Overview:

Over several 4 weeks, students…

* get familiar with the ADW, learn about types of data available
* are introduced to basic qualities of variables: continuous vs. categorical, dependent vs. independent
* brainstorm questions they can answer with data,
* use the Quaardvark search engine to get the data they need.
* Are introduced to, and apply, basic graphing and simple statistical tests to their data
* write up their findings in abbreviated scientific paper form

Week 1: introduction to ADW, data exploration, discussion of variable types, brainstorming, writing assignment (below) distributed.

Week 2: graphing exercise

Week 3: statistics exercise

Week 4: report due

Activities for students the first week in computer lab and topics to address:

Intro to Animal diversity web

Go to computer lab

Look through and explore what is there

Data available – continuous vs categorical

Variables that may affect each other, dependent vs independent

Brainstorm about some ideas

Look up background literature – early intro to literature search as way to explore possible questions. Do these questions make sense? Get a biological basis for each question.

Next time use what is available to formulate 3 questions you can test statistically.

Continuous vs continuous

Continuous vs categorical

Categorical vs categorical

Discussion Assignment due for 5 Sep

Explore the Animal Diversity Web (ADW).

Find data that fit continuous and categorical variables that are available in ADW.

Brainstorm about some ideas

Use the data available to formulate 3 questions you can test statistically. The 3 questions should make use of the data so that they are making these comparisons:

1) Continuous vs continuous

2) Continuous vs categorical

3) Categorical vs categorical

Ideally, your questions will center around a particular topic dealing with a group of animals. For example, you could examine characteristics associated with being a carnivore vs. and herbivore in mammals (you should focus your test on a particular taxonomic group or comparisons between two different groups). The three questions might address these comparisons:

1. You might investigate how mass and lifespan are related in these two groups for a comparison of two continuous variables.
2. You could test to see if the mean mass of carnivores is significantly different from the mean mass of herbivores for a comparison of one continuous and one categorical variable.
3. Finally, you could test if there is a relationship between diet (carnivore vs. herbivore) and breeding cycle (continuous vs. seasonal) for a comparison between two categorical variables.

In each you should write your question so that you can make a prediction with the data. For example, for my first comparison above I might ask. Do carnivores, on average have longer lifespans than herbivores? I can use this question to formulate a statistical test using the data available on ADW.

As we go forward, I’ll continue this example and use it to show how we can use the statistical tests to make comparisons.

**Biology 211: Project 1: Animal Diversity Web hypothesis testing**

**DUE DATE: 3 Oct 2012 at the beginning of discussion**

**Your assignment.** Write (written in your own words), type, and hand in your report. Be sure to include your graphs as figures.

This written assignment will be *an abbreviated form* of a scientific paper. This is an opportunity to gain some expertise in the style of writing, which sections information goes into, and to obtain feedback on writing. I am happy to meet with you if you have questions about your writing.

For THIS assignment you will organize your paper into the typical sections of a journal article – however see notes below on specifics on what to include particular for this assignment.

1) **Introduction:** This section is only two short paragraphs. The first paragraph is a general overview of the project written as if it was the first paragraph of a journal article. See the original handout for guidance on the general goals of the project. Follow the style of articles we have read so far. A second paragraph includes the specific biological questions or predictions for your study.

2) **Methods:** A short paragraph describing your use of Excel as your main methodology to test your questions using graphical and statistical methods, and your use of the Animal Diversity Web as the source of the data. Diagrams can sometimes be helpful here.

3) **Results:** A couple paragraphs describing your results in words. You will specifically refer to and describe the patterns in the graphs (labeled as Figures 1 and 2), report your test statistics, df and p-values. Each figure should have a figure legend, be named figure 1 and 2, and the figure legend should be able to stand alone. Take a peek back at the paper we discussed for style. You will have one scatterplot and one bar graph.

4) **Discussion:** A paragraph describing the conclusions that you draw based on your statistical analyses and interpretations of your graphs. How do these results offer support or do not support your biological hypotheses/questions/predictions? How does this relate back to a broader question about the relationship between the variables?

5) **Abstract:** Although after the title, this should be the first section the reader sees, it should be the last section written. See below for style. Approx 100 words for this assignment.

The written portion of your report should be at minimum TWO FULL pages (one inch margins, double spaced, 12 point font). Additional pages will include your graphs (labeled figures 1, figure 2…), and figure legends (descriptions of what the graph represents).

**Instructions for Authors:**

All scientific journals have a document and set of guidelines that must be followed for a paper to be considered for publications. Below find the Instructions for Authors for Project 1 - 211.

Your paper should also include:

* *Title, name & affiliation of investigator, and name of partner* – at top of the first page.
* The sections Abstract, Introduction, Materials and Methods, Results, Discussion, Acknowledgements and Literature cited should be clearly labeled
* Abstract must include: 1) the research conducted, including the rationale, (2) methods, (3) key results and (4) the main conclusion, including key points of discussion.
* Any Materials and Methods section should be written in sufficient detail such that someone else could replicate of all experiments described.
* *Figures or tables* – may be placed inline with the text or following the text on separate sheets. Refer to all diagrams, graphs and photographs as ‘Figures’.
* *Figure legends should be included.* Key information describing each figure should be in the first sentence of the legend, because this is the text that will be immediately visible. The rest of the legend should be a self-contained, full explanation of the figure, with all abbreviations defined.
* *Literature cited* – immediately following the text. Give the full reference for all sources in the text, and make sure all sources referenced are in the text. Your textbook is an appropriate and sufficient source for Project 1, but see additional requirements for project 2 and 3. Use the format shown below. This section is not applicable for this assignment, but will be important for future projects.

