Introduction to Introductory Statistics
Lab Exercise, Week 5
Take Home (replaces a take home lecture quiz)

Lab by Kelly Pennoyer and Jeff Markert – Version 1.0 – October 2008

Statistics is a mathematical framework for the collection, analysis, interpretation or explanation, and presentation of data.

- **Descriptive statistics**- A set of methods for organizing, summarizing and presenting data (graphs, charts and tables)

- **Inferential statistics** – Body of methods for drawing conclusions about a population based on information available in a **sample** taken from a **population** (hypothesis testing or statistical significance). At its heart, inferential statistics are a set of conventions that allow scientists to decide whether a hypothesis is supported by a data set or not.

  Consider the data from our Osmosis lab a couple weeks ago. It was pretty obvious that worms placed in the high or low salt solutions were very different from worms in the control solutions at the end of the experiment. Most reasonable people would conclude that the different solutions had an effect.

  However what if the animals in the hypotonic and hypertonic solutions had only been just a little bit different from the controls? Statistical methods provide a set of standardized analytical methods that can be used to determine whether a difference is statistically significant the type of result you might get from random noise.

Important Terms[1],[2]

- **Population** – the set of all individuals or items of interest in a statistical study

- **Sample** – Part of the population from which information is collected

- **Parameter** – a descriptive measure or characteristic of the population – the true state of nature

  How is this use of the word population different from the definition used in lecture? Or is it?

  What is the population for the clam worms we studied a couple weeks ago? What is the sample?

  Name some well known parameters

What's the difference between a Parameter and an Estimator?

- **Variable** – A characteristic that varies from one person or thing to another
  What's an important variable in our clam worm data set?

- **Data** – Information obtained by obseving values of a variable
  - Two type of data – Quatitative (Numerical); and Qualitative (Non-Numerical)

Descriptive Measures

***Mean** (Average) - Sum of the observations divided by the number of observations
Generally the best measure of central location
Influenced by extreme observations (outliars)

- **Median** – Middle value of the observations when arranged from smallest to largest (if the number of observations is odd the median is the observation exactly in the middle of the ordered list; if it is even the meadian is the mean of the middle observations in the ordered list)
-
- **Mode** – the value that occurs more frequently in the observations

Examples:

3 15 46 64 64 623

Mean: (3 + 15 + 46 + 64 + 64 + 623)/6 = 815/6 = 135.8

Median: (46 + 64)/2 = 55

Mode: 64

Why do economists talk about median income more often than average income? Hint, what would happen if Bill Gates lived in your neighborhood?

- **Standard deviation** – Measure of variation within a population

Standard Deviation Example[3]*:*

Suppose we wished to find the standard deviation of the data set consisting of the values 3, 7, 7, and 19.

Step 1: find the arithmetic mean (average) of 3, 7, 7, and 19,

(3 + 7 + 7 + 19) / 4 = 9.

Step 2: find the deviation of each number from the mean,

3 - 9 = -6

7 - 9 = - 2

7 - 9 = - 2

19 _ 9 = 10.

Step 3: square each of the deviations, which amplifies large deviations and makes negative values positive,

$(- 6)^2 = 36$

$(- 2)^2 = 4$

$(- 2)^2 = 4$

$(10)^2 = 100$.

Step 4: find the mean of those squared deviations,

(36 + 4 + 4 + 100) / 4 = 36.

Step 5: take the non-negative square root of the quotient (converting squared units back to regular units),

The square root of 36 = 6 = the **Standard Deviation** of the sample. The squared value (36) is referred to as the **Variance**.

- **Normal Distribution** (bell shaped curve) Parametric statistics assume normal distribution.

Hypothesis Testing

As described above, statistics provide a more objective framework for deciding whether two population samples are really different from each other, or whether the results could have occurred by chance. Often, scientists are willing to consider two population samples to be statistically significant if the probability of obtaining a similar result by chance is 5% or less (abbreviated as $P<0.05$). Usually, in these kinds of analyses we develop a special hypothesis called the **Null Hypothesis**. In the case of our worm data, we can test two different null hypotheses. The first that the weight of worms placed in the hypotonic solution *is not s*ignificantly different from the weight of worms in the control solution. We can generate a similar null hypothesis comparing worms in the control solution to the hypertonic solution[4]. If we cannot show that the two samples are

statistically significantly different, we say that we *fail to reject* the null hypothesis. In other words, we find no evidence that the two samples are different from each other in a meaningful way.

