13

Introduction to statistics

INTRODUCTION

The first step toward understanding statistics is to have a firm grasp on graphing and sampling. If one can properly differentiate between the need for a bar graph (when comparing means), versus a best-fit line graph (when looking for a relationship between variables), versus having no graph but a tally table instead, one is on their way to understanding statistics. Statistics is using mathematical formulas, definitions, and computers to predict, define, and tell exactly how one treatment is different from another (Magnusson and Mourao 2004). Statistical inference uses standardized criteria for decision making to help ensure that decisions are not swayed by personal opinion or political pressure (Sinclair et al. 2006).

Statistics may not be easy for beginners (Magnusson and Mourao 2004). One problem is that many statistics courses are taught by mathematicians and their job is to emphasize theory. Things may get too deep, too fast, and never cover examples in one's field. Another problem is that practitioners, the biologists, chemists, and physicists who advise you, may have years of experience and tricks using the techniques they need, but their theory is absent and their language is inconsistent. How can you expect to learn anything from anybody?

A good place to start may be the first page of each chapter of a statistics book. With practice, one can peel through to the next layer with more complex graphs, experiments, and examples. The perfect class, book, or teacher will never emerge for something so personal. Instead, there are books at the library that must be read. Once the basic concepts are learned, the results section of science journals and experimental design books should be read to find specific examples that are of interest. Even if one has taken a statistics course, experimental design may not have been addressed as it is in this chapter.

THE NULL HYPOTHESIS

For every statistical test there is a null hypothesis. **"Null" means** no, nothing, none, nada, zip, or zilch. Depending on the type of experiment conducted, a **null hypothesis** means there is (Table 13.1):

- no difference between means being compared (t-test and ANOVA),
- no difference between observed frequencies and those expected by chance (chi-square),
- no relationship between two variables (correlation and regression).

Corresponding to each of these, the alternative hypothesis is the opposite:

- There is a significant difference between means being compared.
- There is a significant difference between observed frequencies and those expected by chance.
- There is a significant relationship between two variables.

THE PROBLEMS WITH NULL HYPOTHESES

Statistical tests are effective at ruling out null hypotheses, for example, "animals do not move." The trouble for most students is that the null

DOI: 10.1201/9781003271833-13

Table 13.1 Most commonly used statistical tests and their null hypotheses

	t-Test	ANOVA	Chi-squared
What is being compared	Two means	More than two means	Two or more frequencies
Null hypothesis	H_o: $\mu_1=\mu_2$	H_o: $\mu_1=\mu_2=\mu_3$	$freq_1=freq_2$
Type of variable	Continuous	Continuous	Categorical

hypothesis is the opposite of what we would expect. Additionally, the alternative hypothesis is what we are supposed to infer if we reject the null, but this is only inference. There may be more than one alternative hypothesis. Which one are we supposed to choose?

P value

When testing a null hypothesis, the result of the statistical test is a P value. It shows whether the null should be rejected. P has a complex definition, but beginners can think of it as the probability that the null is true.

- if $P<0.05$, reject it. The data supports the alternative hypothesis and one can conclude the means are significantly different.
- if $P>0.05$, fail to reject. There is not enough support for the alternative hypothesis.

When $P<0.05$, beginners could think of the meaning as less than a 5% chance that the null is true. A more accurate way to express it is, “a difference as great as what we found between treatments would be expected less than 5 out of 100 times” (Gotelli and Ellison 2014).

WHY DO WE USE P=0.05 AS THE CUTOFF POINT?

This cutoff seems stringent if we reject only when we would expect it less than 5% of the time. If you used this rule in your everyday life, you would not take an umbrella unless the forecast for rain were at least 95% (Gotelli and Ellison 2014). It means that the evidence must be exceedingly strong for us to reject the null hypothesis. The jury does not issue a guilty verdict unless there is more than 94% surety. We would certainly take precautions if we knew there was a 94% chance of a tornado. The reason the standard is so high is because:

- the convention is based on probability, not certainty. We do not measure whole populations, only samples. The estimate based on sampling is sometimes wrong and noisy. We need to be conservative, which means a high standard.
- two types of errors may occur as illustrated in Table 13.2. In a **type I** (alpha error) the null is rejected when it should not be. In a **type II** (beta error) we fail to reject when we should have rejected.
- A type I error is the worse type because it is a false positive. The researcher has rejected the null hypothesis when it was really true. More significant differences were declared than were actually there. It is like convicting an innocent person for a crime he or she did not commit.
- The problem is, the higher P you choose as your critical value, the more you increase your chance of making a **type I error**.
    - If you use a P value too low, you increase the chance of making a type II error. You let a guilty person go scot-free, perhaps to commit another crime.