Use 12 point font, 1 inch (2.5 cm) margins left, right, top and bottom, single or double spaced accepted.

**Do not use the word prove** (proven etc.). In the data analyses we are doing here, we are finding support in favor of a particular prediction, or failing to find support.

**Data are plural.** So “The data were analyzed using excel”. Data set is singular if you want to mix your language usage a bit.

**Format for presentation of statistical results** in a sentence format: There was a significant positive relationship between shell length and shell width (r = 0.92, df = 43, p<0.001).

The word species is both plural AND singular.

Reference your figures in your text. E.g. We found a positive relationship between herbivory and leaf size (Figure 2).

If the species is in very common usage then the common name will suffice, although the scientific name should still be given at first mention (e.g. soybean (*Glycine max*). For subsequent uses, abbreviate genera to their initial letters e.g. *G. max*, except where this could result in confusion between species, or if the species name is the first word of the sentence.

Make sure scientific names are in italics with only the genus name capitalized. E.g. *Moehringia macrophylla*. Common names should NOT be capitalized.

Page numbers, and author’s name should be in either a header or footer on each page of the document

**CITATION FORMAT** (for later assignments)

*Citations in the text*. Each time an article is cited, the authors' last names and the year should be placed in parenthesis. If the authors’ name(s) are the subject of the sentence, only the year goes in parentheses. When there are three or more authors, all but the first author are abbreviated as et al. (meaning “and others”). Here is an example involving one single-authored paper, one co-authored paper, and one with more than two authors:

Some biologists, like Welch (2002), have said one thing, while others (Podolsky and Wise 2003, Darden et al. 2004) have said another.

*References in the literature cited section*. Use a typical format as shown below:

Journal article:

Darden T, Scholtens BS, Harold A. 2004. Acceptable citation styles for research proposals. Journal of Literature Citation 6:121-126.

Book Chapter

Sancho G, de Buron I, 2002. Citation styles in European Journals. In *International Guidelines for Citation*, ed. Southgate A. CofC Press: Charleston.

Book

Wiseman R. 1999. Stickler for proper citation: a memoir. CofC Press: Charleston.

Tips for Better Writing…including Lab Reports

**General writing tips:**

**Thesis** (i.e., main point you want to make in your paper) should be clear, and evident in the introduction of your paper (or early in first page). In science, the objectives and hypotheses to be explored and tested should be clearly articulated as well.

Spend time to **develop a strong beginning and ending to your paper**—these form lasting first and last impressions, and are often the only parts of your paper that some scientists will read.

**Link the ideas in your introduction to the discussion as you conclude your paper.**

**Support your statements** and assertions with arguments, examples, evidence from the literature.

**Use quotes sparingly** when you can't make the point effectively in your own words or via paraphrasing.

**Comparisons**: If something is bigger, higher etc., be sure to indicate what items are being compared.

NOT: "We detected a higher floral density."

INSTEAD: "We detected a higher floral density in the high water treatment in comparison to the low water treatment.*"*

OR: "Floral density was high in our experiment."

**Watch subject/verb agreement and singular/plural consistency throughout**. e.g.,

NOT: America and their people were saddened by the hurricane.

INSTEAD: The events around the hurricane saddened the American people. (note active voice here also)

In **style**, avoid choppy sentences of similar, short length, or constant use of long, complex sentences. Omit explanations of why you wrote something, or personalized phrases like "as I read the article by Pearcy, I thought that the experiment was complicated." **Read published texts for ideas on style.**

**In lists…use parallel construction** with respect to verb tense, etc. *"They gathered the data, analyzed the data, and wrote a report about the data."*

**Learn to recognize sentence fragments** (phrases missing a subject or verb) **and run-ons** (two sentences that can stand alone but are connected by insufficient punctuation. Use a **; or .** to separate them**, NOT** a **,**

**Edit and streamline text** for clarity.

**Use quotes sparingly**, only when paraphrasing is not appropriate and clearer, or when the quote is particularly distinctive and relevant.

**Topic sentence should include content of sentences within that paragraph**. Errors include too many ideas in one paragraph, or too many short paragraphs which are clearly interrelated as part of one idea or argument.

**Style**

**Study primary journal articles for style as well as content.** This will help you understand the conventions in science (e.g. how to refer to figures and statistics in text of results section) and how others streamline language to create effective arguments and explanations of their research. An excellent example is the article on Papilio for our project.

**Use active voice and inclusive language when possible** to clarify who or what is acting or has acted. Active voice (using I and we) is now common practice in publishable work in the biological sciences. e.g. YES 🡪 "*We collected the plants on…* " versus NO 🡪 "*The plants were collected on…* Active voice makes communication more direct.

**Reduce long, content-poor phrases (constructs):** These interfere with direct communication, flow, and a quicker, clearer understanding of your arguments. Examples to avoid: *It is thought that, due to the fact that, from the position of.*

**Avoid long constructs** such as "due to the fact that." These distract reader from main point and content and can usually be shortened (e.g. "Because….."). The same applies to sentences beginning with "there is/there are”. Theses phrases have no real content.

**Position key concepts and ideas first in sentence for greater impact**, and avoid beginning sentences with there are/is, OR this

e.g. *"Density is high in the high water treatment* *is strong*.

*"Three important results emerged from our experiment"* is stronger than "*There are three important results from our experiment"* and than "*The authors describe three important results in their paper."*

REPETITION is okay! More than okay, it is expected between sections. Your predictions should be restated in the discussion.

