Statistical Analysis Protocol (Biotech Series)

In the biotechnology pathway you will be required to use inferential statistics to analyze your data. Biostatistics (or biometry) is the application of statistics to a wide range of topics in biology. The science of biometry encompasses the design of biological experiments; the collection, summarization, and analysis of data from those experiments. The use of biometry in this course will help to prepare you for your Biotechnology research conducted your senior year, and it will also prepare you for college and career pathways. This protocol contains background information on the most common statistical tests and basic statistical verbiage. Additionally, there are two different dichotomous keys included that will guide you through the decision process to ensure that you use the correct test. Both keys will get you to the same result, so feel free to use either one; however, it is recommended that you use both to affirm you choice of statistics. If viewing this protocol online (click here) for a list of online statistical programs. The Statistics Guide (click here) is a very helpful resource that models the use of biometry with various data sets. It should be used in conjunction with this protocol.

- **P value (P)**: a P value of less than 0.05 is considered a **significant difference**. For example, a P value of 0.04 would mean that 4% of the time or less, we would observe this difference between the control and experimental groups due to chance alone. However, a P value of 0.10 would mean 10% of the time this difference could be due to chance, and not the independent variable in our experiment, and so the difference cannot be considered significant.
- **n:** the sample size.
- **Degrees of freedom (df):** Indicates the number of comparisons being tested and the number of replicates in each group.
- **Mean:** in common usage often called the average
- **Median:** the numerical value separating the higher half of a data sample from the lower half.
- **Mode:** is the value that occurs most often. If no number is repeated, then there is no mode for the list.
- **Range:** the difference between the largest and smallest values.
- **Standard deviation (SD):** measures the amount of variation or dispersion from the average. A low standard deviation indicates that the data points tend to be very close to the mean; a high standard deviation indicates that the data points are spread out over a large range of values.
- **Null Hypothesis (H_o)**: the opposite of the hypothesis you hope to support. If you can refute the null hypothesis statistically, then you can say there is a significant difference between your control and experimental groups, and thus say your hypothesis is supported.
- **Continuous data**: values are evenly spaced, such as age, weight, height, length of time until an event occurs, test scores, etc. This is commonly referred to as parametric data.
- **Ordinal data**: small finite values that are ordered, such as stage of disease, grade in school, numbered preferences, etc. May also be described as discrete or categorical and is non-parametric. Non-parametric statistical procedures are less powerful because they use less information in their calculation.
- **Nominal data**: named characteristics, such as eye color, gender, treatment group, etc. May also be described as discrete or categorical and is non-parametric. Non-parametric statistical procedures are less powerful because they use less information in their calculation.

Dichotomous Key for Choosing a Statistical Test #1

1. **What type of data are you working with?**
    a. My data is continuous, measured on a ratio or interval scale. A good example of this type of data would be measurements of the mass of individual organisms in a population of interest. If your data is continuous, **go to couplet 2.**

    b. My data is nominal, or ordinal. A good example of this type of data would be the counts of the number of individuals of each gender in a population of interest. The categories would be male and female and the counts would be the number of males and the number of females. If your data is discrete or categorical, you should use a **chi-square test.** A bar graph with frequency distributions is suitable for reporting a chi-square test.
2. **What type of question are you asking?**
    a. I am interested in whether there are differences in mean values between two or more groups of observations. A good example of this type of question would be testing for differences in the average circumference of trees measured in two separate areas. If you are testing for differences in mean values between two groups, you will use a **t-test**. Make sure to include the standard deviation with your mean value. If there are more than two groups, you will use an **analysis of variance (ANOVA).** A bar graph with frequency distributions to show data variance is suitable for reporting an ANOVA

    b. I am interested in whether there is a relationship between two of my measured variables. A good example of this type of question would be whether there is a relationship between the percent body fat of a person and their blood cholesterol levels. If you are interested in the relationship between two variables, **go to couplet 3**.

3. **What type of relationship are you looking for?**
    a. I am interested in predicting the value of one variable (my dependent variable) based on a relationship with another variable (my independent variable). A good example of an investigation of this type of relationship would be asking whether insects exposed to warmer temperatures grow more rapidly. The temperature at which the insects were reared would be my independent variable and the rate of growth (in mg/day) over the study period would be my dependent variable. If you are interested in predicting the value of one variable based on the value of a second variable, and you suspect the relationship is linear, you should use **regression analysis**. A fitted line graph on a scatter plot is suitable for reporting this test.

    b. I am interested in whether the values of one variable are associated with the values of a second variable in any way. A good example of an investigation of this type might be asking whether the concentrations of two nutrients in a series of soil samples are associated. We might measure nitrogen (N) and phosphorus (P) to see whether high values of N are associated with high values of P and low values of N are associated with low values of P (a positive relationship between the two variable). Conversely, high N may be associated with low P and low N may be associated with high P (an inverse or negative relationship between the two variables). If you are interested in the association between two variables, you should use **correlation analysis**. A scatter plot is suitable for reporting this test.

Dichotomous Key for Choosing a *Statistical Test #2*

1a. experiment has only 1 dependent variable, no independent variable……..……...go to 2
1b. experiment has at least 1 dependent variable and 1 independent variable……...……go to 4

2a. dependent variable is continuous data…….use *Student's T test for paired data*
2b. dependent variable is ordinal or nominal data…….…….go to 3

3a. dependent variable is ordinal data……...…….…..use *Wilcoxon Signed Rank test*
3b. dependent variable is nominal data…….……..…..……use *Rate,or Proportions*

4a. experiment has 1 dependent variable and more than 1 independent variable, or has more than 1 dependent variable……seek professional help to use *Multivariable or Multivariate analysis*
4b. experiment has 1 dependent and 1 independent variable……..go to 5

5a. dependent and independent variables are nominal data…….use *Chi-square test*
5b. dependent variable is nominal data, independent variable is continuous or ordinal, or dependent variable is continuous …….go to 6

6a. dependent variable is nominal data, independent variable is continuous or ordinal ……..…use *Logistic Regression*
6b. dependent variable is continuous…….……..go to 7

7a. independent variable is continuous…….……use *Correlation and Regression: T test and F test*
7b. independent variable is nominal or ordinal…….……..…….go to 8

8a. independent variable is nominal…….……..…go to 9
8b. independent variable is ordinal……..……...…use *ANOVA or Regression*

9a. there are only 2 groups in the experiment (example: males compared to females) use… *T test for 2 groups*
9b. there are more than 2 groups in the experiment…….. …..use *1 way ANOVA*