Laboratory Three

Introduction to Statistics and Data Analysis

Prelabs

There are **two** prior-to-lab exercises this week. The next week's prelab links are posted each Thursday afternoon. See "Submitting Catalyst Exercises" in this course manual; both exercises are due by Tuesday at 8:00 AM.

1. Study this Introduction to Statistics and Data Analysis lab description, Appendix A in this course manual, and BioSkills 3 (Reading Graphs) and 4 (Using Statistical Tests) in the textbook; then do the Lab 3 Prelab linked to the course website. (2 points)
2. Study the Excel Tutorial instructions in the Other Information folder of the course website, do the tutorial, and then take the Lab Excel Tutorial quiz. (2 points)

Note: Bring course manual Appendices A and B to lab.

Learning Objectives

Understanding basic data analysis and statistical inference is essential for almost any field related to biology. For example, the "Scientific Foundations for Future Physicians" report, issued in June 2009 by the Howard Hughes Medical Institute and the American Association of Medical Colleges, lists eight "competencies" expected of *entering* medical students. The first of these competencies relates directly to material introduced in this lab. Similarly, virtually every graduate school program or job in the biological sciences requires a working knowledge of statistics.

By the end of this lab, students should understand:

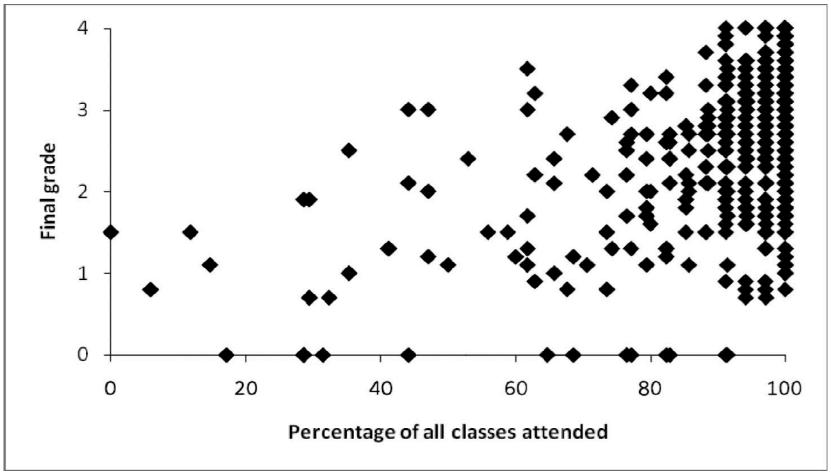
- The difference between categorical and continuous variables, and explanatory vs. response variables
- The role of the mean, standard error, and histograms as descriptive statistics
- The function of a *t*-test; how to perform one in Excel and how to interpret the output; how to make a bar chart with error bars in Excel and how to interpret it
- The function of a linear regression; how to perform one in Excel and how to interpret the output; how to make a scatter plot in Excel and how to interpret it; what an R^2 value means
- The nature of statistical significance; what a *p*-value means

I. Why do statistical testing?

Biologists test hypotheses by doing experiments or making observations and collecting data. In many cases, the data come from two contrasting groups—often an experimental treatment and a control. As you analyze the differences between the two groups, statistical testing helps you determine whether the difference you observed is due to chance or to the different conditions experienced by the groups.

For example, suppose you wanted to know if working for pay affects performance in Biology 180. You hypothesize that working many hours hurts performance, because it leaves less time for studying. You have data from a total of 568 students. The average grade for students who say they worked more than 10 hours per week was 2.46; the average grade for students who said they worked fewer than 10 hours per week for pay was 2.74. The difference is 0.28. Is this difference meaningful, or could it just be due to chance?

Now suppose that you also had data on the proportion of lectures attended by Biology 180 students over the course of the quarter, from two different quarters. When you plot the proportion of total classes attended on the *x*-axis of a graph and grade received on the *y*-axis, the data look like this:



Here, each data point represents one student. It looks as though there might be a trend—students who attend more classes appear to do better. But there's also a lot of variation. Is the trend real? Or could you get data like this just by chance? In this week's lab you'll learn some basic tools to answer questions like these.

