http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/tests-of-means/why-use-paired-t/

WHEN TO USE WHICH STATISTICAL TEST.

Awesome statistics resource

<http://sites.nicholas.duke.edu/statsreview/sampling/>

Also as we repeat experiments again and again (keep gathering more samples), the variance moves towards zero i.e. uncertainty decreases.

**HYPOTHESIS TESTING** IS USED WHEN ONE WANTS TO MAKE DECISIONS REGARDING THE POPULATION ON BASIS OF RESULTS OBTAINED FOR A SAMPLE OF THIS POPULATION.

**HYPOTHESIS TESTING TECHNIQUES ARE: 1) ANOVA TEST i.e. Analysis of Variance test 2) Chi Square test of Independence**

**NESARC DATA SAMPLE:**

Though 13.9 > 13.2 the difference is not significant to reject NULL hypothesis.

Note: Null hypothesis is true until the evidence suggests otherwise.

we get a probability of 0.17 for above. It means that for sample of 100, we would be wrong 17% of the time in rejecting the NULL hypothesis.

How do we conclude what probability value will instill confidence regarding decisions.

This is where p-value or probability value comes into picture.

commonly used cut-off is p< 0.05 or p < 5%.

Basically p-value measures probability of Null hypothesis. Thus p>0.05 means accept NULL hypothesis. Else reject Null hypothesis.

**p-value is measure of number of times we would be WRONG in rejecting the NULL hypothesis (Ho). If we are wrong less than 5% times we reject the NULL HYPOTHESIS.**

**For the NESARC dataset we found p = 17% therefore:**

Categorize your quantitative explanatory variable into 2 categories with level 0 and 1.

ANOVA TEST question:

Is the number of cigarettes smoked associated with depression?

In statistics terms is the mean number of cigarettes smoked is same among individuals with or without depression?

Also as we repeat experiments again and again (keep gathering more samples), the variance moves towards zero i.e. uncertainty decreases.

For confidence interval of 95%, we have t value as 1.96. It changes as the % confidence interval changes.

<http://www.dummies.com/how-to/content/how-to-calculate-a-confidence-interval-for-a-popu0.html>

You estimate the population mean,

mage0.png by using a sample mean, mage1.png plus or minus a margin of error. The result is called a *confidence interval for the population mean,*mage2.png

The formula for a confidence interval for one population mean in this case is

mage5.png

is the critical *t\**-value from the *t*-distribution with *n*– 1 degrees of freedom (where *n* is the sample size).

If you don’t know your population [mean](http://www.statisticshowto.com/mean/)(μ) but you do know the [standard deviation](http://www.statisticshowto.com/what-is-standard-deviation/) (σ), you can find a confidence interval for the population mean, with the formula:  
x̄ ± z\* σ / (√n),

z statistic is (sample mean - pop mean)/sample std

**Z statistic vs T statistic.**

If n > 30 then sample std (s) is almost equal to population std (sigma). Therefore the distribution is normal distribution (bell curve) and we can use z statistic to get a probablity of getting a sample mean (x) that is located at extreme right as seen above.

But if n < 30 then the distribution is a T distribution and one should use t-table to find probability of getting sample mean at the extreme right as seen above.

Box plot :

1st quartile is midpoint or median of 1st 50% values

3rd quartile is median of last 50% values

What is the standard error of the mean?

The standard error of the mean (SE of the mean) estimates the variability between sample means that you would obtain if you took multiple samples from the same population. The standard error of the mean estimates the variability between samples whereas the standard deviation measures the variability within a single sample.

For example, you have a mean delivery time of 3.80 days with a standard deviation of 1.43 days based on a random sample of 312 delivery times. These numbers yield a standard error of the mean of 0.08 days (1.43 divided by the square root of 312). Had you taken multiple random samples of the same size and from the same population the standard deviation of those different sample means would be around 0.08 days.

Use the standard error of the mean to determine how precisely the mean of the sample estimates the population mean. Lower values of the standard error of the mean indicate more precise estimates of the population mean. Usually, a larger standard deviation will result in a larger standard error of the mean and a less precise estimate. A larger sample size will result in a smaller standard error of the mean and a more precise estimate.

Minitab uses the standard error of the mean to calculate the confidence interval, which is a range of values likely to include the population mean.