**AVOIDING PLAGIARISM**

**The bottom line**: *Plagiarism is grounds for receiving an F on the assignment, or an F in the class. Consider this your warning!* You must write in your own words using your own sentence construction based on your own understanding. Do not rearrange parts of sentences written by other authors (including your partner) or simply change, remove, or add words--these all constitute plagiarism. Avoid quoting authors verbatim, even if you put the text in quote marks and cite them (unless you need to comment on their exact words and the meaning would be lost unless you quoted them exactly). Cases of suspected plagiarism will be handled according to the College of Charleston honor code (see pages 11-12 of the CofC Student Handbook <http://www.cofc.edu/about/handbook.pdf>). CofC plagiarism violations range from Class 1 (majority of submitted assignment is intentionally plagiarized) to Class 3 (unintentional plagiarism). If you are unsure about what constitutes plagiarism see the following links. <http://gervaseprograms.georgetown.edu/hc/plagiarism.html>

<http://owl.english.purdue.edu/owl/resource/589/01/>

**Project 1: Managing, exploring, and summarizing data**

Today, you will learn how to manage data, calculate simple descriptive statistics and create graphs in Excel.

**Step 1:** Using the questions that you have proposed, construct graphs predicting what the relationship of the variables will look like (for the instances with 2 continuous variables and with 1 continuous and 1 discreet – the case with 2 discreet will be shown in a table).

Also write out your hypothesis in an if-then statement to accompany your graph or table

|  |  |
| --- | --- |
| ***Hypothesis*** |  |
| ***Predictions (graphs)*** |  |

|  |  |
| --- | --- |
| ***Hypothesis*** |  |

|  |  |
| --- | --- |
| ***Predictions (graphs)*** |  |

**Step 2:** Download the raw data from the Animal Diversity Web and import it into an Excel spreadsheet

**Step 3:** Calculate the means and standard deviations for each continuous variable and each category that you will be comparing.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Category (species, etc.)** | ***Mean*** | ***Standard deviation*** |
|  |  |  |  |
|  |  |  |
|  |  |  |  |
|  |  |  |
|  |  |  |  |
|  |  |  |

**Step 4:** Save your work in your spreadsheet! Make sure that you save on a thumb drive or email the file to yourself.

**Step 5:** Use Excel to create graphs from your means and standard deviations (see other sheet for FAQs about doing graphs).

**Step 6:** Create the table for the comparison with 2 discreet variables.

**Excel Tips and FAQs I. Managing and Summarizing Data**

***Remember to save frequently, but only when you’re sure you haven’t messed up your data set! Saving an original and then modified copy in class is a good idea.***

**FAQ1: How do I sort my data?** Sorting provides a way to rearrange the rows so that all the data for a particular group, like males or females, are put into a continuous block of rows.

1. *Save your spreadsheet before sorting.*
2. Highlight every column that is part of the data set\*.
3. Go to the **Data** menu and select **Sort**.
4. Choose the variable by which you want to sort. For example, if “sex” is in column B and you want to separate the data for males and females, choose the variable “sex” for sorting. If “sex” is in column B and “age” is in column C, and you want to separate data by both sex and age, enter “sex” as the first sorting variable and “age” as the second.

\*Beware: If you don’t highlight all the columns, you will sort some columns but not others, and the data in your rows will turn into a jumbled mess!

**FAQ2: How do I write a function in Excel?** Functions are used to do calculations on numbers in cells. In Excel, a function is typed into a blank cell in the following form:

 = FUNCTION (*argument 1, argument 2, ….*)

You will use functions like AVERAGE and STDEV to calculate means and standard deviations. The “arguments” used by each function, located inside the parentheses, are described below.

 Note that you can also have Excel do mathematical formulas in the same way. For example, if mass is in cell B2 and height is in cell C2, you can enter this formula into a blank cell (for example, D2) to calculate their ratio:

 = B2/C2

**FAQ3: How do I calculate a mean?** The arithmetic mean is a *summary statistic* that provides one estimate of what the average member of a group is like. The mean is useful for making comparisons between groups.

1. Select any blank cell where you want the mean to appear.
2. In that cell, type the function (see FAQ2) that Excel uses to calculate the mean for a group of cells. For example, if cells B2 through B9 contain values for the height of males in your sample, and you want to calculate the height of the average male, you can simply type:

 = AVERAGE (B2:B9)

Shortcut: The colon tells the function to include the continuous block of cells B2 through B9. Instead of typing the *argument*, you could type the rest of the function, position the cursor inside the empty parentheses, and then use the cursor to select and highlight the block of cells that make up the argument. This action will put the correct argument between the parentheses.

Outcome: The result will appear inside the cell, but the function will still be visible in the function window at the top of the spreadsheet (when that cell is selected). To change the function you can type directly in the function window.

**FAQ4: How can I calculate a standard deviation?** The standard deviation is a *summary statistic* that provides one measure of *spread* or *variability* in the data. To compare two groups, one must know not only the mean but also the amount of variability in the data. (After all, the means for two groups will always be at least a little bit different—the question is, are they different enough to conclude that there is a real difference between the groups?) Means are usually reported along with standard deviations or standard errors (see below).

1. Select any blank cell where you want the standard deviation to appear.
2. In that cell, type the function Excel uses to calculate the standard deviation for a group of cells. For example, if cells B2 through B9 contain values for the height of males in your sample, and you want to calculate the standard deviation of male heights, you can type the following (or use the shortcut described above to fill in the argument:

 = STDEV (B2:B9)

ps. Save your work frequently!

