



Statistics

Statistics:

- Probability Basics
- Autocorrelation
- Coefficients
- Correlation
- Confusion Matrix
- Measuring Variability & Spread
- Distributions
- Statistical tests
- Skewness
- Kurtosis
- Central Limit Theorem
- Hypothesis Testing

Probability Basics:

Mode

The **mode** is the value that appears most often in a set of data.

The range is the difference between the lowest value and the highest value.

Range

Median

The **median** is the middle number in a list of numbers ordered from lowest to highest.

The mean is the total of all the values, divided by the number of values.

Mean

Predictor: Predictor variable is an independent variable used in regression analysis

Multicollinearity: In regression, "multicollinearity" refers to predictors that are correlated with other predictors.

Or

Colinearity or Multicollinearity is the state where two variables are highly correlated and contain similar information about the variance within a given dataset.

Multicollinearity occurs when your model includes multiple factors that are correlated not just to your response variable, but also to each other. In other words, it results when you have factors that are a bit redundant.

In Python this can be accomplished by using numpy's `corrcoef` function or

Why multicollinearity is a problem?

If my X_i 's are highly correlated then $|X'X|$ will be close to 0 and hence inverse of $(X'X)$ will not exist or will be indefinitely large. Mathematically, which will be indefinitely large in presence of multicollinearity. Long story in short, multicollinearity increases the estimate of standard error

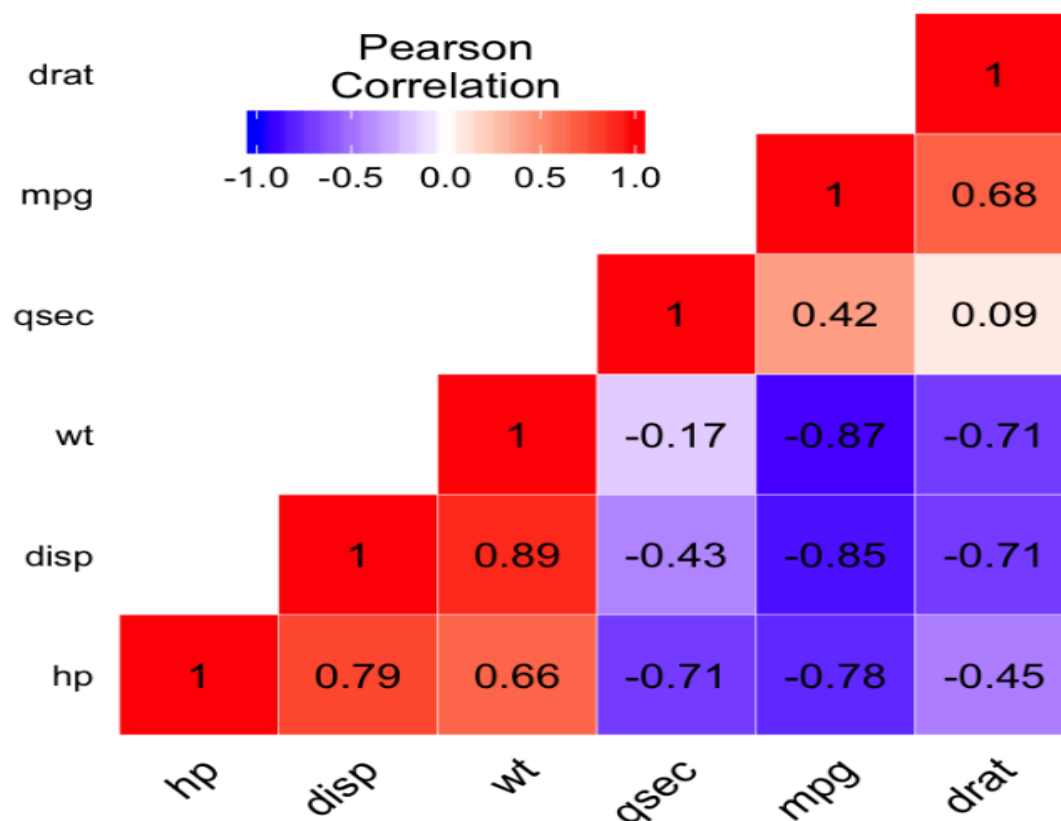
of regression coefficients which makes some variables statistically insignificant when they should be significant.

How can you detect multicollinearity?

There are 2 ways multicollinearity is usually checked

1. Correlation Matrix
2. Variance Inflation Factor (VIF)


Correlation Matrix: A correlation matrix is a table showing correlation coefficients between variables.



VIF (Variance Inflation Factor) Method:

Variance inflation factor (VIF) is the quotient of the variance in a model with multiple terms by the variance of a model with one term alone. It quantifies the severity of multicollinearity in an ordinary least squares regression analysis. VIF value can be interpreted as

- 1 (Non-collinear)
- 1–5 (Medium collinear)
- >5 (Highly collinear)



The values having VIF value above 5 are removed.

Firstly we fit a model with all the variables and then calculate the variance inflation factor (VIF) for each variable. VIF measures how much the variance of an estimated regression coefficient increases if your predictors are correlated.

What is Variance Inflation Factor?


Variance inflation factor (VIF) for an explanatory variable is given $1/(1-R^2)$. Here, we take that particular X as response variable and all other explanatory variables as independent variables. So, we run a regression between one of those explanatory variables with remaining explanatory variables.

```
statsmodels.stats.outliers_influence.variance_inflation_factor(exog, exog_idx)
```

Conclusion

Multicollinearity can significantly reduce the model's performance and we may not know it. It is a very important step during the feature selection process. Removing multicollinearity can also reduce features which will eventually result in a less complex model and also the overhead to store these features will be less.

