Biostatistics

What is Statistics?

Statistics can be divided into two major branches, data analysis and inference. Data analysis helps analyze statistical data (pretty straightforward). Inference allows you to draw conclusion about population parameters (means, standard deviation) from statistical information collected from the sample.

Data Analysis

Population- the whole group one desires to know about
Sample- the subset of a population one studies in order to determine information about the population
Mean- μ is the average for all values of the whole population and $\bar{x}$ of all values in the sample
Median- the middle value of data given by (n+1/2), is a much better measure of center than mean when outliers are involved
Variance of Sample- Measure of the difference of values from mean, where the difference between a value and a mean is squared. The variance is **independent** of sample size.

- The sample variance of a sample is equivalent to the sum of the squares of the differences divided by the sample size minus one. In all biological studies, it is safe to assume that you are dealing with sample variance.
- The population variance is rarely found in biology and is equivalent to sample variance except that you divide by population size
- The **standard deviation** σ is then defined as the square root of variance.

Variance is additive if samples are random. So if you take two samples of height and you add them, the variance will be the sum of the variance of each of the independent sample.

Inference

Inference is one of the most powerful applications of statistics. It allows you to make inferences such as the mean and standard deviation of the population given what you find a sample. For example, if you know that the reputed average score for USABO is 30 points, and you can take a sample of 40 people and find the average score is 40, you can then determine the probability that the actual average score is 30.

Steps to inference:

1. Determine if you have necessary conditions for inference

2. State your null hypothesis (the currently accepted population mean for example) and alternate hypothesis(what you think the mean is)
3. Calculate your parameter
4. Determine the probability of your parameter(multiply by two if your alternate hypothesis is that your value is not null).

T-test

What to do in a t-test: Determine the t-parameter, which represents the number of standard deviations your mean is from the mean distribution of your population. From your t-parameter, you use a t-chart and degree of freedom measurement (which is sample size minus one, or in the case of two samples, the sum of the two sample sizes minus one) and determine the probability of your result happening.

Determining the T parameters of different tests

One Variable T-test - Use this test if you want to compare a sample to a certain mean to see what the chance is that your sample could have that mean within normal variance. Example: your county publishes the normal amount of Fluorine that's supposed to be in your water. You measure the amount of fluorine in the water each day over a course of 10 days. You can use a one variable T-test to figure out if the mean of your measurements is statistically different from the amount that's supposed to be there.

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

$\bar{x}$ is the mean of your sample, μ_0 is the overall mean you are comparing to, s is the standard deviation, and n is the sample size.
Degrees of Freedom = n-1

Two Variable Difference of mean T-test - Use this test if you want to check if two samples have means that are the same as each other within normal variance. Example: You want to test a new drug that decreases facial hair group. Use a two variable T-test to see if there's a real different between the control and treated groups.

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}$$

where

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

X_1 and X_2 are the means of your two populations, S_1 and S_2 are the variances of your two populations, and n_1 and n_2 are the sample sizes.
Degrees of Freedom = $(n_1 – 1) + (n_2 – 2)$

Chi Square Test

Goodness of fit – Use this test if you want to check of data that's sorted in discrete categories fits within the expected distribution. Example: you know how many trees of different species are supposed to be in an area. Using Chi Square, you could test whether the number of trees was within normal range, or whether it was starting to change.

Determine your expected value for each category. Determine the values actually observed in each category. Subtract observed value for each category from expected value, and square the value, and divide by expected value. Add through all categories to determine the chi square parameter.
Degrees of Freedom = n – 1, where n is the number of categories.

Statistical Independence – Use this test if you want to see if two variables are statistically independent. For example, fruit flies can be male or female and have red or white eyes.

First determine your expected values. Your expected value for male fruit flies with red eyes is equal to the proportion of male flies multiplied by the proportion of red-eyed flies (and follow this with the rest of the categories). Then, just follow the procedure for Goodness of Fit.
Degrees of Freedom = (n_1 -1)(n_2 -1), where n_1 is the number of categories for the first trait (two genders) and n_2 is the number of categories for the second trait (two eye colors)

Significance

A result is considered statistically if the p value is less than 0.05 for USABO. To find the p value, look at the degrees of freedom and the t or chi square value calculated and find the p value that corresponds to both of those values with a provided table.

Errors

Type I Error- A type I error is when you reject a true null hypothesis

Type II Error- A type II error is when you fail to reject a false null hypothesis

Biological analogs of Type I and II errors are specificity and sensitivity. Specificity is analogous to a Type I Error, it represents the possibility of not have a disease and still being diagnosed as having it. Sensitivity is equivalent to a Type II Error, it represents the possibility of having the disease and not being diagnosed.

What does it mean if your disease test is highly sensitive and has low specificity?

Confidence Interval

The confidence interval shows the possible values of the actual population mean given the sample population mean within some probability (95% for USABO)

To determine the confidence, from the standard deviation and the t parameter for probability divided by 2 (so look for t parameter at your sample degrees of freedom and $p<0.975$). Multiply t parameter by standard deviation of sample and add and subtract value for confidence interval