**FAQ5: How can I repeat the same function without having to retype it multiple times?** Sometimes you will need to make the same calculation on every row of your spreadsheet. For example, you might want to calculate the ratio of mass to height (see example in FAQ2), and then use this ratio as a new variable. Start by entering the formula in a blank cell on the same row as the data. Then you can copy and paste the function into all the cells in the same column on the other rows where you want to make the same calculation. This step can save a great deal of time if you are making the same calculations for every row in a database.

 [To understand why this works, you must understand the difference between a *relative reference* and an *absolute reference*. In the example from FAQ2, the formula in D2 refers to cells B2 and C2. However, the formula is actually telling Excel, “take the number in the cell two columns to the left and divide it by the number in the cell one column to the left.” This is a relative reference—it does not refer absolutely to B2 and C2, but rather to cells that are in the position of B2 and C2 *relative to* D2. So, when you copy and paste your formula from D2 to D3, the formula will do the calculation on cells in the same *relative* position (namely, B3 and C3).]

 You should always check to be sure you are getting the result you intended. If necessary, you can correct the reference in the function window as explained in the shortcut of FAQ3.

**Excel Tips and FAQs II. Making graphs**

**Tip 1: Before graphing anything, first figure out what your graph will look like.**

* Start by visualizing and making a general sketch, asking yourself the following questions:
	+ Will a bar graph, scatterplot, or some other type of graph be most appropriate for the relationship I want to show?
	+ How will I label the axes be labeled, and what will need to go in a figure legend?
	+ Will I need error bars, and what will they show?
	+ Will black-and-white be sufficient, or will color be useful for conveying information?
* Figure out whether your graph will require raw data or only means and standard deviations.
	+ If you will be displaying means and standard deviations (for a bar graph, for example), you’ll need to organize these descriptive statistics into a mini-table in your spreadsheet.

**Tip 2: Use Excel to make the graph you want while avoiding unnecessary complications**

* Highlight the cells that include the numbers (raw data or means) that you want to plot, *including* the names of groups or variables. Do not include standard errors. Then click on the **Chart Wizard** icon, which looks like a little bar graph, in the tool-bar.
* Select the type of graph you want to produce. Although you are given many choices, you should have already chosen the graph type in Tip 1.
* Click **Next**. If your graph does not show data in the way you expected, try (1) switching the choice between columns and rows, or (2) clicking on the **Series** tab—you might have to correct which cells are referred to for your X and Y variables. You can make this correction by selecting cells, as explained in FAQ 2.
* Click **Next**, and use the tabs to specify axis labels and to alter gridlines and legends.
* Continue clicking **Next** to complete your graph. Then save your spreadsheet.

What to avoid:

* Three-dimensional graphs. If you have two-dimensional data, three dimensional graphs do not add any additional information, and they can actually make it harder to read your graphs.
* Gridlines, background colors, and other distractions. The more you remove from your graph, the more the data will stand out.
* Unnecessary color. Use color to convey information. Most of the time, black-and-white works well.

**FAQ6: How can I add error bars to my graph?** Error bars are included in bar graphs to show standard deviation or other measures of spread

* Right click once on one of the bars in a bar graph. Choose **Format > Data Series**.
* Click on the **Y Error Bars** tab and add **Custom** + and – error bars by placing the cursor inside the correct box and then selecting the cells that contain standard deviations in your data sheet.

**FAQ7: How can I make my graph look more professional?** Compare your original sketch with the graph you’ve made in Excel. Chances are, the Excel graph is not exactly what you had in mind. The labels may not be quite right, the colors may be unappealing, the font may be too small or too large, the legend may be unnecessary, and the gridlines may be annoying.

* Right click on axes to change the range for this axis, increase the font size, etc.
* You can also right click on axes and format them to thicken the axis lines
* Right click on bars or points to change colors, add or modify error bars, add a trendline, etc.
* Right click on the whole plot to change colors or to return to chart options where you can remove gridlines or legends, change labels, etc.
* Explore these options until you have an informative, attractive, and easy-to-read graph.
* Think about the data to ink ratio.

ps. Save your work frequently!



Data from Excel file

|  |  |
| --- | --- |
| Height | Leaf Length |
| 10.1 | 47.6 |
| 10.3 | 68.4 |
| 10.3 | 57.2 |
| 11.5 | 64.3 |
| 11.7 | 34 |
| 12 | 47 |
| 12 | 26.1 |
| 12.3 | 32 |
| 12.6 | 32.6 |
| 13.2 | 78 |
| 13.3 | 60 |
| 13.5 | 27.5 |
| 13.5 | 28.8 |
| 14.2 | 48 |
| 14.58 | 51.9 |
| 14.69 | 63 |
| 14.8 | 37 |
| 14.9 | 63.8 |
| 15 | 71.9 |
| 15 | 72.2 |
| 15.2 | 74 |
| 15.2 | 61 |
| 15.4 | 90 |
| 16.8 | 25 |
| 17.1 | 13.4 |
| 17.3 | 84 |
| 17.5 | 29.5 |
| 17.7 | 79.5 |

How to set up a bar graph with error bars in excel

In the Excel workbook

|  |  |
| --- | --- |
|  | **Treatment** |
| **Trait means** | **Trait** | **Control** | **Low** | **High** |
|  | Leaf Length | 3.5 | 2.6 | 4.4 |
|  |  |  |  |  |
| **Trait stdev** | **Trait** | **Control** | **Low** | **High** |
|  | Leaf Length | 2.5 | 2.4 | 2.7 |



Some suggestions for of a clean look of an Excel Graph (Yes, Excel has a mind of its own – you’ll need to wrangle it to get it to look like you want it to look)

X Y scatterplot graph:

 Thick lines

 Labeled axes

 Large readable font

 Simple shapes

 No gridlines or extraneous color

 Simple yet thick trendline pattern

Bar Graph

 Thick lines

 Labeled axes

 Large readable font

 Simple colors/hatching

 No gridlines or extraneous color

 Thick error bars

**Using data to test hypotheses**

**Statistical Data Analysis**

Today you will learn how to analyze data using statistical tests that are appropriate for different kinds of relationships between variables. You will then apply what you’ve learned to test your predictions from the data you collected on shells.