II. Some basic ideas in data analysis and experimental design

A. Descriptive statistics

As the name implies, descriptive statistics describe data. They give you a feel for what the data say about the quantity you are measuring. There are three basic aspects:

1. What's the average?
   There are several ways to express "central tendency." The **mean** is the arithmetic average, calculated as the sum of all the observations divided by the number of observations. The median is the value such that 50% of the observations are greater and 50% are less.

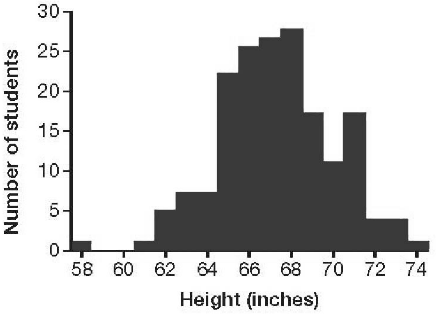
2. How much variation is there?
   There are also several ways of expressing "dispersion" or spread in data. The most common is the **standard deviation**. The higher the standard deviation, the more scattered the data are. If the data you are analyzing have a normal or bell-shaped distribution, then about 95% of the observations are within two standard deviations of the mean; about 68% of the observations are within one standard deviation of the mean.

   A **standard error** is the standard deviation estimated from a sample from a particular population. You'll almost always be analyzing standard errors.

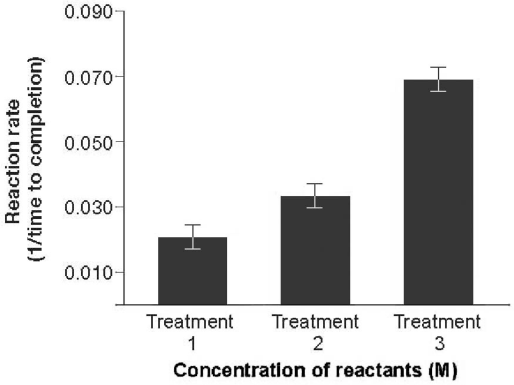
3. Visualizing data

It's often helpful to inspect data visually. Here are 3 approaches.

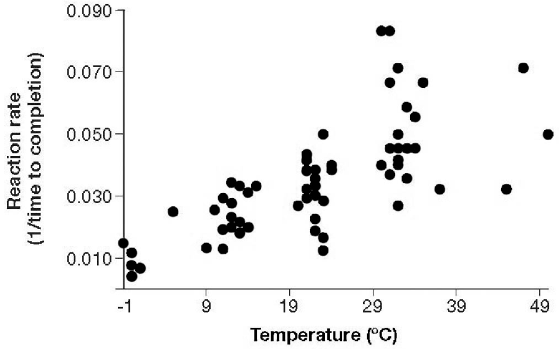
- A **histogram** (or frequency distribution) plots a single variable, with the values observed grouped into bins (or ranges) on the horizontal axis. The vertical axis records the number or frequency of values in each bin. Histograms are a good way to get a feel for how data are distributed over a range of values. For example, this histogram shows the distribution of heights in a class of male students.



- A **bar graph** plots the means from two or more groups—usually with the standard error of the mean indicated by vertical bars.



- A **scatter plot** graphs the relationship between two variables—one on the horizontal axis and one on the vertical axis.



B. Statistical testing: Comparing two means

Do students who work fewer than 10 hours/week really get better grades in Biology180 than students who work 10 or more hours/week? The problem is, your data represent a small sample of all the students who will ever take Biology 180. If you had data from a larger sample, it might be that the apparent difference of 0.28 would disappear—that if you repeated the "experiment" of analyzing Bio 180 grades over many quarters, there would actually be no difference in average grade based on how much people work for pay. A statistical test can tell you how likely this is to be true.

In this case, the statistical test is a four-step process:

1. Specify the **null hypothesis**.

   This is the hypothesis that there is actually no difference between the groups (in this case, that working has no effect on grade).

2. Calculate the **test statistic**.
   - This is a number that quantifies the magnitude of the difference between the groups.
   - There are many different types of test statistics—each is appropriate for a different type of data.
   - In the case of comparing two average values, one of the appropriate test statistics is called Student's *t*. The larger the value of Student's *t*, the larger the difference between the groups.

