A Statistical Primer

This primer will briefly answer the following questions. You should read it in advance, bring questions about it to class, and consult it as you complete the statistical analysis worksheet.

A. What is a population, what is a sample, and why do we need statistics?
B. What is a *statistical* hypothesis?
C. What does a statistical test actually do?
D. How is a statistical test carried out?
E. Do these tests assume anything about my data?
F. What are the possible outcomes of the test?
G. How are statistical results reported?
H. Why does P give the probability of making an error, and why do we set $\alpha = 0.05$?
I. What are one-tailed and two-tailed tests?

A. What is a population, what is a sample, and why do we need statistics?

In research, we want to say something definitive about a ***population***, but we have only a ***sample*** of the population. Statistics provide a quantitative way to express confidence about a conclusion when we have only limited information from a sample.

As a first step, we try to make sure by our sampling method that our sample is ***representative*** of the population as a whole. But what is "the population," and what kind of sample is representative of it? The answers depend on the question. As an example, we might ask "do biology majors score higher than chemistry majors on their first organic chemistry exam?" The "population" could include any and all biology or chemistry majors that are currently taking, have taken, or could take organic chemistry. We could restrict the population of interest in various ways, for example, to students in the USA, or students at CofC, or students in their sophomore year, or students taking the course this year. If we could measure every one of the students in that population, we could say definitively (without statistics) whether there is an average difference in scores between students in the two majors. It is more realistic, however, to measure representative samples. Then we can make a statement about the ***probability*** that the populations of the two types of majors differ based on (1) the sample means and (2) an estimate of confidence that our sample means represent the population means.

To choose a representative sample, we first try to avoid the potential for ***bias***. In most cases, sampling **at random** from *a population* helps to provide an unbiased sample *of that population*. Students sampled at random from CofC, for example, would be an unbiased sample of students at CofC but could be a biased sample of students from the USA (if, for example, one of our programs were stronger than the other in an atypical way). Second, we try to avoid the potential for statistical ***noise*** (or "sampling error") by choosing a sample **as large as practical**, to reduce the possibility of getting a set of values in our sample that is atypical of the population. Avoiding ***bias*** and ***noise*** are two of the major issues in designing an experiment or survey.

Given a representative (unbiased and large) sample, we can then use ***inferential statistics*** to draw conclusions about the defined population. In other words, we can *generalize* the results of analyzing a sample to a larger population that the sample appropriately represents.

B. What is a *statistical* hypothesis?

Imagine a bar graph of average organic chemistry scores for biology and chemistry majors. The two bars will differ in height—it is extremely unlikely that the averages will be

exactly the same. But is this difference in means between *samples* large enough to conclude that the *populations* really differ? To answer this question, we use sample data to determine how likely it is that the difference in means between samples could have been due to *chance* rather than to a *real difference between populations*.

For any statistical test we define two alternative *statistical* hypotheses:

- ***the null hypothesis*** **(H_o)**: the result expected if there were no relationship between variables
- ***the alternative hypothesis*** **(H_a)**: the result expected if there were a predicted relationship between variables (either a difference between groups or a correlation between variables)

Why do we bother setting up such formal alternatives? The answer has to do with how science works, by a process called *falsification*: we assume by default that there is **no relationship** (the null hypothesis) unless we have strong enough evidence to reject the null. This process reflects the *conservative* nature of science—we do not accept a new, alternative idea unless the evidence is highly convincing. In fact, a typical criterion for "rejecting the null hypothesis in favor of the alternative"* is that the relationship must be so convincingly strong that it would occur by chance (that is, because of a chance sampling error) no more than 5% of the time. [Stronger criteria are often applied where the cost of mistakenly rejecting the null hypothesis is high. For example, because the costs of producing and marketing a new drug is high, we might choose to reject the hypothesis that the effects of a new drug differ from those of the current drug on the market if the difference would occur by chance no more than 1% of the time. In that case, it might pay to be even more conservative.]