Make sure to run the multicollinearity test before performing any regression analysis.



Autocorrelation: It is correlation between two successive observations of same variable.

Example: The outcome of current year production is dependent on previous year production (Cotton production over the years).

Checking for autocorrelation: To ensure the absence of autocorrelation we use Ljungbox test.

Null Hypothesis: Autocorrelation is absent.

Alternative Hypothesis: Autocorrelation is present.

```
from statsmodels.stats import diagnostic as diag
diag.acorr_ljungbox(lm2.resid , lags = 1)
(array([ 1.97177212]), array([ 0.16025989]))
```

Since p-value is 0.1602 thus we can accept the null hypothesis and can say that autocorrelation is absent.

Covariance: “Covariance” indicates the direction of the linear relationship between variables. “Correlation” on the other hand measures both the strength and direction of the linear relationship between two variables.

For more info: <https://towardsdatascience.com/let-us-understand-the-correlation-matrix-and-covariance-matrix-d42e6b643c22>

Key Differences between Covariance and Correlation

The following points are noteworthy so far as the difference between covariance and correlation is concerned:

1. A measure used to indicate the extent to which two random variables change in tandem is known as covariance. A measure used to represent how strongly two random variables are related known as correlation.
2. Covariance is nothing but a measure of correlation. On the contrary, correlation refers to the scaled form of covariance.
3. The value of correlation takes place between -1 and +1. Conversely, the value of covariance lies between $-\infty$ and $+\infty$.
4. Covariance is affected by the change in scale, i.e. if all the value of one variable is multiplied by a constant and all the value of another variable are multiplied, by a similar or different constant, then the covariance is changed. As against this, correlation is not influenced by the change in scale.
5. Correlation is dimensionless, i.e. it is a unit-free measure of the relationship between variables. Unlike covariance, where the value is obtained by the product of the units of the two variables.



Coefficients:

Regression coefficients are estimates of the unknown population parameters and describe the relationship between a predictor variable and the response. In linear regression, coefficients are the values that multiply the predictor values. Suppose you have the following regression equation: $y = 3X + 5$. In this equation, +3 is the coefficient, X is the predictor, and +5 is the constant.

The sign of each coefficient indicates the direction of the relationship between a predictor variable and the response variable.

- A positive sign indicates that as the predictor variable increases, the response variable also increases.
- A negative sign indicates that as the predictor variable increases, the response variable decreases.

The coefficient value represents the mean change in the response given a one unit change in the predictor. For example, if a coefficient is +3, the mean response value increases by 3 for every one unit change in the predictor.

Heteroscedasticity:

Heteroscedasticity is a hard word to pronounce, but it doesn't need to be a difficult concept to understand. Put simply, heteroscedasticity (also spelled heteroskedasticity) refers to the circumstance in which the variability of a variable is unequal across the range of values of a second variable that predicts it.

A scatterplot of these variables will often create a cone-like shape, as the scatter (or variability) of the dependent variable (DV) widens or narrows as the value of the independent variable (IV) increases. The inverse of heteroscedasticity is homoscedasticity, which indicates that a DV's variability is equal across values of an IV.

Detecting Heteroscedasticity

Graphical Method: Firstly do the regression analysis and then plot the error terms or residuals against the predicted values (\hat{Y}_i). If there is a definite pattern (like linear or quadratic or funnel shaped) obtained from the scatter plot then heteroscedasticity is present.

Python's Statsmodels includes three of these tests: the **Quandt-Goldfeld** (no longer used), and the **Breusch-Pagan** and **White** tests.

Step1: Import required libraries:

```

from statsmodels.stats.diagnostic import het_breuschpagan
from statsmodels.stats.diagnostic import het_white

import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols

```

Step2: Load your data and run your regression

Step3: In the case of the Breusch-Pagan test, pass it the residuals, and either all of the x variables, or any subset of them that you would like to test specifically.

```

bp_test = het_breuschpagan(statecrime_model.resid,
[statecrime_df.var1, statecrime_df.var2...])

```

Step4: Zip the output array with a list of the following labels:

Step5: Add into Dict and print

```

labels = ['LM Statistic', 'LM-Test p-value', 'F-Statistic', 'F-Test
p-value']

print(dict(zip(labels, bp_test))
print(dict(zip(labels, white_test))

```

What you get are two test statistics and two p-values (textbooks prefer the LM test, but the F test is widely used and basically equivalent). Heteroskedasticity is indicated if $p < 0.05$, so according to these tests, this model is heteroskedastic.

```

{'LM Statistic': 28.005305565415412,
'LM-Test p-value': 0.014204867570584721,
'F-Statistic': 3.131750373433381,
'F-Test p-value': 0.0028924508752700495}

```

White Test Results

```

{'LM Statistic': 11.943695347031433,
'LM-Test p-value': 0.01777490950551206,
'F-Statistic': 3.5167816748485357,
'F-Test p-value': 0.013805007084298618}

```

Breusch-Pagan Test Results

How to rectify?

- Re-build the model with new predictors.
- Variable transformation such as Box-Cox transformation.