**A. What are the different types of variables?** We have distinguished between two different types of variables found in your dataset:

* continuous variables—measurements can take on any value along a continuous range
* categorical variables—values are categories that can be used to divide the dataset into groups

The appropriate statistical analysis, just like the appropriate graphical representation, depends on which types of variables are used for testing your prediction.

**→ Step 1**. Summarize the variable names and types that you used for your three predictions.

|  |  |  |
| --- | --- | --- |
| **Prediction** | **Variable 1** | **Variable 2** |
|  | **Name** | **Type** | **Name** | **Type** |
| 1 |  |  |  |  |
| 2 |  |  |  |  |
| 3 |  |  |  |  |

**B. How well does a sample represent a population?** It is important to recognize that any dataset is collected from a *sample* of all possible individuals that could be measured from an entire *population* of individuals. Because we usually cannot measure every individual, we use statistical tests to build *inference* on whether or not there is a strong enough relationship between variables in our sample to draw conclusions about a pattern in the entire population. We assume, and try to assure by our sampling method, that our sample is *representative* of the population as a whole.

 How is a population defined, and what kind of sample is representative of it? The answers depend on the exact question posed. For example, we might ask, “Are male students taller than female students at the College of Charleston?” The population of interest, as defined by the question, is all students currently at the college. If we could measure every student, we could say definitively whether the average male is taller or shorter than the average female. However, we are more likely to measure only a sample from that population. In that case, we could make only a statement about the *probability* that male and female heights differ based on (1) the sample means and (2) an estimate of how well the sample means are likely to represent the population means.

 To choose a representative sample, we first try to avoid potential *biases*. For example, it would be best to avoid sampling strictly from areas where we expect a disproportionate number of athletes, who might be taller than average. In most cases, a scheme that samples students at random from the population would help to provide a representative sample. Second, we try to choose a sample as large as practical, to avoid the possibility that our sample will provide, by chance, an unusual collection of values from the population.

 Given a representative, unbiased sample that is large enough to test a prediction, we can then use *statistics* and inference to draw conclusions about the entire population of students. In other words, we can *generalize* the results of analyzing a sample to an *appropriate* population.

**C. What is a *statistical* hypothesis?** *Statistics* provide a formal way of deciding whether a biological prediction about a population is likely supported using a sample of data. For example, your bar graphs may have showed at least some difference between groups in the average value. Were these differences, calculated from a *sample*, large enough to conclude that the *populations* actually differ? To answer this question, we use the data to determine how likely it is that the difference between samples was due to *chance* rather than to a *real difference between populations*.

For any statistical test we define two alternative hypotheses:

* *the null hypothesis* (Ho): the result expected if there were no relationship between variables (for quantitative traits). Stated another way, that the differences or relationships observed represent random variation between groups or random associations among variables
* *the alternative hypothesis* (Ha): the result expected if there were a relationship between variables, or the difference in means is too large to be accounted for by random variation among individuals (either a difference between groups or an association between variables)

We assume by default that there is no relationship *until* we have good enough evidence to reject the hypothesis of no relationship. 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” is that the relationship must be so convincingly strong that it should occur by chance (that is, because of a chance sampling of the population) 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.]

Four important notes

1. A statistical test leads to only one of two conclusions: (a) *failure to reject* the null hypothesis, or (b) *rejection* of the null hypothesis *in favor of* the alternative hypothesis. The test does not lead one to accept the null hypothesis nor to prove the null or alternative hypotheses.

2. Rejection of the null hypothesis does not mean that the biological *mechanism* that you described in your prediction is responsible for the relationship between variables. You as a biologist will use inference to make the link between the statistical result and the biological hypothesis. The effect could always be due to some other mechanism you didn’t propose.

3. Rejection of the null hypothesis—a statistical outcome—does not necessarily mean that the effect is an important biological outcome. As a biologist, it is still necessary to consider the magnitude of an effect when judging its biological significance. Some weak relationships may actually be biologically meaningful.

4. The words prove and insignificant are not appropriate when describing the outcome of statistical tests. Instead “provide evidence” and “not statistically significant” are correct.

**→ Step 2.** State the statistical null and alternative hypotheses for each of your biological predictions. Each hypothesis should refer specifically either to differences in means between two groups or to a correlation.

|  |  |
| --- | --- |
| **Prediction** | **Statistical hypotheses** |
| 1 | **Ho:****Ha:** |
| 2 | **Ho:****Ha:** |
| 3 | **Ho:****Ha:** |

**D. How is an appropriate statistical test chosen?** Scientists have access to a large and growing array of statistical tests. Three tests are very commonly used: 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 A & B at the back of this handout for details about these tests.

