T-Tests (Part - 1)

T-tests

A t-test is an analysis of two populations means through the use of statistical examination

Purpose

A fairly common activity in Data Analytics is the comparison of two groups.

Some Background

Hypothesis

A hypothesis is a tentative explanation based on observations you have made. Example: Men's hands are larger than women's hands OR adding fertilizer to a plant makes it grow better OR Crime rates in southern states are higher than northern states.

Null hypothesis

The actual null hypothesis is a more formal statement of your original hypothesis. The null hypothesis is usually written in the following form:
There is no significant difference between population A and population B.

Examples

- **There is no significant difference in hand size between males and females.**
- **There is no significant difference in the growth of fertilized plants vs. unfertilized plants.**
- **There is no significant difference in the crime rates in southern states and northern states.**

Why?

The reason we write it in this form is that scientists are basically skeptics and their goal is to prove a hypothesis false. In fact, you can never really prove that a hypothesis is true.

t-test

We use this statistical test to compare our sample populations and determine if there is a significant difference between their means. The result of the t-test is a 't' value; this value is then used to determine the p-value.

Why statistical tests?

If we cannot use a statistical test (doesn't have to be a t-test) to determine whether a significant difference exists, then it becomes difficult to convince other scientists that your research is worth anything.

p-value

- The p-value is the probability that 't' falls into a certain range.
- In other words this is the value you use to determine if the difference between the means in your sample populations is significant.
- For our purposes, a p-value < 0.05 suggests a significant difference between the means of our sample population and we would reject our null hypothesis.
- A p-value > 0.05 suggests no significant difference between the means of our sample populations and we would not reject our null hypothesis.

Example 1

For example, consider that an analyst wants to study the amount that Pennsylvanians and Californians spend, per month, on clothing. It would not be practical to record the spending habits of every individual (or family) in both states, thus a sample of spending habits is taken from a selected group of individuals from each state. The group may be of any small to moderate size - for this example, assume that the sample group is 200 individuals.

The average amount for Pennsylvanians comes out to $500; the average amount for Californians is $1,000. The t-test questions whether the different between the groups is representative of a true difference between people in Pennsylvania and people in California in general or if it is likely a meaningless statistical difference. In this example, if, theoretically, all Pennsylvanians spent $500 per month on clothing and all Californians spent $1,000 per month on clothing, it is highly unlikely that 200 randomly selected individuals all spent that exact amount, respective to state. Thus,

if an analyst or statistician yielded the results listed in the example above, it is safe to conclude that the difference between sample groups is indicative of a significant difference between the populations, as a whole, of each state.