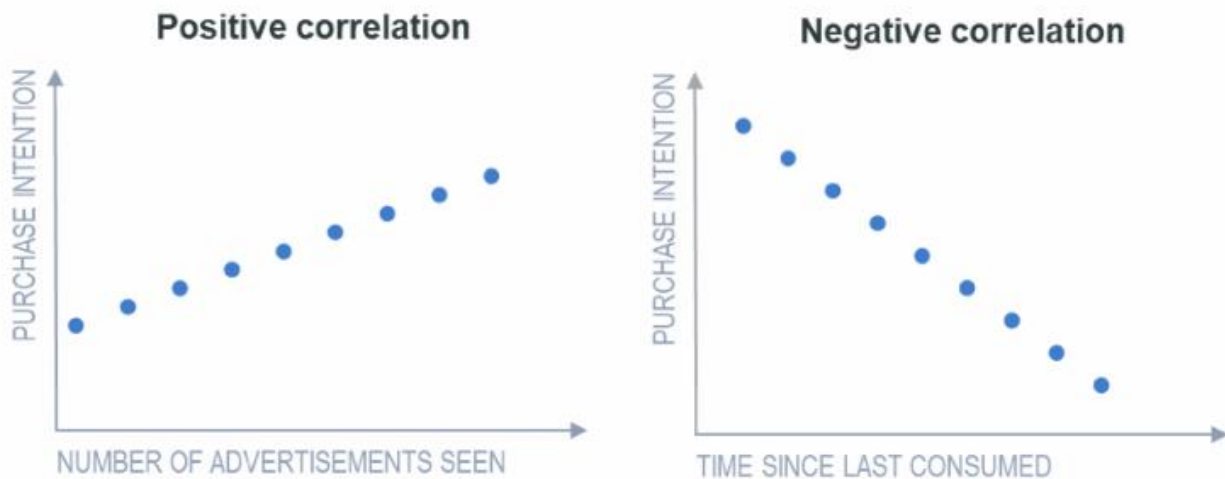
Correlation

Correlation means two variables vary together, if one changes so does the other.

Or

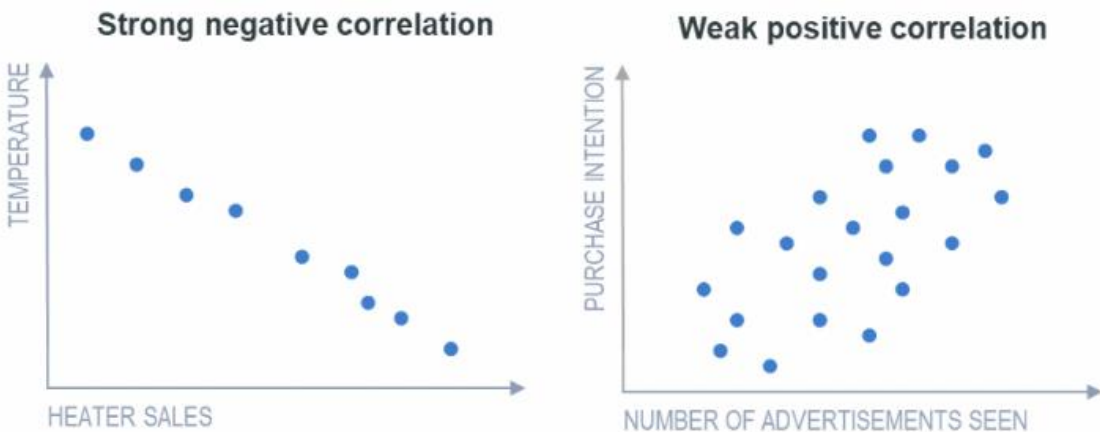
Correlation simply means that there is some type of relationship between two variables. This post will define positive and negative correlation, provide some examples of correlation, explain how to measure correlation and discuss some pitfalls regarding correlation.

When the values of one variable increase as the values of the other increase, this is known as **positive correlation**. When the values of one variable decrease as the values of another increase to form an inverse relationship, this is known as **negative correlation**.



An example of positive correlation may be that the more you exercise, the more calories you will burn.

