Statistics in Biology

BI-102 (Fall 2007)
Professor J.A. LaMack
Modified by Dr. C. S. Tritt

Why do we need statistics?

- *Every* measurement you will ever make has some error associated with it.
- Statistics allow us to determine whether differences that we observe were real or just due to random error.
- In general, the more data you take, the better you can make such decisions.

Null Hypothesis

- All statistical tests begin with the formation of a *Null Hypothesis*.
- This is generally a statement that you would like to demonstrate is *false*.

Case 1: Discrete Treatment Levels

- There are two or more "categories" that the independent variable is divided into.
- Multiple measurements of the dependent variable are made at each level (these are called replicates).

Scenario

- Hypothesis: there is no difference in the height that people can jump before and after 8:00am.
- Dependent variable: height
- Independent variable: time of day (2 levels)
- Various potential control and confounding variables

Data Set 1 (3 replicates)

Height (inches)	
Before 8:00am	After 8:00am
22	25
26	28
27	26
Average = 25.00	Average = 26.33

Is there a difference?

Answer: not that we can find

Data Set 2 (18 replicates)

Is there a difference?

Answer: yes

Now we can state that there is a difference. This data included the same first three subjects, we just included more total subjects. Furthermore, the averages are exactly the same.

Height (inches)	
Before 8:00am	After 8:00am
22	25
26	28
27	26
22	25
26	28
27	26
22	25
26	28
27	26
22	25
26	28
27	26
22	25
26	28
27	26
22	25
26	28
27	26
Average = 25.00	Average = 26.33

How to do the Statistics

- For this type of categorical data, we will use a test called a *t-test* to determine if there is a difference between the two categories.
- The null hypothesis for a t-test is that there is *no* difference between the two groups.

The p-value

- Most statistical tests yield a result known as a *p-value*. This is the number you look at to make your decision.
- The p-value reflects the likelihood that you are making a mistake if you reject the null hypothesis.
- If $p < 0.05$, it is considered safe to reject the null hypothesis ("statistical significance").
- So, for a t-test, if $p < 0.05$, you can conclude that the two groups are different.

Demonstration: How to do a t-test

- Follow along…open the Statistics Demo and go to the Raw Data Set 1 worksheet (using the bottom tabs)
- Results
  - Data set 1: p = 0.492
  - Data set 2: p = 0.034

Be careful of the conclusions you make

- If $p < 0.05$, you can claim to have found a difference. *Always include the p-value when making a conclusive statement*.
- If $p > 0.05$, you can't say why
  - There may have been an effect that you just couldn't discern because your data were too noisy or not enough data points
  - There may have been no difference
- See handout for examples of acceptable and unacceptable written conclusions.

Display this type of data using a bar graph with error bars

Use Chart Tools | Layout | Error Bars to display error bars.

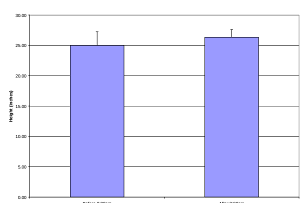

Figure 1: Jumping heights measured in subjects before and after 8:00am. Error bars represent standard deviation. n = 19.

Case 2: Continuous Independent Variable with a Predicted Linear Effect

- Sometimes, the independent variable has an infinite number of possible levels, rather than 2 or 3 distinct "categories".
- If we predict that there is a linear relationship between the dependent and independent variables, we can perform an analysis called "linear regression".

Linear Regression Basics

- We are fitting the data to a mathematical model:
  $Y = \beta_0 + \beta_1 X$
  - Y: Dependent variable
  - X: Independent variable
  - β_1: The slope of the line
- If Y actually depends on X, then the slope of the line will be non-zero.
- Therefore, the null hypothesis for linear regression is: $\beta_1 = 0$.
- R-squared: a second measure often used—only tells how close the data were to the regression line, not the strength of the relationship between the variables.

Scenario

- We hypothesize that jumping height depends on body weight.
- To test the hypothesis, we gather a group of subjects and measure the body weight and jumping height of each.
  - Dependent variable: jumping height
  - Independent variable: body weight
  - Control & confounding variables: many possible, but body height would be one

Demonstration: How to do perform linear regression analysis

- Follow along using the Raw Data Set 2 worksheet
- Warning: you may have to add the Analysis Toolpak
- Results:
  - $p = 1.26 \times 10^{-8}$
  - R-squared = 0.825
  - slope = -0.104
- Conclusion: since $p < 0.05$, reject the null hypothesis that there is no relationship. We conclude that there is a negative linear relationship between jumping height and body weight.

Data are plotted with the regression line

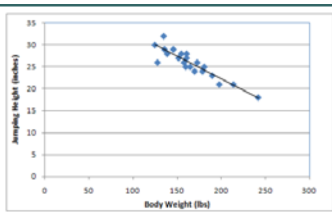

Figure 3: Relationship between jumping height and body weight. The solid line is the linear regression fit to the data collected.