C. What does a statistical test actually do?

For any statistical test, we start with a simple assumption. We then evaluate whether there is strong enough evidence *to reject* this assumption.

Assumption: **The null hypothesis is true.**

If the null hypothesis **were true**, we would *expect* that a test statistic calculated from the data (a measure of the strength of the relationship between variables) to equal *zero*: a null hypothesis states, for example, that there is no correlation ($r = 0$) or no difference between two group means ($t = 0$). However, because we calculate a test statistic from just a limited (and therefore imperfect) sample of a population, it would not be surprising to get *small* differences from zero (positive or negative) in our test statistic ***just by chance,*** even if the null hypothesis were true. The question is, how large can the test statistic get before we begin to suspect that it is *not* due to chance (and that the null hypothesis, in fact, is probably *not* true?)

Imagine repeating the same data collection 1000 times, *still assuming the null hypothesis is true*. Each time, you take a new sample from your population and calculate a new test statistic. If you compiled all your test statistics, they would form a normal distribution centered at zero (see right). Most values should be close to zero, and fewer would be extreme (large or small). That is, there is a high probability of getting a

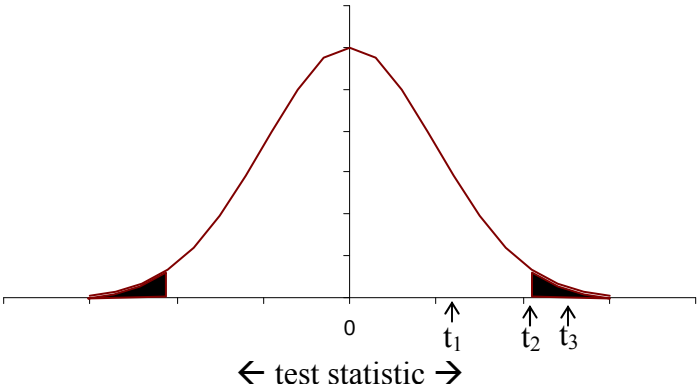

Fig. 1. Distribution of a test statistic given the assumption that the null hypothesis is true. Note that the probability specified by α is distributed equally between the two tails (for a two-tailed test).

small value and a low probability of getting an extreme value, *just by chance*. [Q: Statisticians can generate this probability distribution just by knowing the sample size. *How would you expect the width of the curve to change depending on the size of the samples used to calculate the test statistic?*]

The problem is, in research you often have the result of only one such experiment. So, what is the probability of getting the test statistic that *you* got (remember, assuming the null hypothesis is true) *just by chance*? That probability (called the **P-value**) can be found by seeing where your test statistic falls on this distribution. In Fig. 1, test statistic $\mathbf{t_2}$ falls a certain distance from 0, associated with a certain probability (**P)** of getting a value that extreme *just by chance* even if the null hypothesis were true. The value $\mathbf{t_1}$ is closer to 0, so has a higher probability, while $\mathbf{t_3}$ is further from 0, so has a lower probability of occurring by chance *assuming the null hypothesis were true*. Small test statistics have high probabilities of occurring by chance (and large P-values, see Section D), and large test statistics have low probabilities of occurring by chance (and low P-values), again *assuming that the null hypothesis is true*.

Because test statistics fall along a continuum, we need some way to say when our value is so extreme—that is, when it has such a small probability of having occurred by chance—that we question our assumption that the null hypothesis is true. That criterion is based on **α (alpha)**, a threshold probability value that we choose in advance. We use that threshold to judge when we have enough evidence to reject our assumption that the null hypothesis is true. When $\mathbf{P < \alpha}$, we conclude that the probability of our large test statistic occurring by chance is too small to stick with the null hypothesis, and instead we **reject the null hypothesis in favor of the alternative**. Conversely, when $\mathbf{P > \alpha}$, we **fail to reject the null hypothesis.**

By convention in biology, α is usually set at 0.05 (see section I). That is, we decide to reject the null hypothesis only when we expect a test statistic *as large as ours* no more than 5% of the time by chance. In Fig. 1, the shaded areas under the curve together account for 5% of the distribution of test statistics expected by chance (each tail has 2.5% of the probability). The **critical value** is the test statistic associated with the probability $\alpha = 0.05$. In this example, the critical value for our test is at $\mathbf{t_2}$, so anything equal to or larger than $\mathbf{t_2}$ (or equal to or smaller than $-\mathbf{t_2}$) provides enough evidence to reject our initial assumption that the null hypothesis is true.