Experience has shown that P=0.05 is usually the right balance between type I and II errors for most situations as long as at least 30 replicates in each treatment were taken.

The extent that a statistical test minimizes type II errors is called **power**. The power of the test increases with sample size. For almost all tests, there is sufficient power when there are at least 30 replicates in each treatment. In some situations, especially epidemiology, avoidance of type II may be more important. Ask your research adviser.

Table 13.2 Delineation of type I and II errors

	Analysis indicated that we should fail to reject H_0	Analysis indicated that we should fail to reject H_0
H_0 *true*=in reality there is no difference $\mu_1=\mu_2$	Our analysis is correct	Type I error (alpha)
H_0 *false*=in reality there is a difference $\mu_1 \neq \mu_2$	Type II error (beta)	Our analysis is correct

USING CERTAIN WORDS

- Be careful about using "prove." Just because $P<0.05$ and you reject the null, it does not prove the null hypothesis is false.
- Be careful about using "accept the null hypothesis." It is equivalent to "proving." You may "reject" the null or "fail to reject," but you may not "accept the null" unless you do a power analysis to precisely calculate the probability of a type II error.

AN OBSESSION WITH REJECTION–STATISTICAL VERSUS SCIENTIFIC SIGNIFICANCE

A preoccupation with null rejection can overshadow more important concerns (Sinclair 2006). A data set with a small number of replicates, or a faulty null **hypothesis** in the first place, should always be treated with caution or suspicion. Statistical significance does not necessarily imply scientific significance.

MEANS COMPARISONS

A **t-test** is used to determine if there is a statistical difference between two means. Computer software from R, **Excel**, **Instat**, or others can be used to calculate t. A difference exists between independent and paired t-tests. Paired t-tests have more power to detect change and should be used whenever appropriate, but most comparisons are independent and not paired. **Paired t-tests** can be used when pairs of sampling units are correlated. In other words, a plot in a forest with a large value the first year is likely to have a large value the second year, which means they are correlated. Other examples include the right side of bilateral animals compared to the left, or comparisons between identical twins. If you are not sure whether a **t-test** should be paired, it probably should not. Proceed as if it should not. A t-test can also be either one-tailed or two. If you are not sure, assume it is two-tailed.

PARAMETRIC VERSUS NON-PARAMETRIC TESTS

The most commonly used and most accurate statistical tests are based on specified distributions in their histograms (especially normal or Poisson). When they are based on these known distributions they are considered **parametric** tests. These include **t-test, chi-square,** analysis of **variance**, and others. They have certain rules (assumptions) that must be met, otherwise the results are invalid. The assumptions of a **t-test** are:

1. **data are normally distributed**: In other words, if a frequency histogram was drawn for each treatment, each histogram would form a bell-shaped curve (Figure 13.1). This will almost always be true if there is a sample size of 30 replicates for each treatment because of the Central Limit Theorem.

2. **the variance (or SD) of one treatment is approximately equal to the variance (or SD) of the other treatment**: This is the more important of the first two assumptions.

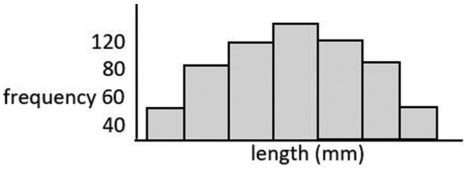

Figure 13.1 A bell-shaped curve in a frequency **histogram** showing a normal (Gaussian) curve. Each bar represents a range of values, with most values grouped tightly around the mean.

Variance refers to how much variation there is among the samples in one treatment.

3. **observations are independent**: One replicate does not in any way influence the value in another replicate and there is true replication, not pseudoreplication.

WHICH IS THE BEST CHOICE, PARAMETRIC OR NON-PARAMETRIC?

Despite the restrictions, you should always try to use parametric tests if possible. Non-parametric should be used only if a correction or transformation does not meet the assumptions. Parametric tests are far more accurate and do the most to reduce both type I and II errors If you use a non-parametric test you should justify your use of it, otherwise we may think poorly of your choice of statistical test.

WHAT IF THE FIRST TWO ASSUMPTIONS OF THE T-TEST ARE NOT MET?

There are two choices.