**→ Step 3.** Based on information in the Appendices and on the variables listed in step 1, record the appropriate statistical test for each prediction, as well as the sample size.

|  |  |  |
| --- | --- | --- |
| **Prediction** | **Appropriate statistical test** | **Sample size** |
| 1 |  |  |
| 2 |  |  |
| 3 |  |  |

**E. How is a statistical test applied?** Regardless of which test is used, the procedure is similar:

(1) calculate a **test statistic**,

(2) compare the test statistic to a **critical value**,

(3) determine a ***P*-value** (or ***P***) based on comparing the test statistic to other critical values based on degrees of freedom, 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 the data from your sample. For our three tests mentioned above, the names of the test statistics are *r*, *t*, and χ2 (“chi-squared”), respectively.
* What is a **critical value**? A value that can be looked up in a table (or by Excel). Critical values are calculated by statisticians based on the type of test, the **degrees of freedom**, and **alpha**. If your test statistic is greater than the critical value, then the data from which you calculated the test statistic are “extreme,” and any relationship you found between two variables is unlikely to be due to chance.
* What are the **degrees of freedom (d.f.)**? A number based on the sample size of the data used (see Appendix B for how to calculate for each test).
* What is a ***P*-value**? ***P*** is the probability that a relationship between variables measured *from your sample* is due to chance rather than to an actual relationship *in the population*.
* What is **alpha**? The *upper limit* on the risk you are willing to take that a relationship between variables measured in your *sample* is due to chance rather than to an actual relationship in the *population*. If ***P***< **alpha**, then the probability of such an error *with your data* is less than the upper limit you set, and you can reject the null hypothesis with confidence. Alpha is typically set equal to 0.05 (a 5% chance of rejecting the null hypothesis when the relationship you found is actually due to chance).

**F. 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 considering:

* Data for continuous variables are assumed to have a normal (bell-shaped) distribution, with many more measurements close to the average value and progressively fewer measurements at more extreme values. This kind of distribution is typical of many types of data.
* Data points are assumed to be independent of one another. For example, when measuring heights of students at College of Charleston, we assume that a large number of measurements are not taken from any single family, because family members are likely to be similar in height and therefore do not represent independent measures of height.
* In addition, 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.

**→** **Step 4.** Use what you have learned from the information above and from Appendix A and B to plan and execute statistical tests of your predictions. Use Excel to carry out your analyses, aided by the accompanying Excel FAQ and the statistical tables supplied.

 For each prediction, run the appropriate test and record the relevant information used to reach a statistical conclusion and an answer to the original biological question:

**Prediction 1:**

|  |  |
| --- | --- |
| Statistical test used |  |
| Test statistic calculated (use correct symbol) |  |
| Critical value for this test |  |
| Degrees of freedom for this test |  |
| Alpha (see part E for recommended value) |  |
| P-value from the test (see Excel FAQs) |  |
| Conclusion: reject or fail to reject null hypothesis? |  |

Use this conclusion to evaluate your prediction and to answer the original question posed:

**Prediction 2:**

|  |  |
| --- | --- |
| Statistical test used |  |
| Test statistic calculated (use correct symbol) |  |
| Critical value for this test |  |
| Degrees of freedom for this test |  |
| Alpha (see part E for recommended value) |  |
| P-value from the test (see Excel FAQs) |  |
| Conclusion: reject or fail to reject null hypothesis? |  |

Use this conclusion to evaluate your prediction and to answer the original question posed:

**Prediction 3:**

|  |  |
| --- | --- |
| Statistical test used |  |
| Test statistic calculated (use correct symbol) |  |
| Critical value for this test |  |
| Degrees of freedom for this test |  |
| Alpha (see part E for recommended value) |  |
| P-value from the test (see Excel FAQs) |  |
| Conclusion: reject or fail to reject null hypothesis? |  |

Use this conclusion to evaluate your prediction and to answer the original question posed:

**Appendix A. Which statistical test should I use for this shell dataset?**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 association for high measures of one variable to be associated with high (or low) measures of another variable?  | Ho : there is no association between variablesHa: 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 predictor &1 continuous response | t-test | Is there statistical evidence that the mean of one group is significantly greater than or less than the mean of a second group?  | Ho : there is no difference in the mean between groupsHa: there is a difference (positive or negative) in the mean between groups | If the difference between means (relative to variation around the mean) is larger 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 particular category in the other variable? | Ho : there is no association between two categorical variablesHa: 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 B. 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 individuals for which you have paired measurements of the two variables (in most cases for the shells N=30)  | N–2 |
| t-test | *t* | N = total number of measurements summed across both groups (in most cases if you have 2 species and 30 measurements per species =60, then df =60-2 = 58) | N–2 |
| Chi-square test | χ2 | N = total number of subjects measuredC1 & C2 = number of categories in variables 1 & 2 | C1-1 x C2-1 |

**Excel Tips and FAQs III. Statistical tests**

Excel provides a way to carry out the basic statistical tests that you will use in this class. Below are instructions for doing a **Correlation Analysis**, **t-test**, **and Chi-square test**. One easy way to compute these statistics is to write simple functions in cells, as we explained in Excel FAQ I. You will also use the statistical tables (**Tables 1-3**) attached to this handout to evaluate the test-statistics generated from these functions.

**FAQ8: How do I perform a Correlation Analysis using Excel?** Correlation analysis provides a way to test the hypothesis of a relationship between two continuous variables.

