BIO 3A LABORATORY
Statistical Methods in Biology: Descriptive Statistics and the t-Test

Objectives

- To understand the tendency of numbers in a collected data set to be grouped around a specific value
- To be able to calculate the common measures of central tendency: mean, median and mode
- To understand the concept of "variation within a population"
- To understand the difference between precision and accuracy
- To estimate the confidence interval of a mean
- To use a t-test to test for a statistically significant difference between two means

Introduction

In biology we use statistical methods to help describe the characteristics of a data set (or set of measurements), to generalize about the whole population from a subset of samples, and to analyze sample data and differences between data sets.

The concept of creating a sample is integral to this procedure. A sample should be clearly representative of the population as a whole. If the sample is truly a random subset of the whole population, then we can draw fairly accurate conclusions related to that population.

Central Tendency Theorem

As you will see, there is a tendency for data to group themselves around a specific value. This is called "central tendency." One of the easiest measures of this tendency is called the **mean** (the average). It is generally expressed as

$$\overline{X} = \sum X / n$$

In this equation, the mean, "X bar", is equal to the sum of the measurements, divided by the total number of the measurements.

There are two other common measures of central tendency. The **median**, which is simply the middle measurement in a ranked list of the data; and the **mode**, is the most commonly occurring measurement in the sample.

Variation in the Population

The mean, median and mode are only a partial description of the data set. We also need to know how variable the data are. The **range** is one measure of data variability. The range is the difference between the largest and smallest measurement in the data set. The range may be a biased estimate of the distribution.

Precision and Accuracy

To most people, these two terms are equivalent. However, in statistics they have very different meaning. **Accuracy** is the closeness of the measurement to the true value. If the measurements are consistently high or low, the data are said to be biased. **Precision** is the closeness of the measurements to each other (or agreement with each other). If the deviation from the mean is small, then the precision is high.

Deviation from the mean

It is possible to calculate the variation of the data set from the mean. This is can be done is several ways. One of the most common is the **standard deviation** (sd). This measure carries the same units as the original measure and is an estimate of the variation (standard deviation) in the entire population. It is calculated as

$$sd = \sqrt{SS/N}$$

where N is equal to the number of samples and SS is the sum of the squared deviations from the mean

$$SS = \sum (\overline{X} - X)^2$$

This calculation is done by subtracting each measurement from the mean, squaring the result and adding up all of the results.

Confidence in an estimation

When we calculate the mean of a sample, we infer that this is an estimate of the entire population mean. We may want to know the precision of this estimate of the population mean. This can be expressed as the confidence interval, which is calculated from the standard error of the mean. The standard error of the mean is

$$se = sd / \sqrt{N}$$

where se is the standard error, s is the standard deviation and n is the sample size. Using the standard error, you can calculate the confidence interval (an interval or range with a stated confidence). This interval is usually stated as a percentage, i.e. 95% confidence interval. 95% of the data will fall within this interval. The 95% confidence interval is

$$95\% \text{ confidence interval} = \overline{X} \pm \text{s.e.}$$

where the confidence interval equals the mean plus or minus (that's the interval) the standard error times "t" which is a value called Student's t, obtained from a statistical table.

Comparing Statistical Populations: Two sample testing

Suppose you have data on a similar measurement for both males and females. Are they significantly different? In biology the level of significance is usually stated at 95%. In other words, is the overlap less than 5% (look at the confidence intervals). A typical way of making this comparison is called a t-test.

Statistical Methods Part I

We will complete this as a group in lab and then you will have to opportunity to complete it on **your own**. Use Excel for all of these exercises. You will create and turn in two figures as indicated in number 2 and 6 below.

1. Record the height and sex of all members of the lab class (we most likely did this the prior meeting)
2. Create a frequency histogram for combined height data (all males and females together → do not separate them into categories yet) with a complete figure caption. **This is Figure 1 for Part 1.**
3. Create a frequency histogram for male height data and female height data
4. Calculate mean, median, mode and range for each of these samples
5. Calculate standard deviation and standard error for each sample, and the 95% confidence interval.
6. On the histogram, place the mean, median, mode, standard deviation and confidence interval. Do the confidence intervals overlap? In other words, use Student's t-test to determine if males are significantly different from females (95% confidence, $p < 0.05$) on the height parameter. This will result in ONE figure. In the figure caption, state the results of the t-test. **This is Figure 2 for Part 1.**

Due Date: You will need to turn in Figure 1 (from #2 above) and Figure 2 (from #6 above) at the beginning of the next class meeting.

Statistical Methods Part II

You should work in groups of two for the data collection portion of this exercise. Use the form on the next page to collect your data. Measurements should be in centimeters (cm), measured with an accuracy of 0.1 cm.

Each student must turn in **their own** work for #3 below. Check the website for the due date of Part II.

1. Measure and record on the data sheet the sex and the lengths of both fifth (little) fingers of 100 individuals on campus.
2. Enter all data in Excel before the next class meeting. Turn in the raw data as an Excel file during the next class meeting.
    a. Do not put a space after each value
    b. Do not include units with each value
3. You must first select one of the following two possible hypotheses:

Possible hypotheses for fifth finger lab (You must choose one of these):

- The combined length of male fifth fingers is significantly greater than the combined length of female fifth fingers (Note: "combined" means add the length of the right and left for each subject)
- Mean length of the right fifth finger is significantly different from mean length of the left fifth finger.

4. Test one of the hypotheses above using *your data* (that's the 100 measurements you made) and *the combined data* (for all classes). These data sets will be posted on the website.

100 data points (the data you collected):

1. State the hypothesis you have chosen
2. Compute the Descriptive Statistics for the two means you are testing
3. On Figure 3, show the two frequency histograms for the two data sets you are testing
4. On each frequency histogram draw the mean and 95% CI
5. Run the appropriate t-test between the means and find the p value.

Large data set (this is the data set collected by all sections):

Now, for the large, whole class data set (downloadable from the website) repeat steps 2 through 5. This will be Figure 4 for Part II.

Finally, write a short paragraph in which you state the hypothesis and its significance with the smaller and larger data sets. What is effect of the larger number of samples?

Due Date: From Part II, You will need to turn in Figure 3 (from 100 data points you collected) and Figure 4 (the large data set from all sections) at the beginning of the following class meeting. You also need to include the short paragraph on "What is the effect of the larger number of samples?" Please see the website for the due date.

Fifth Finger Length Data (**in cm**) Date:_____

Researchers:_____

No.	Sex	Right	Left
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			
21			
22			
23			
24			
25			
26			
27			
28			
29			
30			
31			
32			
33			
34			
35			
36			
37			
38			
39			
40			
41			
42			
43			
44			
45			
46			
47			
48			
49			
50			

No.	Sex	Right	Left
51			
52			
53			
54			
55			
56			
57			
58			
59			
60			
61			
62			
63			
64			
65			
66			
67			
68			
69			
70			
71			
72			
73			
74			
75			
76			
77			
78			
79			
80			
81			
82			
83			
84			
85			
86			
87			
88			
89			
90			
91			
92			
93			
94			
95			
96			
97			
98			
99			
100			