D. How is a statistical test carried out?

Scientists have access to a dizzying array of statistical tests. Fortunately, many simple analyses can be performed with knowledge of just three tests: the **correlation analysis**, the **t-test**, and the **chi-square analysis**. Which test to use depends on whether the variables are continuous or categorical. **See APPENDIX 1** to determine when to use each of these tests.

Regardless of which test is used, the procedure is similar:
(1) calculate a **test statistic**,
(2) compare the test statistic to the threshold **critical value** (found in a statistical table),
(3) determine a ***P*-value** by comparing the test statistic to other values in the table, and
(4) reach a conclusion to reject the null hypothesis only if ***P*** is less than **alpha**.

Here are the details:

- What is a **test statistic**? A single value computed from your data. For the three tests you will use in this class, the test statistics are *r* (for correlation analysis), *t* (for a t-test), and χ^2 (for a "chi-squared" test). Excel will calculate them or help you to calculate them.

- What is a **critical value**? A threshold value that can be looked up in a table. If your test statistic exceeds the critical value, then the data from which you computed the test statistic are highly different from what you would expect if the null hypothesis were true, so any relationship you found between variables is unlikely to be due to chance. Critical values are tabulated based on the type of test, the **degrees of freedom**, and **alpha**.
- What are the **degrees of freedom**? A number based on the sample size of the data (**see APPENDIX 2** for how to calculate *df* for each test). Because larger samples give greater power to detect a relationship between variables, sample size affects the critical value.
- What is **alpha**? A probability value chosen *before* the data are analyzed. Because the choice of alpha determines the critical value, it also acts as a threshold for decisions about the null hypothesis: when the probability of getting your results by chance is *less than* alpha (that is, your test statistic is *greater than* the critical value), the null hypothesis is rejected in favor of the alternative. Alpha is typically set at 0.05 (see section H).
- What is the ***P*-value**? A probability value calculated from your data. ***P*** is the probability that the relationship between variables measured *from your sample* is due to chance rather than to an actual relationship *in the population*. It is also, therefore, the probability that you are making an error by rejecting your null hypothesis (see Section H).

E. Do these tests assume anything about my data?

Yes, but many tests work even with small violations of these assumptions, so we will not worry here about testing them. The kinds of statistical tests you will use make just a few basic assumptions that are worth knowing about:

- Data points are assumed to be independent of one another. For example, when measuring scores on an organic chemistry exam, we assume that each student's score is independent of the scores of other students (not always the case!).
- Any statistical test fits the data to some kind of model, which is an ideal representation of how the data are patterned. Real data points always deviate from the ideal model. The size of those deviations (known as residuals) are assumed to have a normal (bell-shaped) distribution, with many more measurements close to the average and progressively fewer toward the tails. This kind of distribution will be true for many types of data.
- The **t-test** assumes that measurements for the two groups you are comparing have equal standard deviations. If the standard deviations you calculated are not terribly different, you probably meet this assumption. If they are terribly different, a version of the t-test is available that can account for this difference in standard deviations.

F. What are the possible outcomes of the test?

- Conventionally there are two possible outcomes: (a) *failure to reject* the null hypothesis, or (b) *rejection* of the null hypothesis *in favor of* the alternative hypothesis. It is incorrect to state that the test leads acceptance of the null hypothesis or proof of either hypothesis.
- Rejection of the null hypothesis does not necessarily imply a mechanism for the relationship. The effect could be due to some other mechanism you didn't propose.
- Rejection of the null hypothesis—a statistical outcome—does not necessarily mean that the effect has great biological significance. As a biologist, it is still necessary to consider the magnitude of an effect when judging its biological importance.

