How many variables are you interested in?

One Variable
What kind of variable is it?

Two Variables
What kind of variables are they?

Both Categorical (two grouping variables)
(e.g. school year & favorite ice cream, or Political party & Ethnicity)

One Quantitative, One Categorical
(e.g. height & gender, or income & political party)

Both Quantitative
(e.g. height & weight, or income & age)

Both Categorical

Hypothesis test:
χ^2 test of Independence
Are the observed proportions for one variable independent from the other variable?

H_0: Variable A is independent of variable B
H_a: Variable A is not independent of variable B

If A is independent, than the proportion in each level of A will be the same for each level of B as they are in general.

Test statistic: χ^2 -value

$$\chi^2 = \sum \frac{(Obs. - Exp.)^2}{Exp.}$$

In this case, sum is over all cells in the contingency table (A in rows, B in Col)

$$df = (J - 1) \cdot (K - 1)$$

where df is the degrees of freedom and J is the number of groups in variable A (rows), and K is the number of groups in variable B (cols).

p-value: Compare χ^2 value to χ^2 distribution with degrees of freedom, df (done with χ^2 table or technology)

tests are only one-sided (one-tailed)

One Quantitative, One Categorical

The Categorical Variable has Exactly Two Levels
(e.g. treatment/control, or male/female)

2-sample t-test

Paired t-test
Two measurements for each case
e.g. pre & post test, left & right hand

Assumptions/Conditions
1. The sample is random
2. The differences between measurements are normally distributed

Standard Error

$$SE = \frac{s_{\text{diff}}}{\sqrt{n}}$$

where s_{diff} is the standard deviation of the differences and n is the number of cases

Hypothesis test: Paired t-test
Is the average difference between measurements in the population different from 0?

$H_0 : \mu_{\text{diff}} = 0$

$H_a : \mu_{\text{diff}} \neq 0$

where μ_{diff} = Average of $x_2 - x_1$

Test statistic: t value

$$T_{df} = \frac{\bar{x}_{\text{diff}} - 0}{SE}$$

$$df = n - 1$$

p-value: Compare T_{df} to t-distribution with degrees of freedom, df (t-table)
Use 1 or 2 tailed, depending on H_a

Confidence Interval:

$$\bar{x}_{\text{diff}} \pm t^* \cdot SE$$

where t^* is the cutoff value corresponding to the confidence level.
Use a t-table and appropriate df e.g. for 95% CI, with $df = 10$ use $t^* = 2.23$

Independent Samples t-test
Comparing two independent groups
e.g., treatment & control, male & female

Assumptions/Conditions
1. The groups are independent
2. The sample is random
3. The quant. variable is normally distributed

Standard Error
SE can be calculated in a couple of different ways, i.e. Pooled or Unpooled.

When the s.d. for each group is roughly equal use pooled, otherwise use unpooled.

Consult another reference for these formulas.

Hypothesis test: 2 sample t-test
Is the average in one group of the population different from another group?

$H_0 : \mu_1 = \mu_2$

$H_a : \mu_1 \neq \mu_2$

Test statistic: t-value

$$T_{df} = \frac{\bar{x}_1 - \bar{x}_2}{SE}$$

(SE and df calculated differently depending on the sample sizes of the two groups and the standard deviations)

p-value: Compare T_{df} to t-distribution with degrees of freedom, df (t-table)
Use 1 or 2 tailed, depending on H_a

The Categorical Variable has more than two levels
(e.g. car manufacturer, or year in college)

ANOVA
Test to see if all of the group means are the same.

Assumptions/Conditions
1. Groups (levels) are independent
2. The quant. variable is normally distributed
3. Population variances are equal across levels

Hypothesis test: ANOVA
Is the average in at least one group of the population different from all the other groups?

$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots$

H_a : At least one mean is different

Test statistic: F-value

$$F = \frac{MSG}{MSE}$$

$$df_G = J - 1 \ (\text{Groups})$$
$$df_T = n - 1 \ (\text{Total})$$
$$df_E = df_T - df_G \ (\text{Error})$$

Calculations:

$$SSE = SST - SSG$$
$$MSG = SSG/df_G$$
$$MSE = SSE/df_E$$

p-value: Compare F to F-distribution with degrees of freedom, df_G and df_E using technology.
Only 1 tailed tests.

If null hypothesis is rejected, use t-tests to compare pairs of groups. Adjust the significance level by dividing by the number of tests done to avoid compounding Type I Errors

---

Both Quantitative

Simple Linear Regression

Assumptions/Conditions
1. "linearity"
2. Independence of residuals (random resid plot)
3. Normality of residuals
4. Equal variances (equal spread around the line)

Linear Model: $\hat{Y}_i = \beta_0 + \beta_1 \cdot X_i$

Residual: $e_i = Y_i - \hat{Y}_i$

r (correlation) gives strength of the relationship
R^2 gives percent of variation explained by the model
β_1 (slope) measures the size of the relationship, e.g. how much Y is expected to change for a unit change in X
β_0 (intercept) is the value expected when $X = 0$
e_i measures the deviation between the model and the observed values.

Hypothesis test: Regression
Is the slope (or intercept) in the population different from 0?

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

Test statistics:
The entire model is tested with an F-statstic.
The slope and intercept are tested with t-tests.
F and T values are calculated using technology.

p-value: Compare to F and t distributions using technology.
p-values are interpreted in the usual way.