3. Determine the probability of getting the value of the test statistic just by chance. If the null hypothesis is correct and you did the "experiment" many times, how often would you get a value for the test statistic that is as large or larger than the one you actually got?
   - The answer comes from a reference distribution, a mathematical function that specifies the probability of getting each value of the test statistic when the null hypothesis is true.
   - When you use a computer program to calculate Student's *t*, it will also tell you the probability associated with that value, based on the reference distribution. This will be reported as the "*p* value."
   - The *p* value means that you would only get a difference as large as the one you observed, just by chance, a proportion *p* of the time.

4. Decide whether the result is "**statistically significant**."
   - By convention, biologists accept a difference as significant or "real" if there is a 5% or less chance of observing it by chance (meaning that $p < 0.05$).
   - If the probability that the observed difference is due to chance is more than 5%, then biologists accept the null hypothesis that there really is no difference between the groups.

C. Statistical testing: Comparing two variables

Do students who go to class every day do better than students who miss many lectures? Stated another way, is attendance in class a meaningful predictor of performance? Again, the data you have represent a small sample of all the students who will ever take Biology 180. If you had data from a larger sample, it might be that the apparent positive correlation would disappear—that if you repeated the "experiment" of analyzing Bio 180 grades over many quarters, there would actually be no relationship between attendance and grades. A statistical test can tell you how likely this is to be true.

Again, you can break the statistical test into a four-step process:

1. Specify the **null hypothesis**.

   In this case, the null is that there is no relationship between attendance and course grade.

2. Use a regression analysis to calculate the **line of best-fit.**

   A line of best-fit is a mathematical function that minimizes the total amount of distance between the data points and the line. In Bio180, most of the functions we'll be using are linear—meaning that they have the form $y = ax + b$, where a is the slope of the line and b is the y-intercept. (You'll be fitting lines using a technique called linear regression.)

   - If the slope of the line is 0, then there is no relationship between x and y.
   - If the slope is positive, it means that y increases as x increases.
   - If the slope is negative, it means that y decreases as x increases.

3. Determine the probability of getting the slope on the line of best-fit just by chance.

   If the null hypothesis is correct and you did the "experiment" many times, how often would you get a value for the slope that is as large or larger than the one you actually got?

   - As with a *t*-test, the answer comes from a reference distribution. This is a mathematical function that specifies the probability of getting each slope when the null hypothesis is true.
   - When you use a computer program to do a regression analysis, it will also tell you the probability associated with that slope, based on the reference distribution. This will be reported as the "*p* value."
   - The *p* value means that you would only get a slope as large as the one you observed, just by chance, a proportion *p* of the time.

4. Decide whether the result is "statistically significant."

   - By convention, biologists accept a relationship as significant or "real" if there is a 5% or less chance of observing a slope that large by chance (meaning that $p < 0.05$).
   - If the probability that the observed slope is due to chance is more than 5%, then biologist accept the null hypothesis that there really is no relationship between the groups.

Exercise 1: Comparing two means

Student names: ______________________________

____

Your TA will return your experimental design (termite) lab report. Review your hypothesis and results. Working in pairs, perform a *t*-test of that data. Complete this report and **attach a bar chart with error bars comparing the means**.

Question concerning termite behavior:	
Why is this question interesting?	
Hypothesis:	
Null hypothesis:	
Data collected last week (summarize here, or your TA may ask you to attach the termite lab's data table)	

Statistical test results and discussion:

Variable	Mean	Standard error	*n* (Sample size)
p-value:			

Was your hypothesis accepted? Why or why not?	
What does the *p*-value you obtained tell you about your results?	
Do your conclusions and interpretations today differ from those last week? Why or why not?	

Exercise 2: Comparing two variables

Working with the same partner, examine the results from the Personal Data Questionnaire in "Master Data." Develop a question you'd like to answer that requires testing the relationship between two **continuous** variables. Then do the analysis, complete this report, and **attach a scatter plot with best-fit line. Write the p-value on your graph.**

Question	
Why is this question interesting?	
Hypothesis	
Null hypothesis	

Which two variables are you comparing?

Explanatory (independent) variable:	*n* =
Response (dependent) variable:	*n* =

Regression analysis results:

R^2:	*p*-value for the slope:

Discussion:

Was your hypothesis supported? Why or why not?	
What does the R^2 value you obtained tell you?	
What does the *p*-value you obtained tell you?	