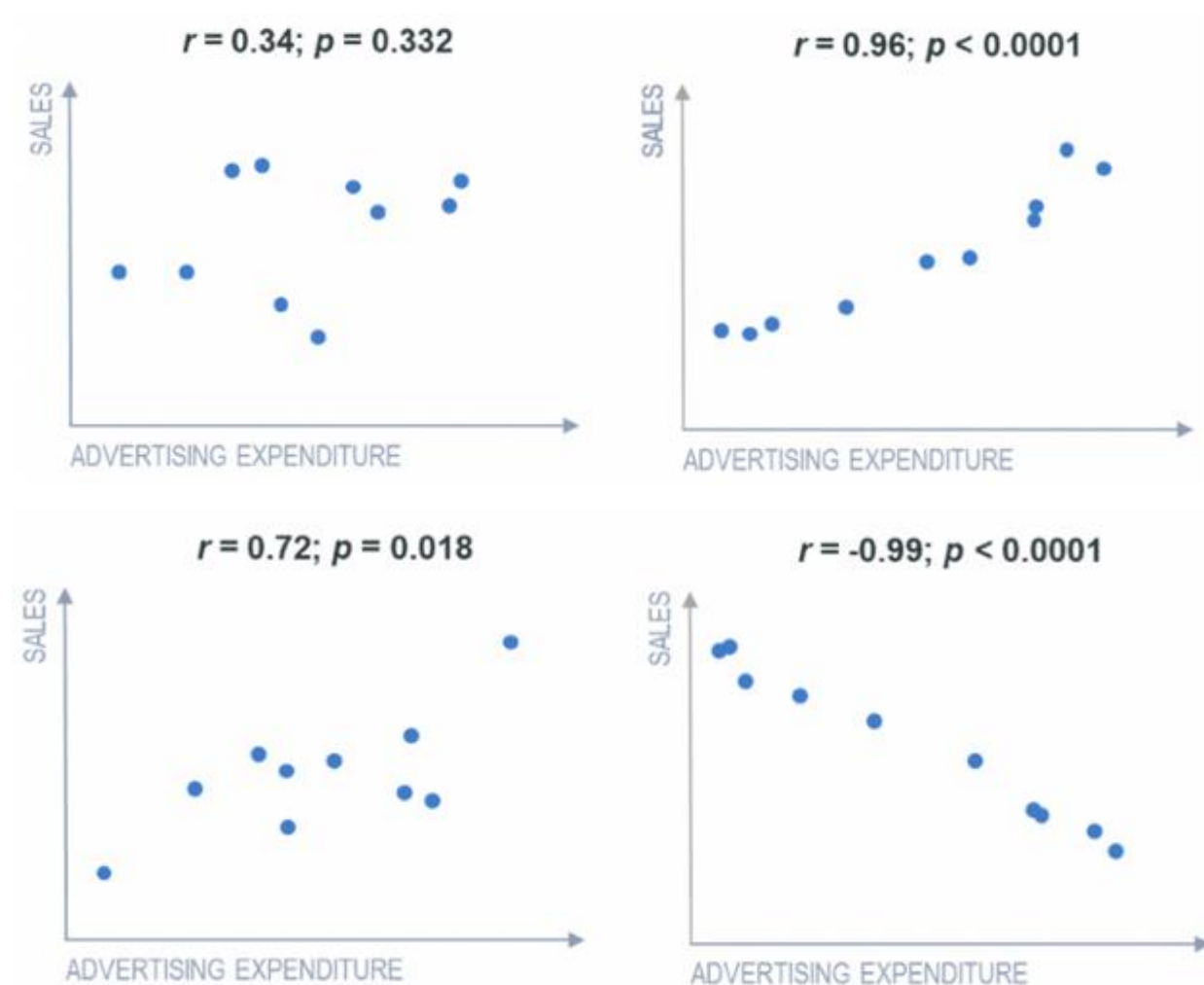
Where it is possible to predict, with a reasonably high level of accuracy, the values of one variable based on the values of the other, the relationship between the two variables is described as a strong correlation. A weak correlation is one where on average the values of one variable are related to the other, but there are many exceptions.



Pearson's Product-Moment Correlation

The most common measure of correlation is Pearson's product-moment correlation, which is commonly referred to simply as the *correlation*, the *correlation coefficient*, or just the letter *r* (always written in italics):

- A correlation of 1 indicates a perfect positive correlation.
- A correlation of -1 indicates a perfect negative correlation.
- A correlation of 0 indicates that there is no relationship between the different variables.
- Values between -1 and 1 denote the strength of the correlation, as shown in the example below.