1. Some statistical packages make corrections for unequal variances such as using the Welch correction within Instat, but you have to say yes for this choice when prompted. Most statistical software does not provide the option.
2. Each number in the data sets can be transformed by ln x, sqr root of x, or ln (x+0.1). This is like changing units from miles to km. The natural log transformation is the most common and has the particular effect of making the standard deviations relatively smaller and therefore more equal. This seems like a magic trick, but it is as valid as changing units from miles to km. Note that when dealing with percentages, the first two assumptions are almost always violated. The best transformation to use with percentages is to take the arcsin square root of each of the data points in each treatment. Although you may transform the data for analysis, you should report the results in the original units when making graphs or tables.
3. If the problem is not corrected when the test is run again on the transformed data, the researcher should use a **non-parametric test**. These have less stringent assumptions. **Parametric** means that the probability fits a specific distribution, almost always implying a bell-shaped (normal) curve. Non-parametric tests are usually based on ranks, a far less accurate way to assess differences. The non-parametric alternative for comparing two means is a Wilcoxon signed rank test or Mann-Whitney U test. The alternative for comparing more than two means is a Mann-Whitney U test or Kruskal-Wallis.

WHAT IF THE RESEARCHER IS COMPARING MORE THAN TWO MEANS?

An analysis of variance (ANOVA) using an *F* test is employed when comparing the means of more than two treatments (Figure 13.2). The basis of an ANOVA is different than of a t-test. In an F test, F is the ratio of the variance among groups over the variance within. The idea is that if the variance among is much greater than within, the treatments must be significantly different. This makes intuitive sense and is why this is called analysis of variance.

To do this in practice, the sum of the squared deviations among treatments $\sum\left(\bar{x}-x_1\right)^2$ is divided by the sum of the squared deviations within treatments $\sum\left(\bar{x}-x_1\right)^2$to produce an F value. This F value is checked against a table of values to determine P. Simple.

The computer output for an ANOVA is in a format as in Table 13.3 (Dytham 2011). The business end of the table, of course, is the P value. In this case P>0.05, which indicates there is no significant difference among treatment means. We

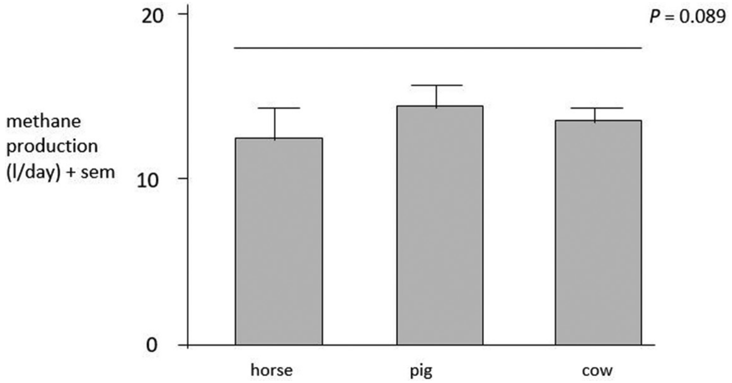

Figure 13.2 Methane production by type of farm animal (n=5) when fed standard diets administered for 6 weeks.

Table 13.3 An example of a standard ANOVA table produced in a standard computer output

Levene Test for Homogeneity of Variances					
Statistic	**df1**	**df2**	**2-tail Sig.**		
0.7616	1	8	0.408		
Analysis of Variance					
Source	**D.F.**	**Sum of squares**	**Mean Squares**	**F Ratio**	**P**
Between	1	16.3840	16.3840	15.3624	0.089
Within	8	8.5320	8.5320	1.0665	
Total	9	24.9160			

HOW TO TEST THE T-TEST ASSUMPTIONS

ASSUMPTION 1 (NORMALITY ASSUMPTION)

1. Some statistics packages check this automatically and will not let you proceed if the assumption has been violated. Some of the other statistical packages test it on request or provide results automatically with the final P value, but they do not stop you from proceeding to the end.
2. If you have 30 samples or more, assume the assumption is met.
3. Seat of the pants rule if you do not know another way: when there are less than 30 replicates per treatment, use a number line to examine the mean and individual values for each treatment. If the individual values are evenly distributed around the mean, assume the distribution is normal for that treatment.

ASSUMPTION 2 (EQUAL VARIANCE ASSUMPTION)

1. Some statistics packages check this automatically. In other software, check the results of the Levene test or other computerized test on the printout for results of the equal variance assumption.

2. **Seat of the pants rule if you do not know another way**: calculate the variance or standard deviation for each treatment. (Excel can do this.) If the standard deviations of the treatments are roughly equal (within 30%), consider the assumption to be met.

