**Types of Errors in Hypothesis Testing**

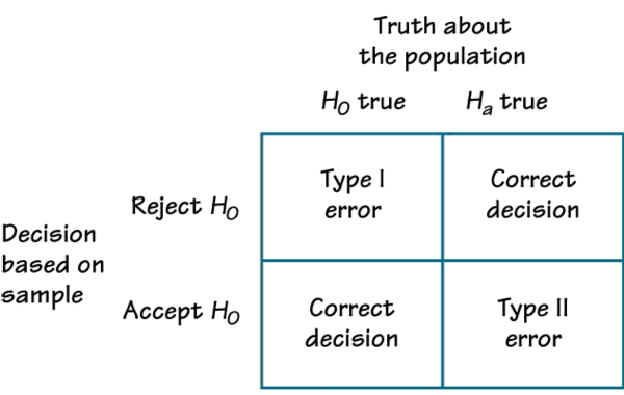
Now we have defined a basic Hypothesis Testing framework. It is important to look into some of the mistakes that are committed while performing Hypothesis Testing and try to classify those mistakes if possible.

Now, look at the Null Hypothesis definition above. What we notice at the first look is that it is a statement subjective to the tester like you and me and not a fact. That means there is a possibility that the Null Hypothesis can be true or false and we may end up committing some mistakes on the same lines.

There are two types of errors that are generally encountered while conducting Hypothesis Testing.

* **Type I error**: Look at the following scenario – A male human tested positive for being pregnant. Is it even possible? This surely looks like a case of False Positive. More formally, it is defined as the incorrect rejection of a True Null Hypothesis. The Null Hypothesis, in this case, would be – Male Human is not pregnant.
* **Type II error**: Look at another scenario where our Null Hypothesis is – A male human is pregnant and the test supports the Null Hypothesis.  This looks like a case of False Negative. More formally it is defined as the acceptance of a false Null Hypothesis.

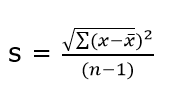
The below image will summarize the types of error :



## T-tests

T-tests are very much similar to the z-scores, the only difference being that instead of the Population Standard Deviation, we now use the Sample Standard Deviation. The rest is same as before, calculating probabilities on basis of t-values.

The Sample Standard Deviation is given as:



where n-1 is the Bessel’s correction for estimating the population parameter.

Another difference between z-scores and t-values are that t-values are dependent on Degree of Freedom of a sample. Let us define what degree of freedom is for a sample.

**The Degree of Freedom –** It is the number of variables that have the choice of having more than one arbitrary value. For example, in a sample of size 10 with mean 10, 9 values can be arbitrary but the 1oth value is forced by the sample mean.

Points to note about the t-tests:

1. Greater the difference between the sample mean and the population mean, greater the chance of rejecting the Null Hypothesis. Why? (We discussed this above.)
2. Greater the sample size, greater the chance of rejection of Null Hypothesis.

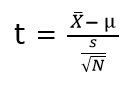
## 7. Different types of t-tests

#### 7.1 1-sample t-test

This is the same test as we described above. This test is used to:

* Determine whether the mean of a group differs from the specified value.
* Calculate a range of values that are likely to include the population mean.

For eg: A pizza delivery manager may perform a 1-sample t-test whether their delivery time is significantly different from that of the advertised time of 30 minutes by their competitors.



where, **X(bar)** = sample mean

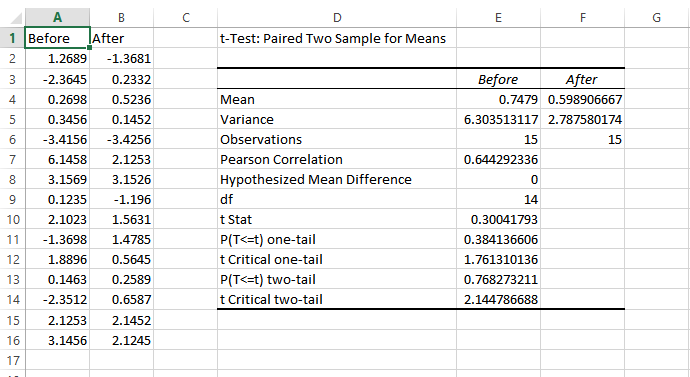
**μ** = population mean

**s** = sample standard deviation

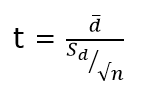
**N** = sample size

#### 7.2 Paired t-test

Paired t-test is performed to check whether there is a difference in mean after a treatment on a sample in comparison to before. It checks whether the Null hypothesis: The difference between the means is Zero, can be rejected or not.



