Data Explorer

hhmi BioInteractive

Data Explorer

Tool

Statistical Analysis Selection Guide

Before you choose a statistical analysis, identify the variables you are interested in analyzing. Choose a statistical analysis based on the variables and your purpose.

Describe important features of the data

Descriptive statistics

- **What it uses:** One categorical or numerical variable.
- **When to use it:** To find numbers (mean, median, etc.) that summarize features of the variable and its distribution.

Analyze one group

Chi-square test of goodness of fit

- **What it uses:** One categorical variable.
- **When to use it:** To test whether sample data for the variable come from a specific population distribution.
- **Assumptions:**
  - The observations are independent and randomly selected.
  - All the expected counts are greater than or equal to 5.
  - The degrees of freedom (df) calculated by *Data Explorer* are appropriate for your hypothesis. *Data Explorer* does not calculate the correct df for Hardy-Weinberg and other intrinsic hypotheses in which df is reduced because the data are used to estimate parameters.

Compare groups

Chi-square test of independence

- **What it uses:** Two categorical variables.
- **When to use it:** To test whether the variables are associated. Compares counts observed in the samples to those you'd expect if the variables were independent.
- **Assumptions:**
  - The observations are independent and randomly selected.
  - All the expected counts are greater than or equal to 5.

Two-sample *t* test assuming equal variance (Student's *t* test)

- **What it uses:** One numerical variable and one categorical variable (with only two values).
- **When to use it:** To test whether two populations have the same *mean* for a variable. The variable is assumed to have the same variance in the populations.
- **Assumptions:**
  - The observations in each sample are independent and randomly selected.
  - The numerical variable is normally distributed in each population. You can examine the distribution of each sample using the histogram in the "Visualize" tab.
  - The variance (standard deviation) of the numerical variable is the same in both populations.

Two-sample *t* test *not* assuming equal variance (Welch's *t* test)

- **What it uses:** One numerical variable and one categorical variable (with only two values).
- **When to use it:** To test whether two populations have the same *mean* for a variable. The variable is *not* assumed to have the same variance in the populations.

- **Assumptions:**
  - The observations in each sample are independent and randomly selected.
  - The numerical variable is normally distributed in each population. You can examine the distribution of each sample using the histogram in the "Visualize" tab.

Paired *t* test

- **What it uses:** Two groups of paired observations (for example, observations before and after a treatment).
- **When to use it:** To test whether the mean difference between paired observations is 0.
- **Assumptions:**
  - Each observation in the first group is paired with an observation in the second group.
  - The pairs of observations are independent and randomly selected.
  - The differences between the pairs of observations are normally distributed in the population. You can examine the distribution of each group by adding a column of the differences in the data sheet and using the histogram in the "Visualize" tab.

One-way ANOVA assuming equal variances (Fisher's ANOVA)

- **What it uses:** One numerical variable and one categorical variable.
- **When to use it:** To test whether two or more populations have the same *mean* for a variable. The variable is assumed to have the same variance in the populations.
- **Assumptions:**
  - The observations in each sample are independent and randomly selected.
  - The numerical variable is normally distributed in each population. You can examine the distribution of each sample using the histogram in the "Visualize" tab.
  - The variance (standard deviation) of the numerical variable is the same in all the populations.

One-way ANOVA *not* assuming equal variances (Welch's ANOVA)

- **What it uses:** One numerical variable and one categorical variable.
- **When to use it:** To test whether two or more populations have the same *mean* for a variable. The variable is *not* assumed to have the same variance in the populations.
- **Assumptions:**
  - The observations in each sample are independent and randomly selected.
  - The numerical variable is normally distributed in each population. You can examine the distribution of each sample using the histogram in the "Visualize" tab.

Mann-Whitney *U* test (independent samples)

- **What it uses:** One numerical variable and one categorical variable (with only two values).
- **When to use it:** To test whether two populations have the same *median* for a variable.
- **Assumptions:**
  - The observations in each sample are independent and randomly selected.
  - The distributions of the numerical variable in both populations have the same shape. You can examine the distributions of each sample using the histogram in the "Visualize" tab. To compare the shape of the distributions, focus on whether there is one hump or multiple humps and whether the shape is skewed to the left or the right.

Wilcoxon signed rank test (paired samples)

- **What it uses:** Two groups of paired observations (for example, observations before and after a treatment).
- **When to use it:** To test whether the median difference between paired observations is 0.
- **Assumptions:**
  - Each observation in the first group is paired with an observation in the second group.

- The pairs of observations are independent and randomly selected.
- The distribution of the differences between the two populations is symmetrical. You can examine the distribution of the differences between the two groups by adding a column of the differences in the data sheet and using the histogram in the "Visualize" tab.

Correlate variables

Linear regression

- **What it uses:** Two numerical variables.
- **When to use it:** To model the relationship between the variables. Generates an equation that predicts values of one variable (*Y*) based on the other variable (*X*). This assumes that changes in *X* lead to changes in *Y*.
- **Assumptions:**
  - The observations are independent of each other.
  - For any value of the independent variable (*X*), the values of the dependent variable (*Y*) are normally distributed and have the same standard deviation (variance).
  - The relationship between the variables fits a straight line.

Linear correlation

- **What it uses:** Two numerical variables.
- **When to use it:** To determine the strength and direction of the relationship between the variables. Does not assume that changes in one variable lead to changes in the other.
- **Assumptions:**
  - The observations are independent of each other.
  - For any value of one variable, the values of the other variable are normally distributed and have the same standard deviation (variance).
  - The relationship between the variables fits a straight line.