ASSUMPTION 3 (INDEPENDENCE ASSUMPTION)

There is no computer program that can tell you this. Use common sense and judgment. Was there true replication, or was it pseudoreplicated? Do the samples in one treatment or replicate influence the others? If this assumption is not met, the analysis must be abandoned. There is no choice.

fail to reject the null hypothesis. Note that when we refer to differences between two means we use "between." When we refer to differences among more than two means we use "among." We must check the assumptions of ANOVA just as we did in the t-test, and the assumptions are the same. For some statistics software, the assumptions are checked automatically. The computer output will now include something like Table 13.3 with a Levene's test:

The "2-tail Sig." value is the P value for the Levene's test. It tells us there is no significant difference in the variances among treatments, thus we have met the equal variance assumption. This is good and we can proceed to validly accept the information in the analysis of variance table. If the P value for the Levene's test were less than 0.05, we would have to take corrective action such as transforming the data. It would not be valid for us to accept the rest of the results in the analysis of variance table.

POST HOC TESTS TO COMPARE PAIRS OF MEANS

If the P value for our ANOVA results tells us that some treatment means are different from others, how do we know which mean is different from which when we have more than two means? To determine differences between means we need a **post hoc test**. Post hoc = after the fact. This is also sometimes called ***a posteriori*** = after the fact. This is also called a **means-comparison test**. These compare every mean to every other mean and provide P values for every pairwise comparison.

The one thing that is not valid is to complete multiple t-tests to compare every mean to every other mean. This compounds the probabilities and renders the tests invalid. In other words, we are no longer testing our null at the 0.05 level. If we conduct two t-tests, we are now testing the null at the 0.05 × 0.05 = 0.025 level. If we use three t-tests, it is at the 0.05 × 0.05 × 0.05 level and so forth. Reducing our *P* standard will drive up our chance of making a type II error.

There are several **post hoc tests** available, Least Significant Difference (LSD), Student Newman Keuls (SNK), Sheffe, Tukey's, and Duncan Multiple Range. Mathematical research shows that the best one to use is Tukey's. Here is why: LSD uses multiple t-tests, which we just reported was invalid; it produces a result that is too conservative = too many type IIs. Duncan's Multiple Range is too liberal with too many type I = worst kind of error. Thus, Duncan's Multiple Range test should never be used. It means too many innocent people in jail. Tukey's has fewest type II or type I errors – use this.

The computer output with Levene's test, the ANOVA table, and the post hoc test is in Table 13.4.

According to our Tukey's test (Table 13.4) there are no significant differences between any of our pairwise comparisons among means because none of the P values were less than 0.05. Actually, we would probably not have run the post hoc tests on this analysis in the first place because we did not find a significant P value when the overall ANOVA was run.

HOW DO I SIGNIFY PAIRWISE SIGNIFICANT DIFFERENCES ON MY GRAPH?

Look again at the figure in our methane example. There is a horizontal line over all three bars. The conventional rule is that this horizontal line is placed over the top of all treatments that are not significantly different. In this case the original P = 0.089. Because it was not less than 0.05, there

Table 13.4 Example of standard ANOVA table with output for Levene test and post hoc test

Levene Test for Homogeneity of Variances			
Statistic	**df1**	**df2**	**2-tail Sig.**
0.7616	1	8	0.408

Analysis of Variance					
Source	**D.F.**	**Sum of squares**	**Mean Squares**	**F Ratio**	**P**
Between	1	16.3840	16.3840	15.3624	0.089
Within	8	8.5320	8.5320	1.0665	
Total	9	24.9160			

Tukey-Kramer Pairwise Comparisons 2-tail Sig.			
	horse	pig	cow
horse	0.00		
pig	0.072	0.00	
cow	0.099	0.124	0.00

were no significant differences. A line can be drawn over all three treatments.

Consider an example when there is a significant difference among the means (Figure 13.4). The mean for horse is significantly different from the means for pig and cow, but pig and cow are not significantly different from one another. Notice that the graph has a line over the treatments that are not significantly different. Notice that to make this convention work, the means must be placed in order from lowest to highest, or highest to lowest on the graph, thus the order of the bars has been rearranged in this example.

The convention of placing a line above the treatments that are not significantly different can be used for some very sophisticated differences. In Figure 13.3, the lines tell us the mean for horse is significantly different from every other mean. The mean for snake is different from every other mean. Cow and pig are not different from each other. Pig, hamster, and lion are not different from each other. Remember, this convention will only work when you order your means from lowest to highest or highest to lowest. Altogether there are 11 significant pairwise differences depicted in Figure 13.3. Can you name them all?