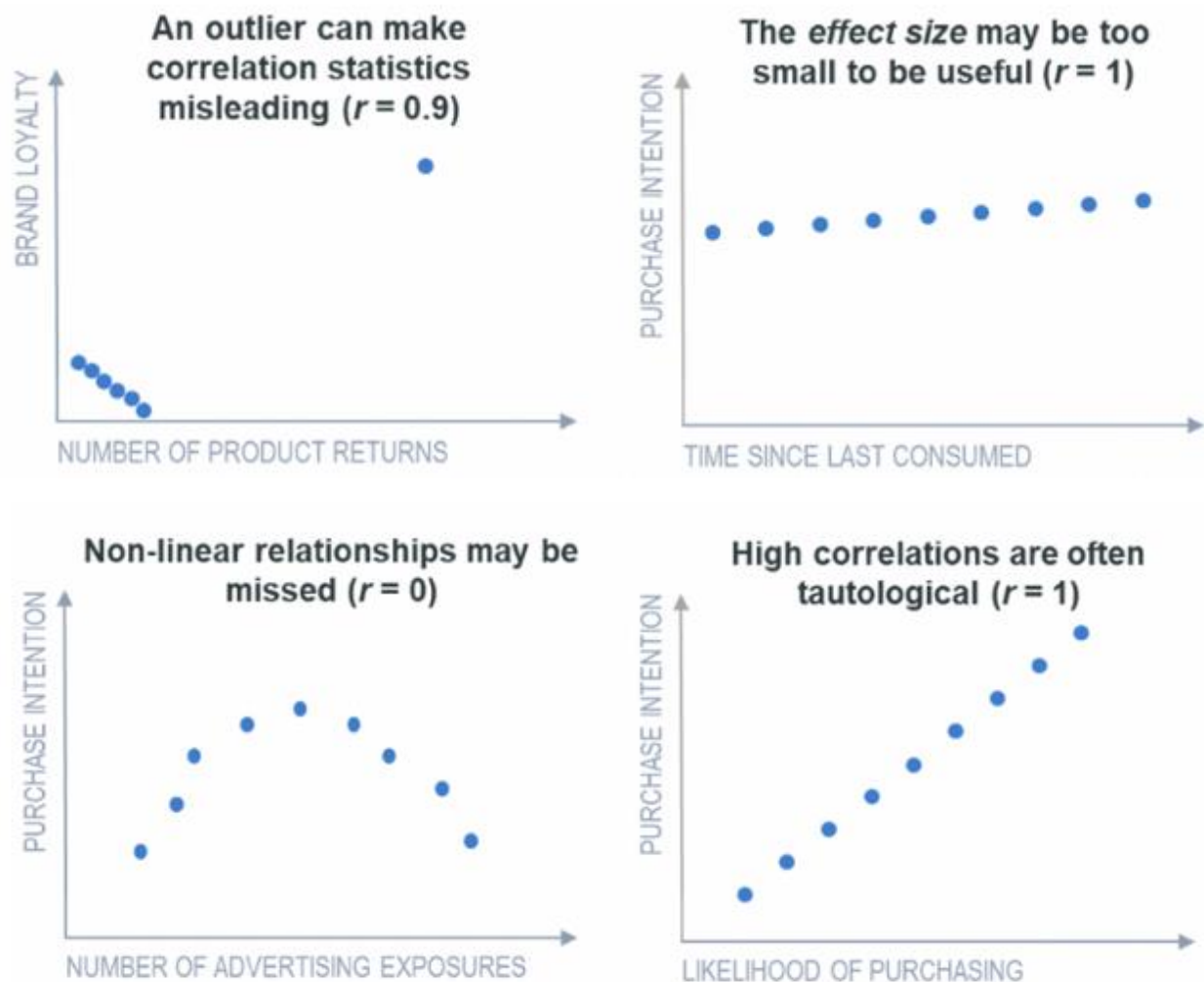


Misinterpreting correlations:

Just about all the common problems that can render statistical analysis meaningless can occur with correlations.

One example of a common problem is that with small samples, correlations can be unreliable. The smaller the sample size, the more likely we are to observe a correlation that is further from 0, even if the true correlation (obtained if we had data for the entire population) was 0. The standard way of quantifying this is to use p-values. In academic research, a common rule of thumb is that when p is greater than 0.05, the correlation should not be trusted.

Another problem, illustrated in the top-left chart below, is that a single unusual observation (outlier) can make the computed correlation coefficient highly misleading. Yet another problem is that correlations show only the extent to which one variable can be predicted by another, and they do not pick up situations where the difference in the predictive values is too small to be considered useful (to use the jargon, situations where the effect size is small), as shown in the top-right chart below.



Yet another problem with *correlation* is that it summarizes the linear relationship, and if the true relationship is nonlinear, then this may be missed. One more problem is that very high correlations often reflect tautologies rather than findings of interest.

Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

		Actual	
		True	False
Predicted	Positive	True Positive HIT	False Positive False Alarm (Type 1 Error)
	Negative	True Negative Correct Rejection	False Negative Miss (Type 2 Error)

True Positive: You predicted positive and it's true.

You predicted that a woman is pregnant and she actually is.

True Negative: You predicted negative and it's true.

You predicted that a man is not pregnant and he actually is not.

False Positive (Type 1 Error): You predicted positive and it's false.

You predicted that a man is pregnant but he actually is not.

False Negative (Type 2 Error): You predicted negative and it's false.

You predicted that a woman is not pregnant but she actually is.

Just Remember, We describe predicted values as Positive and Negative and actual values as true and false.


Recall: Out of all the positive classes, how much we predicted correctly. It should be high as possible.

Or

It attempts to answer, what proportion of actual positives was identified correctly?

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

Precision: It attempts to answer, what proportion of positive identifications was actually correct?



$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

Accuracy: it is the measure of all the correctly identified cases. It is most used when all the classes are equally important.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{(\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative})}$$

F1 Score: F1 Score is needed when you want to seek a balance between Precision and Recall using the harmonic mean as it punishes the extreme values more.

The F-Measure will always be nearer to the smaller value of Precision or Recall.

$$F1 = 2 \times \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

To summarize the differences between the F1-score and the accuracy,

- Accuracy is used when the True Positives and True negatives are more important while F1-score is used when the False Negatives and False Positives are crucial
- Accuracy can be used when the class distribution is similar while F1-score is a better metric when there are imbalanced classes as in the above case.
- In most real-life classification problems, imbalanced class distribution exists and thus F1-score is a better metric to evaluate our model on.

High recall, low precision: This means that most of the positive examples are correctly recognized (low FN) but there are a lot of false positives.

Low recall, high precision: This shows that we miss a lot of positive examples (high FN) but those we predict as positive are indeed positive (low FP)

Visualizing Recall and Precision:

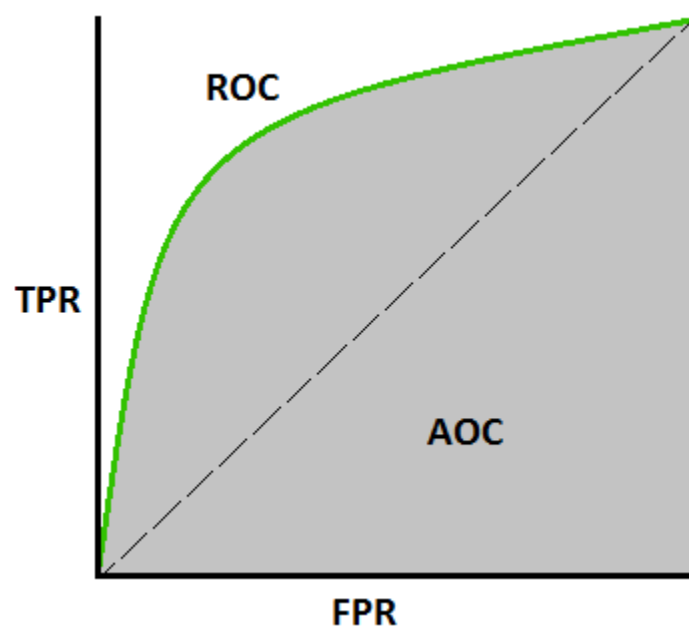
- **Confusion matrix:** shows the actual and predicted labels from a classification problem
- **Receiver operating characteristic (ROC) curve:** plots the true positive rate (TPR) versus the false positive rate (FPR) as a function of the model's threshold for classifying a positive
- **Area under the curve (AUC):** metric to calculate the overall performance of a classification model based on area under the ROC curve

When we need to check or visualize the performance of the multi - class classification problem

What is AUC - ROC Curve?

AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing between patients with disease and no disease.

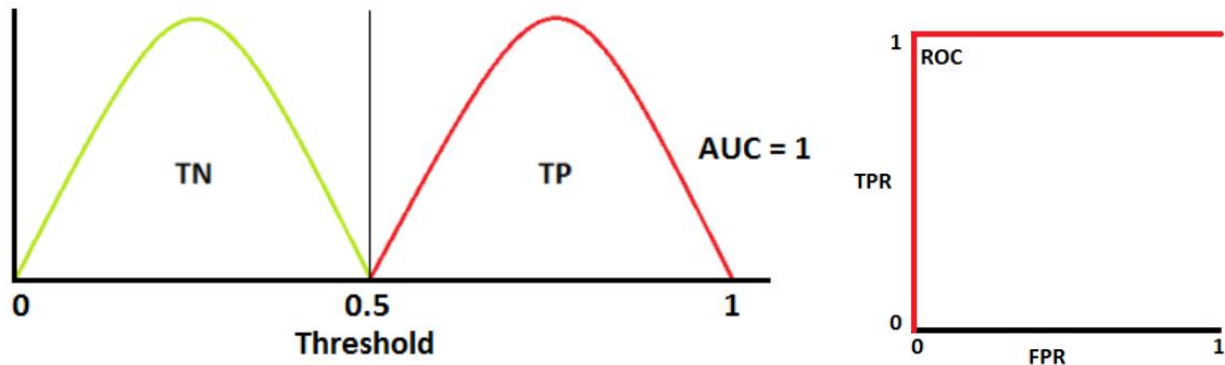
The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis.



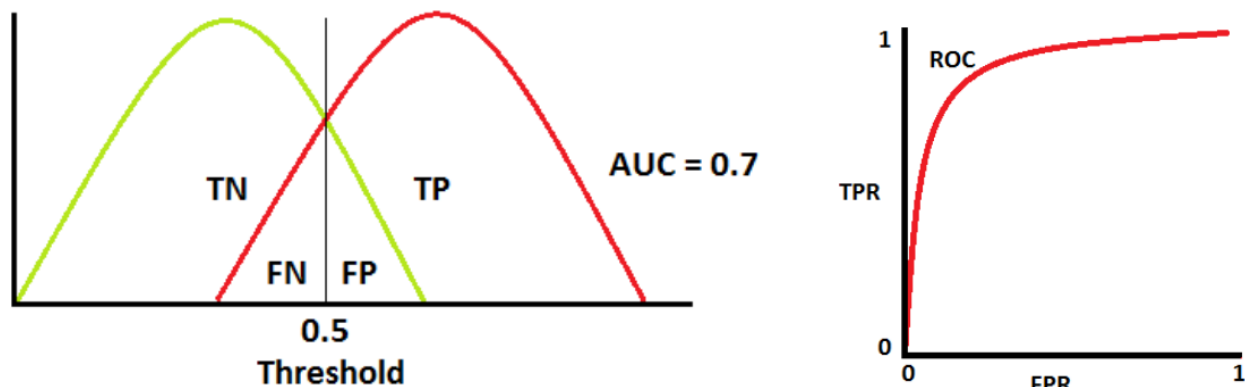
How to speculate the performance of the model?

An excellent model has AUC near to the 1 which means it has good measure of separability. A poor model has AUC near to the 0 which means it has worst measure of separability. In fact it means it is reciprocating the result. It is predicting 0s as 1s and 1s as 0s. And when AUC is 0.5, it means model has no class separation capacity whatsoever.

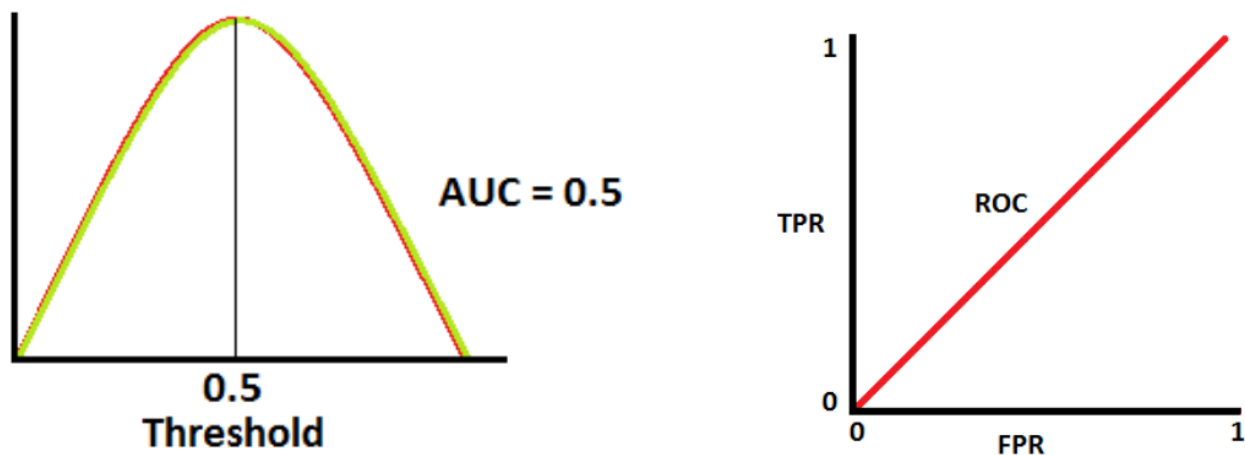
1. This is an ideal situation. When two curves don't overlap at all means model has an ideal measure of separability. It is perfectly able to distinguish between positive class and negative class.