The above example suggests that the Null Hypothesis should not be rejected and that there is no significant difference in means before and after the intervention since p-value is not less than the alpha value (o.o5) and t stat is not less than t-critical. The excel sheet for the above exercise is available [here](https://drive.google.com/open?id=0ByAvlBzuj2TgQ0M0U2lVZnZuams).



where, **d (bar)** = mean of the case wise difference between before and after,

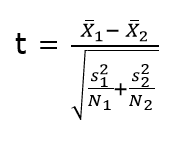
https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/01/30060715/Image_1.png= standard deviation of the difference

**n** = sample size.

#### 7.3 2-sample t-test

This test is used to determine:

* Determine whether the means of two independent groups differ.
* Calculate a range of values that is likely to include the difference between the population means.



The above formula represents the 2 sample t-test and can be used in situations like to check whether two machines are producing the same output. The points to be noted for this test are:

1. The groups to be tested should be independent.
2. The groups’ distribution should not be highly skewed.

where, **X1 (bar)** = mean of the first group

https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/01/30060909/Image_2.png**=**represents1st group sample standard deviation

https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/01/30060913/Image_3.png= represents the 1st group sample size.

#### 7.4 Practical example

We will understand how to identify which t-test to be used and then proceed on to solve it. The other t-tests will follow the same argument.

**Example:** A population has mean weight of 68 kg. A random sample of size 25 has a mean weight of 70 with standard deviation =4. Identify whether this sample is representative of the population?

#### Step 0: Identifying the type of t-test

Number of samples in question = 1

Number of times the sample is in study = 1

Any intervention on sample = No

Recommended t-test = 1- sample t-test.

Had there been 2 samples, we would have opted for 2-sample t-test and if there would have been 2 observations on the same sample, we would have opted for paired t-test.`

#### Step 1: State the Null and Alternate Hypothesis

**Null Hypothesis:**The sample mean and population mean are same.

**Alternate Hypothesis:** The sample mean and population mean are different.

#### Step 2: Calculate the appropriate test statistic

df = 25-1 =24

t= (70-68)/(4/√25) = 2.5

Now, for a 95% confidence level, t-critical (two-tail) for rejecting Null Hypothesis for 24 d.f is 2.06 . Hence, we can reject the Null Hypothesis and conclude that the two means are different.

You can use the t-test calculator [here](http://www.danielsoper.com/statcalc/calculator.aspx?id=98).

## 8. ANOVA

ANOVA (Analysis of Variance) is used to check if at least one of two or more groups have statistically different means. Now, the question arises – Why do we need another test for checking the difference of means between independent groups? Why can we not use multiple t-tests to check for the difference in means?

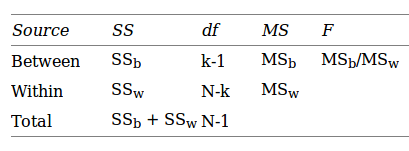
The answer is simple. Multiple t-tests will have a compound effect on the error rate of the result. Performing t-test thrice will give an error rate of ~15% which is too high, whereas ANOVA keeps it at 5% for a 95% confidence interval.

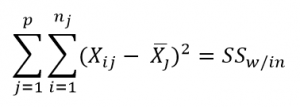
To perform an ANOVA, you must have a continuous response variable and at least one categorical factor with two or more levels. ANOVA requires data from approximately normally distributed populations with equal variances between factor levels. However, ANOVA procedures work quite well even if the normality assumption has been violated unless one or more of the distributions are highly skewed or if the variances are quite different.

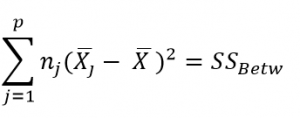
ANOVA is measured using a statistic known as F-Ratio. It is defined as the ratio of Mean Square (between groups) to the Mean Square (within group).

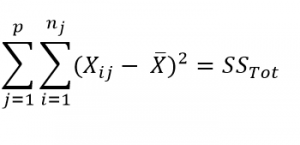
Mean Square (between groups) = Sum of Squares (between groups) / degree of freedom (between groups)

Mean Square (within group) = Sum of Squares (within group) / degree of freedom (within group)









Here, **p** = represents the number of groups

**n =**represents the number of observations in a group

https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/01/30062117/image_41.png=  represents the mean of a particular group

**X (bar)** = represents the mean of all the observations

Now, let us understand the degree of freedom for within group and between groups respectively.

Between groups : If there are k groups in ANOVA model, then k-1 will be independent. Hence, k-1 degree of freedom.

Within groups : If N represents the total observations in ANOVA (∑n over all groups) and k are the number of groups then, there will be k fixed points. Hence, N-k degree of freedom.

#### 8.1 Steps to perform ANOVA

1. Hypothesis Generation
   1. Null Hypothesis : Means of all the groups are same
   2. Alternate Hypothesis : Mean of at least one group is different
2. Calculate within group and between groups variability
3. Calculate F-Ratio
4. Calculate probability using F-table
5. Reject/fail to Reject Null Hypothesis

There are various other forms of ANOVA too like Two-way ANOVA, MANOVA, ANCOVA etc. but One-Way ANOVA suffices the requirements of this course.