TABULAR COMPARISONS: COMPARING FREQUENCIES THROUGH CHI-SQUARE TEST

It is appropriate to use a chi-square test when:

- frequencies are being compared, not means.
- categorical data are being used, not continuous data.
- the null is that the observed and expected frequencies are not different.

Frequencies are not means. They are the number of each organism or object and can only occur as whole numbers, not decimals as is possible for a mean. It is how many times a coin lands heads or tails, how many individual beetles reproduce or not, or whether organisms are present or not. It is the abundance of something, a unitless number, the number of times something occurs. You keep a tally when assessing frequency.

EXAMPLE HYPOTHESES TESTED IN CHI-SQUARE

- Is the observed and expected frequency the same?
- Are phenotypic ratios in a monohybrid cross the same as the expected 3:1 frequency?
- Are sex ratios the same as what we would expect?

Comparisons among frequencies are useful in genetics and some chemistry research, but with few exceptions they are rarely appropriate for answering field biology questions (Magnusson and Mourao 2004). Too many individuals or plots have to be sampled to record something like presence

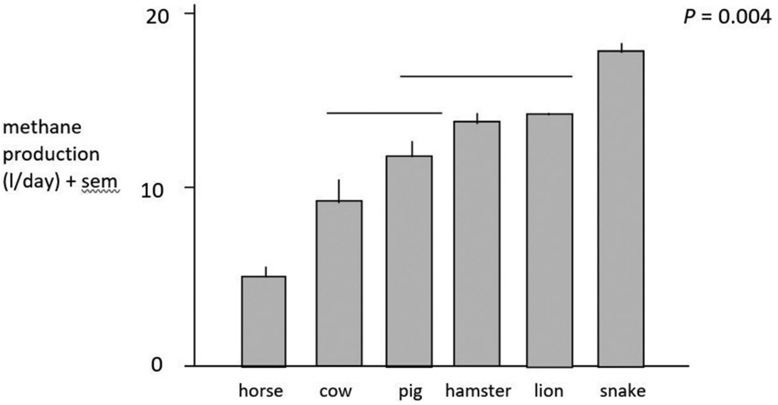

Figure 13.3 Methane production by type of farm animal (n=5) when fed standard diets administered for 6 weeks. Lines above bars indicate no significant difference in pairwise comparisons.

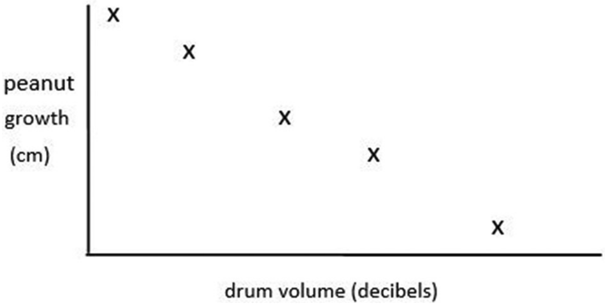

Figure 13.4 Peanut growth versus drum volume showing negative correlation.

and absence. It ends up being too expensive and time consuming to visit 1,000 Chestnut trees and determine whether they do or do not have reproductive structures. A thousand test tubes, however, may be reasonable.

Ecologists often find means and the variation among samples to be more appropriate and more enlightening for small sample sizes. Chi-squared tests are very sensitive. They almost always show statistical significance to the point where the results become meaningless. If you are doing an ecological project and you find yourself using frequencies, see if you can turn your question into something using a mean. This is almost always possible. Leave the chi-squared tests for the laboratory researchers with many more replicates.

CORRELATION AND REGRESSION

For both correlation and regression we usually draw a graph with two axes, and plot points. This is called a scatter plot (Figure 13.4) or if it has a line with it, a line graph. If the two variables are correlated, the data set will slope either one way or the other, positive correlation or negative correlation.

- In a graph plotting drum volume versus peanut growth, we interpret a positive correlation as "the greater the volume, the greater the peanut growth."
- We interpret a negative correlation as "the less the volume, the greater the peanut growth" or "the more the peanut growth, the less the volume."