1. Click in an empty cell where you want to calculate the test statistic, ***r***.
2. Type the function

 =CORREL(array1,array2)

 where array1 = the range of cells that contain values for one continuous variable

 array2 = the range of cells that contain values for the second continuous variable

1. Press return.
2. Re-select the cell that contains *r*, select **Copy** on the **EDIT** menu, select **Paste Special** on the **EDIT** menu, choose **Paste…Values** and press **OK**.\* This step pastes the number instead of the function back into the same cell.
3. A positive value of *r* indicates a positive correlation, while a negative value indicates a negative correlation. To see if the correlation is significant, compare the absolute value of your *r* to the critical value in **Table 1**. To find the critical value you will need the **degrees of freedom** and **alpha** (see handout 1.3, part E).
4. If *r* is greater than the critical value for your chosen **alpha**, you can reject the null hypothesis in favor of the alternative. To report the ***P-value*** for your test, determine where ***r*** falls relative to the other critical values on the same row for your *df* (the possible choices are P > 0.05, P < 0.05, P < 0.01, P < 0.001).

\*Beware! If you don’t use **Copy** and **Paste Special…Values**, the formula you typed could produce a different number when you next sort your data!

**FAQ9: How do I perform a t-test in Excel?** The t-test provides a way to test for differences in the mean value between two groups.

1. Check the standard deviations of the measurements from the two groups that you want to compare. If they are similar, use the formula for equal variances below. If they are very different, use the formula for unequal variances.
2. Click in an empty cell where you want to calculate the test statistic, ***t***.
3. Type the function

 =TINV(TTEST(*array1,array2*,2,2),*df*) for equal variances

 **or**

 =TINV(TTEST(*array1,array2*,2,3),*df*) for unequal variances

 where *array1* = the range of cells that contain the numerical data for one group

*array2* = the range of cells that contain the numerical data for the 2nd group

 *df* = degrees of freedom (see Appendix B)

1. Press return.
2. Re-select the cell that contains *t,*  select **Copy** on the **EDIT** menu, select **Paste Special** on the **EDIT** menu, choose **Paste…Values** and press **OK**.\* This step pastes the number instead of the function back into the same cell.
3. Now compare your ***t*** to the critical value in **Table 2**. To find the critical value you will need the **degrees of freedom** and **alpha** (see handout 1.3, part E).
4. If ***t*** is greater than the critical value for your chosen **alpha**, you can reject the null hypothesis in favor of the alternative. To report the ***P-value*** for your test, determine where ***t*** falls relative to the other critical values on the same row for your *df* (the possible choices given this table are P > 0.05, P < 0.05, P < 0.02, P < 0.01, P < 0.001).

\*Beware! If you don’t use **Copy** and **Paste Special…Values**, the formula you typed could produce a different number when you next sort your data!

**FAQ10: How do I perform a Chi-square test in Excel?** The Chi-square test provides a way to test for an association between two categorical variables. The test requires that you do a simple count of the number of subjects that fall into groups that are combinations of the categories. For example, if you are testing for an association between sex and survival, you would count the number of male survivors, female survivors, male non-survivors, and female non-survivors.

1. Save your work before sorting.
2. **Sort** your data so that you can easily count by eye the number of subjects in each of your groups. When you sort be sure to include entire rows, and to follow all the other directions for sorting in FAQ1 (last week’s FAQs).
3. Enter these counts into a block of cells. Use column and row labels to indicate the groups. See the example below. These numbers are your **Observed** values.
4. Now, sum across each row and down each column by using the formula:

 = sum(cell1:cell2) – or however many cells in the row or column

 where cell1 and cell2 are the sets of adjacent cells that make up the rows and columns of your table. This should produce sums that are placed to the side of or below your rows and columns, respectively. See the example below.

1. Enter the total sum of all subjects in all groups into the lower right corner box to complete the table.
2. Next, create a table below that will contain the **Expected** values for each box. The expected value for each cell is calculated using the following formula:

 = Rsum\*Csum/Tsum

 where Rsum is the sum on the row where the cell is located

 Csum is the sum of the column where the cell is located

 Tsum is the total sum in the lower right hand corner of your table.

 Repeat this step for each of your **Observed** values to produce a corresponding **Expected** value. See the example below.

1. Next, create a table below, where each cell contains the following formula using the corresponding cells above:

 = (**Observed**-**Expected**)2 / **Expected**

 Repeat this step for each of your **Observed** and **Expected** value pairs to complete the table. See the example below.

1. Finally, take the sum of the cells in the last table you computed. This sum is your Chi-square test statistic. Compare your Chi-square value to the critical values in the Chi-square table (separate sheet). If your value of Chi-square is greater than the critical value for alpha = 0.05, you can reject the null hypothesis with confidence. To report the ***P-value*** for your test, determine where your Chi-square statistic falls relative to the other critical values in the table(the possible choices are P > 0.05, P < 0.05, P < 0.025, P < 0.02; P < 0.01, P < 0.005, P < 0.0025, P<0.001).

Count of

female survivors

Sum of all

non-survivors

Sum of all

individuals

Sum of all

females

Expected value calculated for female non-survivors

Chi-square value for female survivors calculated from formula 7

Sum of all individual

chi-square values

**Table 1**. Critical values of the correlation coefficient, *r*, for different degrees of freedom (*df*) and probabilities. Ignore the sign (+ or –) on your calculated *r* in order to compare with the critical value in the table. **If your *df* are not in the table, use the next smaller value**. Critical values are within the box of double lines.