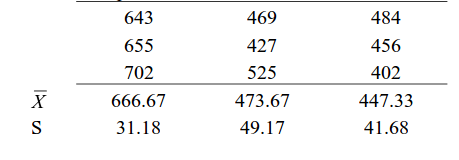
Practical applications of ANOVA in modeling are:

1. Identifying whether a categorical variable is relevant to a continuous variable.
2. Identifying whether a treatment was effective to the model or not.

#### 8.2 Practical Example

Suppose there are 3 chocolates in town and their sweetness is quantified by some metric (S). Data is collected on the three chocolates. You are given the task to identify whether the mean sweetness of the 3 chocolates are different. The data is given as below:

                                                                 Type A                    Type B                   Type C



Here, first we have calculated the sample mean and sample standard deviation for you.

Now we will proceed step-wise to calculate the F-Ratio (ANOVA statistic).

#### Step 1: Stating the Null and Alternate Hypothesis

**Null Hypothesis:**Mean sweetness of the three chocolates are same.

**Alternate Hypothesis:**Mean sweetness of at least one of the chocolates is different.

#### Step 2: Calculating the appropriate ANOVA statistic

In this part, we will be calculating SS(B), SS(W), SS(T) and then move on to calculate MS(B) and MS(W). The thing to note is that,

Total Sum of Squares [SS(t)] = Between Sum of Squares [SS(B)] + Within Sum of Squares [SS(W)].

So, we need to calculate any two of the three parameters using the data table and formulas given above.

As, per the formula above, we need one more statistic i.e Grand Mean denoted by X(bar) in the formula above.

**X bar** = (643+655+702+469+427+525+484+456+402)/9 = 529.22

**SS(B)**=[3\*(666.67-529.22)^2]+ [3\*(473.67-529.22)^2]+[3\*(447.33-529.22)^2] = 86049.55

**SS (W)** = [(643-666.67)^2+(655-666.67)^2+(702-666.67)^2] + [(469-473.67)^2+(427-473.67)^2+(525-473.67)^2] + [(484-447.33)^2+(456-447.33)^2+(402-447.33)^2]= 10254

**MS(B)** = SS(B) / df(B) = 86049.55 / (3-1) = 43024.78

**MS(W)** = SS(W) / df(W) = 10254/(9-3) = 1709

**F-Ratio** = MS(B) / MS(W) = 25.17 .

Now, for a 95 % confidence level, F-critical to reject Null Hypothesis for degrees of freedom(2,6) is 5.14 but we have 25.17 as our F-Ratio.

So, we can confidently reject the Null Hypothesis and come to a conclusion that at least one of the chocolate has a mean sweetness different from the others.

You can use the F-calculator [here](http://stattrek.com/online-calculator/f-distribution.aspx).

**Note:** ANOVA only lets us know the means for different groups are same or not. It doesn’t help us identify which mean is different.To know which group mean is different, we can use another test know as Least Significant Difference Test.

## 9. Chi-square Goodness of Fit Test

Sometimes, the variable under study is not a continuous variable but a categorical variable. Chi-square test is used when we have one single categorical variable from the population.

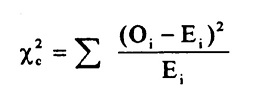
Let us understand this with help of an example. Suppose a company that manufactures chocolates, states that they manufacture 30% dairy milk, 60% temptation and 10% kit-kat. Now suppose a random sample of 100 chocolates has 50 dairy milk, 45 temptation and 5 kitkats. Does this support the claim made by the company?

Let us state our Hypothesis first.

Null Hypothesis: The claims are True

Alternate Hypothesis: The claims are False.

Chi-Square Test is given by:



where, https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/01/30062343/Image_5.png= sample or observed values

https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/01/30062444/Image_61.png= population values

The summation is taken over all the levels of a categorical variable.

https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/01/30062444/Image_61.png= **[n \* https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/01/30062619/image_7.png]**  Expected value of a level (i) is equal to the product of sample size and percentage of it in the population.

Let us now calculate the Expected values of all the levels.

E (dairy milk)= 100 \* 30% = 30

E (temptation) = 100 \* 60% =60

E (kitkat) = 100 \* 10% = 10

Calculating chi-square = [(50-30)^2/30+(45-60)^2/60+(5-10)^2/10] =19.58

Now, checking for p (chi-square >19.58) using [chi-square calculator](http://stattrek.com/online-calculator/chi-square.aspx), we get p=0.0001. This is significantly lower than the alpha(0.05).

So we reject the Null Hypothesis.

## 10. Regression and ANOVA

If you have studied some basic Machine Learning Algorithms, the first algorithm that you must have studied is Regression. If we  recall those lessons of Regression, what we generally do is calculate the weights for features present in the model to better predict the output variable. But finding the right set of feature weights or features for that matter is not always possible.