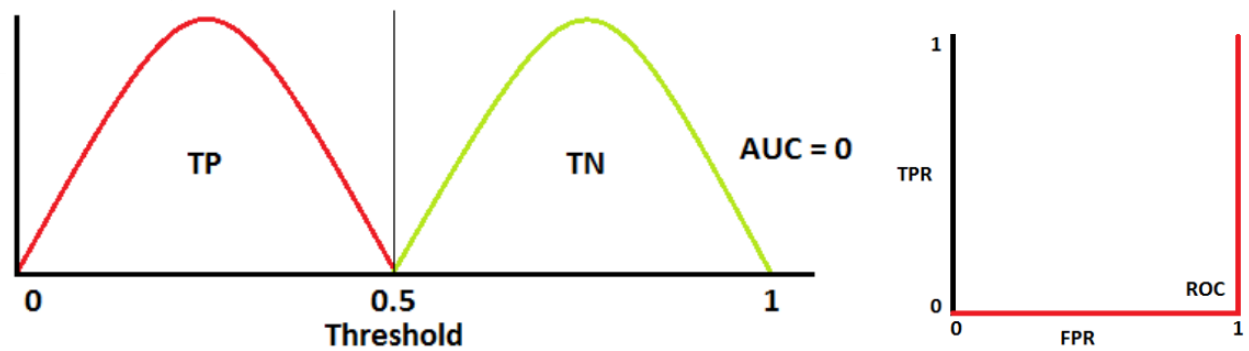
2. When two distributions overlap, we introduce type 1 and type 2 error. Depending upon the threshold, we can minimize or maximize them. When AUC is 0.7, it means there is 70% chance that model will be able to distinguish between positive class and negative class.



3. This is the worst situation. When AUC is approximately 0.5, model has no discrimination capacity to distinguish between positive class and negative class.



4. When AUC is approximately 0, model is actually reciprocating the classes. It means, model is predicting negative class as a positive class and vice versa.



Defining terms used in AUC and ROC Curve.

1. TPR (True Positive Rate) / Recall / Sensitivity:

$$\text{TPR / Recall / Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

2. Specificity:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

3. FPR:

$$\begin{aligned} \text{FPR} &= 1 - \text{Specificity} \\ &= \frac{\text{FP}}{\text{TN} + \text{FP}} \end{aligned}$$

More on ROC and AUC curve:

<https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>

<https://towardsdatascience.com/understanding-the-roc-and-auc-curves-a05b68550b69>