Table 13.5 Statistical tests available for comparing samples

Purpose	Parametric test	Non-parametric test
Comparing Frequencies from Two or More Treatments, categorical, not continuous data		Chi-square test OR G tests OR Fisher's exact test OR Cochran-Mantel-Haenszel
Comparing Means from Two Treatments, Continuous, not categorical data		
• Samples independent	Independent sample t-test	Mann-Whitney U test
• Samples paired	Paired t-test	Wilcoxon's signed rank test
Comparing Means from More than Two Treatments, Continuous, not categorical data	Analysis of variance with Tukey's – could add Welch's correction for unequal variance A post hoc means-comparison test should follow a significant ANOVA	Kruskal-Wallis test OR Mann-Whitney U test with Bonferroni correction A Post hoc means-comparison test should follow a significant procedure
Comparing Means from More than Two Treatments, one factor involved		
• Samples independent as in independent t-test	Analysis of variance for independent samples	
• Samples blocked as in paired t-test	Analysis of variance with blocking	
• Replicates nested at different spatial scales	Nested analysis of variance	
• Sampling is repeated on the same replicates over time	Repeated measures analysis of variance	

(Continued)

Table 13.5 (*Continued*) Statistical tests available for comparing samples

Purpose	Parametric test	Non-parametric test
• Only one replicate	Before-after controlled impact (BACI) design	
Comparing Means from More than Two Treatments, Two Factors Involved		
	Two-way analysis of variance	
• All treatments fully crossed	Factorial analysis of variance, also called orthogonal design	
• Only one replicate for each treatment	Experimental regression analyzed using regression, not ANOVA	
Correlation between Two or More Variables		
• Neither variable is clearly independent or dependent	Correlation analysis	
One variable is clearly the dependent variable	Regression analysis	
• Two or more independent variables	Multiple regression analysis	
• Two or more independent variables and two or more dependent variables	Multivariate analysis	
• Independent variable is categorical rather than continuous	Logistic regression	

HOW ARE REGRESSION AND CORRELATION DIFFERENT?

Correlation is plotting two variables and looking for a pattern. The researcher did not specify which variable is on the x axis and which is on the y. There is no predictor and response variable, no cause and effect. Regression does specify cause (independent (x) variable) and effect (dependent (y) variable) because the researcher knows which is causing which.

A best-fit line can be added to correlation or regression. This is done to minimize the average distance of the points from the line and can be done mathematically or by eyeball. For linear regression, the squared vertical distances from a line are generally minimized (Magnusson and Mourao 2004). For correlation, the horizontal and vertical distance of each point is minimized and this is called least-squares (also referred to as Model II regression) (Gotelli and Ellison 2014).

More advanced students may be interested to know that least-squares regression is logically and mathematically the very same as ANOVA. The distances to the best-fit line are "residuals." The variation about the line is the residual variation not explained by our model (the line). A plot of residuals after an ANOVA can be enlightening to establish how much variation exists within or between treatments. It is also used to calculate r^2.

SOME OTHER DIFFERENCES BETWEEN REGRESSION AND CORRELATION

In regression one can draw a line in the form y=mx+b to predict y values based on x. A prediction is not appropriate in correlation. In regression one can obtain a P value and a significance test. It tests the null that one variable does not depend on the other. In other words, **P** is the probability that the best-fit line has a slope of zero. In correlation this is not appropriate.

THE STRENGTH OF THE CORRELATION CAN BE MEASURED

Strength is signified by **r** which stands for the Pearson's product-moment correlation. It varies from –1 (negative) to +1 (positive), with 0 indicating no correlation at all. It represents the percent variation in one variable explained by the other. In other words, it represents consistency in the data. We can ask, is there a high degree of error, or do the data points make a straight line? We can also calculate a P value that tests the H_0: one variable is not related with the other. It tells us something about the strength of relationship but not the slope.

WORDS OF CAUTION ABOUT CORRELATION

Pearson's correlation assumes that both variables are normally distributed. This assumption is almost never heeded. Correlations are frequently used, but do not report their statistical assumptions. Even if a correlation exists, it does not imply cause and effect (one variable causes the other). Consider "the higher the drum volume, the greater the peanut growth." All a correlation can do is establish a possible pattern but nothing else (Table 13.5).

REFERENCES

Dytham, C. 2011. *Choosing and Using Statistics*. Blackwell Science, Malden, MA.

Gotelli, N.J., and A.M. Ellison. 2014. *A Primer of Ecological Statistics*. Sinauer, Sunderland, MA.

Magnusson, W.E., and G. Mourao. 2004. *Statistics without Math*. Sinauer, Sunderland, MA.

Sinclair, A.R.E., J.M. Fryxell, and G. Caughley. 2006. *Wildlife Ecology, Conservation, and Management*. Wiley-Blackwell, Hoboken, NJ.