It is highly likely that that the existing features in the model are not fit for explaining the trend in dependent variable or the feature weights calculated fail at explaining the trend in dependent variable. What is important is knowing the degree to which our model is successful in explaining the trend (variance) in dependent variable.

Enter ANOVA.

With the help of ANOVA techniques, we can analyse a model performance very much like we analyse samples for being statistically different or not.

But with regression things are not easy. We do not have mean of any kind to compare  or sample as such but we can find good alternatives in our regression model which can substitute for mean and sample.

Sample in case of regression is a regression model itself with pre-defined features and feature weights whereas mean is replaced by variance(of both dependent and independent variables).

Through our ANOVA test we would like to know the amount of variance explained by the Independent variables in Dependent Variable VS the amount of variance that was left unexplained.

It is intuitive to see that larger the unexplained variance(trend) of the dependent variable smaller will be the ratio and less effective is our regression model. On the other hand, if we have a large explained variance then it is easy to see that our regression model was successful in explaining the variance in the dependent variable and more effective is our model. The ratio of Explained Variance uand Unexplained Variance is called F-Ratio.

Let us now define these explained and unexplained variances to find the effectiveness of our model.

**1. Regression (Explained) Sum of Squares** – It is defined as the amount of variation explained by the Regression model in the dependent variable.

Mathematically, it is calculated as:

where, https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/01/30063257/Image_8.png[hat] = predicted value and

**y(bar)** = mean of the actual y values.

**Interpreting Regression sum of squares –**

If our model is a good model for the problem at hand then it would produce an output which has distribution as same to the actual dependent variable. i.e it would be able to capture the inherent variation in the dependent variable.

**2. Residual Sum of Squares** – It is defined as the amount of variation independent variable which is not explained by the Regression model.

Mathematically, it is calculated as:

where, https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/01/30063257/Image_8.png = actual ‘y ‘ value

**f(x)** = predicted value

**Interpretation of Residual Sum of Squares –**

It can be interpreted as the amount by which the predicted values deviated from the actual values. Large deviation would indicate that the model failed at predicting the correct values for the dependent variable.

Let us now  work out F-ratio step by step. We will be making using of the Hypothesis Testing framework described above to test the significance of the model.

While calculating the F-Ratio care has to be taken to incorporate the effect of degree of freedom. Mathematically, F-Ratio is the ratio of **[Regression Sum of Squares/df(regression)] and [Residual Sum of Squares/df(residual)].**

We will be understanding the entire concept using an example and [this excel sheet](https://drive.google.com/file/d/0ByAvlBzuj2TgV0RDM0FORkQ3YW8/view).

#### Step 0: State the Null and Alternate Hypothesis

Null Hypothesis: The model is unable to explain the variance in the dependent variable (Y).

Alternate Hypothesis: The model is able to explain the variance in dependent variable (Y)

#### Step 1:

Calculate the regression equation for X and Y using Excel’s in-built tool.

#### Step 2:

Predict the values of y for each row of data.

#### Step 3:

Calculate y(mean) – mean of the actual y values which in this case turns out to be 0.4293548387.

#### Step 4:

Calculate the Regression Sum of Squares using the above-mentioned formula. It turned out to be 2.1103632473

The Degree of freedom for regression equation is 1, since we have only 1 independent variable.

#### Step 5:

Calculate the Residual Sum of Squares using the above-mentioned formula. It turned out to be 0.672210946.

Degree of Freedom for residual = Total degree of freedom – Degree of freedom(regression)

=(62-1) – 1 = 60

#### Step 6:

F-Ratio = (2.1103632473/1)/(0.672210946/60) = 188.366

Now, for 95% confidence, F-critical to reject Null Hypothesis for 1,60 degrees of freedom in 4. But we have F-ratio as 188, so we can safely reject the Null Hypothesis and conclude that model explains variation to a large extent.

## 11. Coefficient of Determination (R-Square)

It is defined as the ratio of the amount of variance explained by the regression model to the total variation in the data. It represents the strength of correlation between two variables.

We already calculated the Regression SS and Residual SS. Total SS is the sum of Regression SS and Residual SS.

Total SS = 2.1103632473+ 0.672210946 = 2.78257419

Co-efficient of Determination = 2.1103632473/2.78257419 = 0.7588

## 12. Correlation Coefficient

This is another useful statistic which is used to determine the correlation between two variables. It is simply the square root of coefficient of Determination and ranges from -1 to 1 where 0 represents no correlation and 1 represents positive strong correlation while -1 represents negative strong correlation.

This [cheat sheet on Statistics](http://web.mit.edu/~csvoss/Public/usabo/stats_handout.pdf) by MIT covers all the concepts of Statistics required for deep learning.