|  |  |
| --- | --- |
| **df** | **Probability (α or *P*)** |
|  | **0.05** | **0.01** | **0.001** |
| **1** | 0.997 | 1.000 | 1.000 |
| **2** | 0.950 | 0.990 | 0.999 |
| **3** | 0.878 | 0.959 | 0.991 |
| **4** | 0.811 | 0.917 | 0.974 |
| **5** | 0.755 | 0.875 | 0.951 |
| **6** | 0.707 | 0.834 | 0.925 |
| **7** | 0.666 | 0.798 | 0.898 |
| **8** | 0.632 | 0.765 | 0.872 |
| **9** | 0.602 | 0.735 | 0.847 |
| **10** | 0.576 | 0.708 | 0.823 |
| **11** | 0.553 | 0.684 | 0.801 |
| **12** | 0.532 | 0.661 | 0.780 |
| **13** | 0.514 | 0.641 | 0.760 |
| **14** | 0.497 | 0.623 | 0.742 |
| **15** | 0.482 | 0.606 | 0.725 |
| **16** | 0.468 | 0.590 | 0.708 |
| **17** | 0.456 | 0.575 | 0.693 |
| **18** | 0.444 | 0.561 | 0.679 |
| **19** | 0.433 | 0.549 | 0.665 |
| **20** | 0.423 | 0.457 | 0.652 |
| **25** | 0.381 | 0.487 | 0.597 |
| **30** | 0.349 | 0.449 | 0.554 |
| **35** | 0.325 | 0.418 | 0.519 |
| **40** | 0.304 | 0.393 | 0.490 |
| **45** | 0.288 | 0.372 | 0.465 |
| **50** | 0.273 | 0.354 | 0.443 |
| **60** | 0.250 | 0.325 | 0.408 |
| **70** | 0.232 | 0.302 | 0.380 |
| **80** | 0.217 | 0.283 | 0.357 |
| **90** | 0.205 | 0.267 | 0.338 |
| **100** | 0.195 | 0.254 | 0.321 |

Looking at the table, you should be able to answer the following questions:

1) As you go from higher to lower *probabilities* for a given *df*, does it become harder or easier to reject the null hypothesis for a given *r*?

2) As you go from lower to higher *df* at a given *probability*, does it become harder or easier to reject the null hypothesis for a given *r*?

**Table 2**. Critical values of Student’s t-statistic for different values of Probability (α or *P*) and degrees of freedom (df). Ignore the sign (+ or –) on your calculated *t* for comparison with the critical value in the table**. If your *df* are not in the table, use the next smaller value.**

|  |  |
| --- | --- |
| **df** | **P (α or *P*)** |
|  | **0.05** | **0.02** | **0.01** | **0.001** |
| **1** | 12.706 | 31.821 | 63.657 | 636.619 |
| **2** | 4.302 | 6.965 | 9.925 | 31.599 |
| **3** | 3.182 | 4.541 | 5.841 | 12.924 |
| **4** | 2.77645 | 3.747 | 4.604 | 8.610 |
| **5** | 2.570 | 3.365 | 4.032 | 6.869 |
| **6** | 2.447 | 3.143 | 3.707 | 5.959 |
| **7** | 2.365 | 2.998 | 3.499 | 5.408 |
| **8** | 2.306 | 2.896 | 3.355 | 5.041 |
| **9** | 2.262 | 2.821 | 3.250 | 4.781 |
| **10** | 2.228 | 2.764 | 3.169 | 4.587 |
| **11** | 2.201 | 2.718 | 3.106 | 4.437 |
| **12** | 2.179 | 2.681 | 3.055 | 4.318 |
| **13** | 2.160 | 2.650 | 3.012 | 4.221 |
| **14** | 2.145 | 2.624 | 2.977 | 4.141 |
| **15** | 2.131 | 2.602 | 2.947 | 4.073 |
| **16** | 2.120 | 2.583 | 2.921 | 4.015 |
| **17** | 2.110 | 2.567 | 2.898 | 3.965 |
| **18** | 2.101 | 2.552 | 2.878 | 3.922 |

*continued*

|  |  |
| --- | --- |
| **df** | **P (α or *P*)** |
|  | **0.05** | **0.02** | **0.01** | **0.001** |
| **19** | 2.093 | 2.539 | 2.861 | 3.883 |
| **20** | 2.086 | 2.528 | 2.845 | 3.850 |
| **21** | 2.080 | 2.518 | 2.831 | 3.819 |
| **22** | 2.074 | 2.508 | 2.819 | 3.792 |
| **23** | 2.069 | 2.500 | 2.807 | 3.768 |
| **24** | 2.064 | 2.492 | 2.797 | 3.745 |
| **25** | 2.060 | 2.485 | 2.787 | 3.725 |
| **26** | 2.056 | 2.479 | 2.779 | 3.707 |
| **27** | 2.052 | 2.473 | 2.771 | 3.690 |
| **28** | 2.048 | 2.467 | 2.763 | 3.673 |
| **29** | 2.045 | 2.462 | 2.756 | 3.659 |
| **30** | 2.042 | 2.457 | 2.750 | 3.646 |
| **40** | 2.021 | 2.423 | 2.704 | 3.551 |
| **50** | 2.009 | 2.403 | 2.678 | 3.496 |
| **60** | 2.000 | 2.390 | 2.660 | 3.460 |
| **80** | 1.990 | 2.374 | 2.639 | 3.416 |
| **100** | 1.984 | 2.364 | 2.626 | 3.390 |
| **1000** | 1.962 | 2.330 | 2.581 | 3.300 |
| **inf** | 1.960 | 2.326 | 2.576 | 3.291 |

**Table 3**. Critical values of Chi-square (**χ2**) for different values of Probability (α or *P*). These values are appropriate for analysis of a 2x2 table (with 1 degree of freedom).

|  |  |
| --- | --- |
| **Probability (α or *P*)** | **Critical value of χ2** |
| 0.05 | 3.84 |
| 0.025 | 5.02 |
| 0.02 | 5.41 |
| 0.01 | 6.63 |
| 0.005 | 7.88 |
| 0.0025 | 9.14 |
| 0.001 | 10.83 |