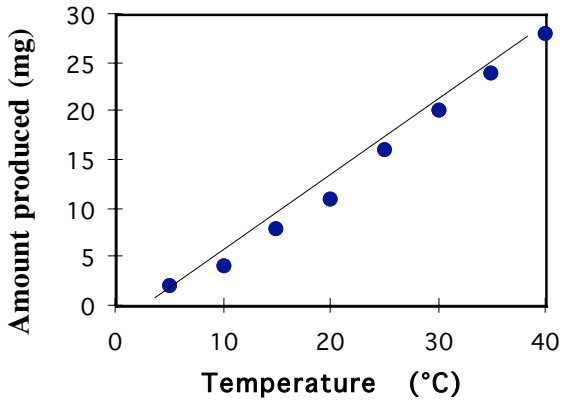

Fig. 3. Chemical production as a function of reaction temperature.

Kinds of data

Data come in two main types, qualitative and quantitative. **QUALITATIVE** data represent states of being. Qualitative data are **NOMINAL** when they indicate what category a sample belongs in (e.g., is it blue, red, or green?). Sometimes qualitative data have a natural order to them, in which case they are called **ORDINAL** (e.g., small, medium, large). **QUANTITATIVE** data are from measurements on a particular scale (e.g., weight of a frog, or number of seeds in a fruit, temperature, or percentage of a population that is sick). Nonparametric statistical tests are appropriate for nominal and ordinal data; parametric tests make assumptions about the distribution of the data, and are appropriate for asking questinos about quantitative data.

Oh, and a note. Datum is the singular of data; therefore "data" is always plural (i.e., "The data were consistent with the prediction that ...." (NOT "the data was consistent"!)).

A cheat sheet for interpreting some common statistics in Environmental Studies

Note: this is NOT a substitute for having a good grasp on statistics, but an aid for when you are a little rusty.

ANOVA Analysis of Variance. Parametric test for comparison of the means of three or more groups (see *F*).

χ^2 or ***Chi square*** Non-parametric test statistic often used for tests of frequency data, such as tests of association between two or more events (contingency tables) or whether two distributions are different. The bigger the better, but *P* depends on the degrees of freedom.

Correlation Measure of association of trends in two variables, how much a change in one variable can be used to predict a change in another; does not necessarily imply a causal relationship (See R^2).

df or **degrees of freedom** Indicates the number of comparisons being tested and the number of replicates in each group.

F Test statistic for comparison of three or more means, such as in an ANOVA; the bigger the better, but *P* depends on the degrees of freedom.

Interaction term In a multiple ANOVA, used to test whether the effects of one factor (say, sex) influences the effect of another factor (say, diet). Denoted as Sex X Diet in the ANOVA table.

N The sample size.

P or ***p*-value** The probability of rejecting the null hypothesis when in really is true; it indicates statistical significance of the test; 0 to1, the smaller the better; by convention $P \leq 0.05$ is considered significant.

R^2 or *r* Correlation coefficient; $R^2 = r^2$; Indicates how well a model fits the data (such as a regression model); R^2 is the better measure of fit, *r* also indicates whether it is a positive or negative association; R^2 ranges 0 to 1, the closer to 1 the better the fit. An R^2=0.65 means the model explains 65% of the variation in the data.

Regression Fits a model to explain trends in the relationship between two or more variables; $y = mx + b$ is a linear regression. Implies a causal relationship where a change in x leads to a change in y, with slope *m* and intercept *b*..

t Student's *t*; a test statistic for comparing two means, or one mean from a particular value; the bigger the better, but *P* depends on the degrees of freedom.

Dichotomous key for choosing statistical tests

This key is a guide to help you select an appropriate statistical test. Consult a statistics text for more details.

1a. Question about associations among variables . . . 2
1b. Question about differences among groups . . . 5

2a. The variables vary together but one does not depend on the other 3
2b. The variables vary together, and the value of one is supposed to depend on the value of the other (includes a prediction or expectation of cause-effect) . . . 4

3a. Parametric . . . . . **Pearson Correlation**
3b. Nonparametric . . . **Spearman Rank Correlation** or **Kendal's Correlation**

4a. Parametric . . . . **Linear Regression**
4b. Nonparametric . . . Not available

5a. Question about differences between frequency distributions . . . 6
5b. Question about differences between means or variances 7

6a. Comparison between an observed frequency distribution and a theoretical distribution . . . **Chi-squared (χ^2) or G test, "Goodness of Fit"**
6b. Comparison between two or more frequency distributions . . . **Chi-squared (χ^2) or G test of independence**

7a. Question about differences between sample variances . . . 8
7b. Question about differences between sample means . . . 9

8a. Question about differences between two variances . . . . . **F test**
8b. Question about differences among three or more variances. . . **Bartlett Test**

9a. Question about difference between two means 10
9b. Question about difference among three or more means . . . 13

10a. Parametric 11
10b.Nonparametric . . . 12

11a. The two samples are grouped in natural pairs . . . **Paired t test**
11b. The two samples are independent . . . . **Non-paired t-test**

12a. The two samples are grouped in natural pairs . . . **Wilcoxon signed-rank test**
12b. The two samples are independent . . . **Mann-Whitney U** or **Wicoxon tests**

13a. Parametric . . . . **Analysis of Variance (ANOVA)**
13b. Nonparamtric . . . **Kruskal-Wallis Test**