Student's t test: A common parametric statistic that we will be using frequently in this class. A convenient way to think of a t-statistic is to consider it to be a way of isolationg signal from noise[5]. In our case, the signal would be the mean value in the control and experimental samples. The noise is the variablity around these means. If the variance of two samples is very very large, the 'noise' from these distributions is much more important than any signal we might get from differences in sample means.

Uses:

- A test of whether the mean of a normally distributed population has a value specified in a null hypothesis.
- A test of the null hypothesis that the means of two normally distributed populations are equal. Given two data sets, each characterized by its mean, standard deviation and number of data points you can use the student's t test to determine if the means are distinct.
    - Unpaired – individuals are randomly assigned into two groups, measured after an intervention (experimental manipulation) and compared with the other group.
    - Paired – Each member of the group has a unique relationship with a paticular member of the sample (before or after an intervention).

From the t value, a p-value can be found using a table of values of the student's t – test. These tables are based on theoretical calculations that consider the mean and variance of your samples (signal & noise). We'll talk more obout these kinds of tables when we do the corn genetics lab after Columbus Day. These days, we usually don't look at these tables directly, instead we rely on electronic versions. Microsoft Excel and other stats programs can calculate the t-statistic for you and compare this value to the table automatically – for our purposes, we suggest using the calculator on this web site:

http://graphpad.com/quickcalcs/ttest1.cfm

We also provide directions on doing t-tests in Excel at the end of this lab.

If the t-statistic is below the threshold chosen for statistical significance (usually $p < 0.05$) the null hypothesis (there is no difference between the two groups) is rejected and the alternate hypothesis is not rejected.

A result is called statistically significant if it is unlikely to have occurred by chance, we normally say that a result *is* significantly different if the p-value is less than 0.05.

Assignment (due October 20th)

Complete the questions in the text above before you leave today.

The, take a look at the worm data from last week. Use the data from from the ten minute mark. First, calculate the mean, median, and standard deviation the hypotonic, hypertonic, and control treatments.

Then do two separate t-tests, using Microsoft Excel or the link above. For the first, comparison, determine whether there is a statistically significant difference between worms in the control solution and the hypotonic solution. Next, determine whether there is a difference between worms in control solution and the hypertonic solution.

Appendix – Excel Methods

Getting a p-value using the students t-test in excel.

Start by putting your data into two rows or two columns

Click on an empty cell, then click on Insert and select "function" from the menu

You can choose to select from a category of functions- pick Statistical

Choose Ttest

Aray 1 will be- Solution A %Change 10 min. Highlight all of the group data.

Aray 2 will be- Solution B %Change 10 min. Highlight all of the group data.

Tails- 2 (two tailed distribution)

Type – 2 (unpaired t-test)

This will give you a p-value, if the p-value is less than 0.05 you reject the null hypothesis (the null hypothesis should always be that there is no change or no difference)

Get a Standard deviation of the mean for each of the time intervals, and solutions (mean, mode median etc are all similar)

Click on the insert menu tab, choose 'Insert Function'

You can choose to select from a category of functions- pick Statistical

Choose Stdev

Number 1 Highlight values for Solution A, 10min %Change

Number 2 is left blank

References

[1] Zar, J.H. 1999. Biostatistical Analysis. 4th edition. Prentice Hall, Upper Saddle River, NJ
[2] http://udel.edu/~mcdonald/statintro.html
[3] http://en.wikipedia.org/wiki/Standard_deviation
[4] We should mention that we could compare all three at once using a method called ANOVA, which is basically multiple simultaneous t-tests, but that's more than we need to get into for this exercise!
[5] http://www.socialresearchmethods.net/kb/stat_t.php