G. How are statistical results reported?

To report the outcome of a statistical test, one states a conclusion along with the test statistic, degrees of freedom, and P-value (the latter three often in parentheses). For example:

> "There was no significant difference between the means of the two groups ($t = 0.45$, $df = 134$, $P > 0.05$)".

H. Why does *P* give the probability of making an error, and why do we set $\alpha = 0.05$?

We have established that the ***P*-value** is the probability of getting a test statistic as extreme as ours by chance if the null hypothesis were true. If ***P*** is small enough (less than alpha), we decide to reject the null hypothesis in favor of the alternative. But of course there is still some probability (given by ***P***) that the null hypothesis *is* true and we just happened to get one of those extreme sampling errors. The P-value is therefore a statement of *confidence* about a decision to reject the null. For example, if the test concludes that $P < 0.02$, we have strong confidence that *less than* 2% of the time, with a test statistic as large as the one we calculated, we will be making an error by rejecting the null hypothesis. This type of error—rejecting the null hypothesis when it is in fact true, known as a "false positive"—is called a **Type-1 error**. Alpha is therefore the upper limit on the type-1 error we are willing to tolerate when performing the test.

A second type of error, called a **Type-2 error** (or a "false negative"), involves failing to reject the null hypothesis when it is in fact false (in the whole population). In science, we generally guard against Type-1 errors more than against Type-2 errors. The reason is that science is conservative—it does not accept new ideas (alternative hypotheses) until there is strong evidence. Setting α higher (e.g., 0.1) would lead to rejecting the null hypothesis more often, but with a higher number of false positives. Setting α lower (e.g. 0.01) would reduce the false positives, but could make it unreasonably hard to reject the null hypothesis, and create more false negatives. The value $\alpha = 0.05$ is a compromise. In some fields, however, a lower value for α is chosen because there are especially high costs of a Type-1 error. For example, a pharmaceutical company might want especially strong evidence that a new drug works significantly better than the current drug before it invests millions of dollars in R&D. It would set a low $\alpha = 0.01$ in order to conservatively guard against concluding there is a real difference between drugs in case there isn't.

Note: information in section I is optional for this course, but will be interesting and useful for those who want to build their understanding of statistical testing.

I. What are one-tailed and two-tailed tests?

The example above describes null (H_0) and alternative (H_a) hypotheses associated with a *two-tailed test*. Consider the goal of a t-test, which is used to test for a difference between means of two groups **a** (μ_a) and **b** (μ_b):

TWO-TAILED TEST

Hypothesis	In words	In symbols
H_0	There is no difference between means of the two groups	$\mu_a \neq \mu_b$
H_a	There is a difference between means of the two groups	$\mu_a = \mu_b$

We use a **two-tailed test** when we make no strong prediction about which of the two means will be greater. That is, either an extreme positive or negative test statistic would provide evidence to reject the null hypothesis (see Fig. 1 above).

A **one-tailed** test is used instead when we have decided, *before* looking at the data, that we have a *directional prediction*: for a t-test, that the mean for group 1 is bigger (or smaller) than for group 2, or for a correlation analysis, that there is a positive (or negative) correlation between variables. In this case, the null and alternative hypotheses are slightly different, as shown here for a t-test:

ONE-TAILED TEST

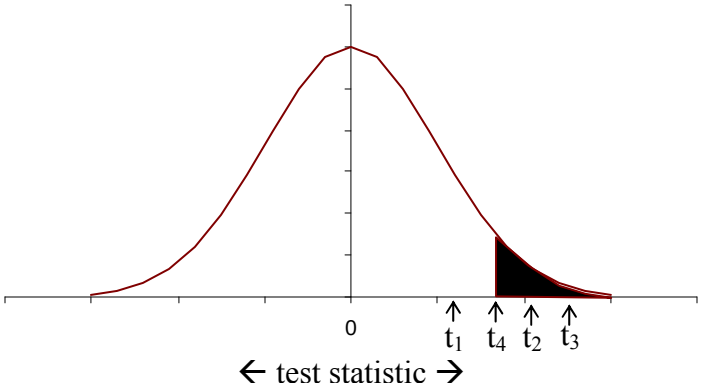

Fig. 2. Distribution of a test statistic under the assumption that the null hypothesis is true. Note that the probability specified by α is concentrated in only one tail (for a one-tailed test).

Hypothesis	**In words**	**In symbols**
H_0	The mean of group **a** is not larger than the mean of group **b**	$\mu_a \leq \mu_b$
H_a	The mean of group **a** is larger than the mean of group **b**	$\mu_a > \mu_b$

Notice that the null hypothesis still expresses a null expectation ("is not") but that the criteria now include *less than or equal to*, whereas the criterion for the alternative hypothesis is *greater than*. The two hypotheses together must cover all possible outcomes, just as they did the two-tailed test.

Which test you choose depends entirely on which prediction you make *before the data are collected or analyzed*. When there is a *directional prediction*, the one-tailed test provides a key advantage: it gives greater power to reject the null hypothesis, because all of the probability α is now concentrated in *one tail* of the distribution. As a result, the critical value for a one-tailed test in Fig. 2 is $\mathbf{t_4}$, which is smaller (easier to exceed) than $\mathbf{t_2}$, the value that was associated with a two-tailed test. Any value of your test-statistic falling between $\mathbf{t_4}$ and $\mathbf{t_2}$ would now lead you to reject your null hypothesis, which was not true for a two-tailed test.

When stating a conclusion from either test, one should be true to the null and alternative hypotheses chosen before looking at the data. That is, if one rejects the null hypothesis, the conclusion is exactly the alternative, and if one fails to reject the null, the conclusion is limited to the null. This means, for example, that if you carry out a two-tailed test and reject the null hypothesis, you can conclude only that the means are not different, not that one is specifically larger than the other. And if you carry out a one-tailed test and your test statistic falls in the opposite tail from where your alternative hypothesis predicted, you can only still fail to reject the null hypothesis. If you want to test only the directional prediction, use a one-tailed test.

APPENDIX 1. Which statistical test should I use?

If you are testing the relationship between…	**use this test…**	**to answer this question…**	**involving these statistical hypotheses…**	**to reach this kind of conclusion…**
2 continuous variables	Correlation analysis	Is there a statistical tendency for high measures of one variable to be associated with high (or low) measures of another variable?	H_o : there is no association between variables H_a: there is an association (positive or negative) between variables	If the association is stronger than is likely by chance, the variables are said to be significantly positively (or negatively) correlated.
1 categorical variable & 1 continuous variable	t-test	Is there statistical evidence that the mean of one group is significantly different from the mean of a second group?	H_o : there is no difference in the average between groups H_a: there is a difference (positive or negative) in the average between groups	If the difference between means (relative to the standard error) is more extreme than expected by chance, then the difference is said to be statistically significant.
2 categorical variables	Chi-square test	Is there a statistical tendency to belong to a particular category in one variable if a subject belongs to a particular category in the other variable?	H_o : there is no association between two categorical variables H_a: there is an association (positive or negative) between the two categorical variables	If the association is stronger than is likely by chance, the variables are said to be significantly associated with one another.

APPENDIX 2. Test statistic, calculation of sample size and degrees of freedom for different tests

Statistical test	**Test statistic**	**Sample size**	**Degrees of freedom**
Correlation Analysis	r	N = number of subjects with paired measurements of the two variables	N–2
t-test	t	N = total number of measurements taken in both groups	N–2
Chi-square test	χ^2	N = total number of subjects measured C_1 & C_2 = number of categories in variables 1 & 2	C_1-1 x C_2-1