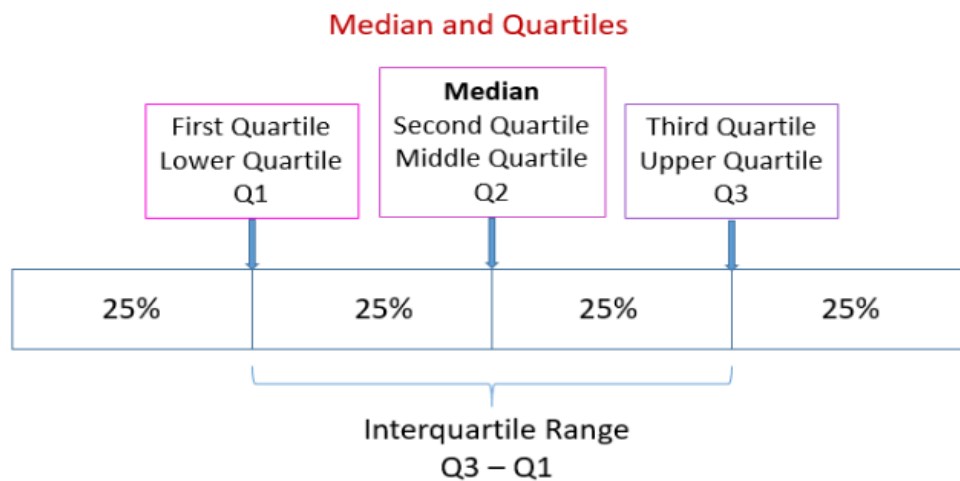
Measuring variability and spread

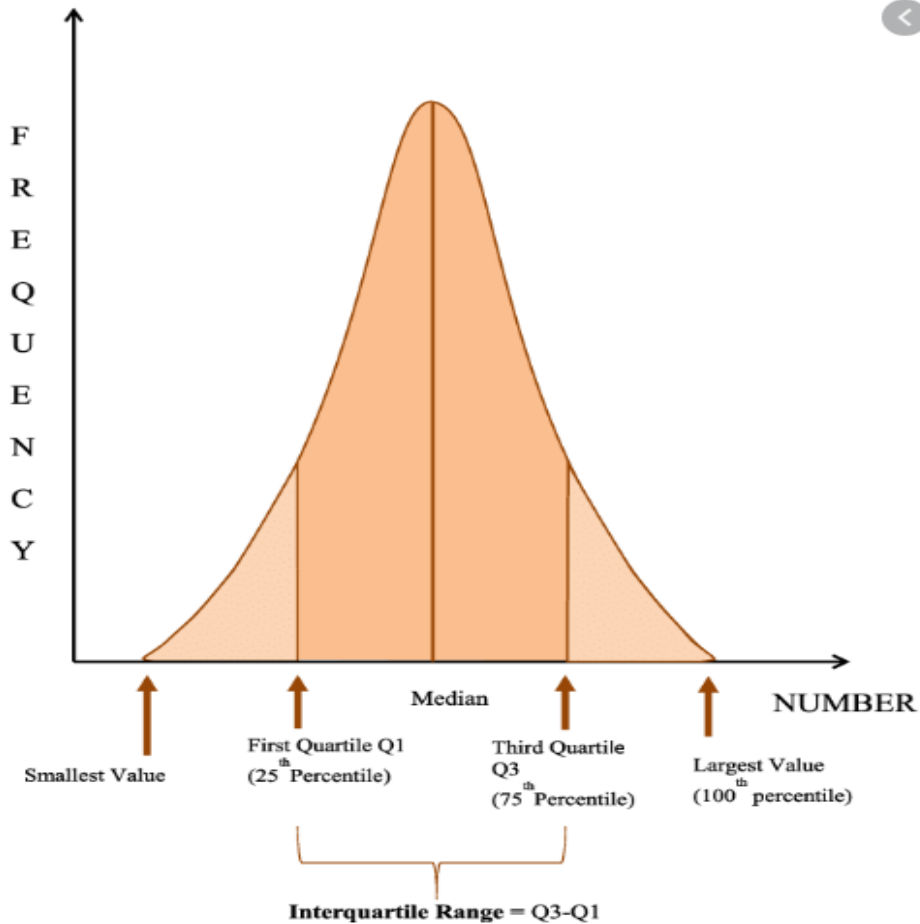
- Min and Max
- Quartiles
- Interquartile Range IQR
- Box plot or Whisker plot
- Variance
- Standard Deviation
- z-score

Quartiles:

A quartile divides a sorted data set into 4 equal parts, so that each part represents 1/4 of the data set

1. The lowest 25% of numbers.
2. The next lowest 25% of numbers (up to the median).
3. The second highest 25% of numbers (above the median).
4. The highest 25% of numbers.





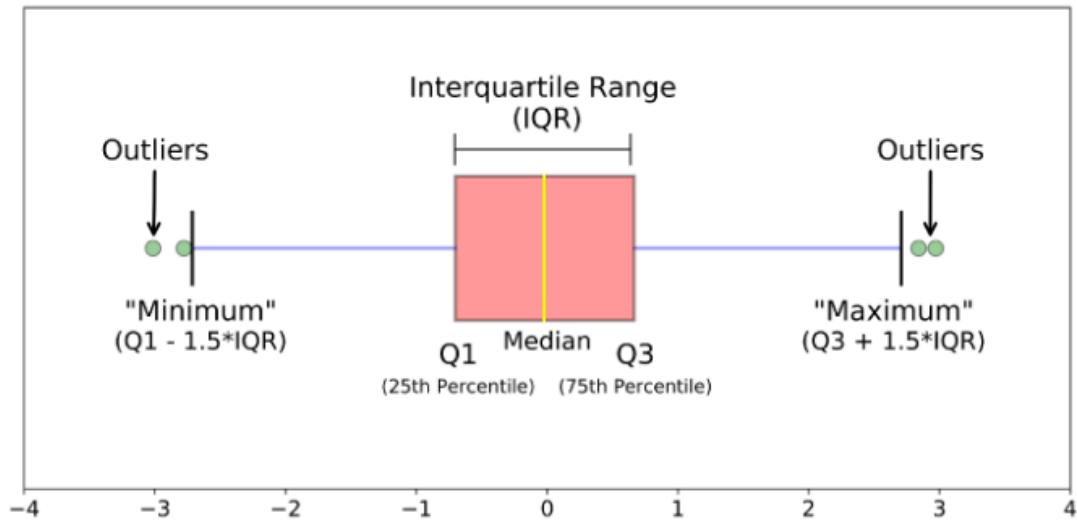
Interquartile Range IQR:

Interquartile range is defined as the difference between the upper and lower quartile values in a set of data

$$\text{Interquartile Range} = Q_3 - Q_1$$

Box plot or Whisker plot:

It displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum.



Variance:

In statistics, variance refers to the spread of a data set. It's a measurement used to identify how far each number in the data set is from the mean.

While performing market research, variance is particularly useful when calculating probabilities of future events. Variance is a great way to find all of the possible values and likelihoods that a random variable can take within a given range.

A variance value of zero represents that all of the values within a data set are identical, while all variances that are not equal to zero will come in the form of positive numbers.


The larger the variance, the more spread in the data set.

A large variance means that the numbers in a set are far from the mean and each other. A small variance means that the numbers are closer together in value.

How to Calculate Variance: Variance is calculated by taking the differences between each number in a data set and the mean, squaring those differences to give them positive value, and dividing the sum of the resulting squares by the number of values in the set.

The formula for variance is as follows:

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$



In this formula, X represents an individual data point, \bar{x} represents the mean of the data points, and N represents the total number of data points.

Advantage of Variance

One of the primary advantages of variance is that it treats all deviations from the mean of the data set in the same way, regardless of direction.

This ensures that the squared deviations cannot sum to zero, which would result in giving the appearance that there was no variability in the data set at all.

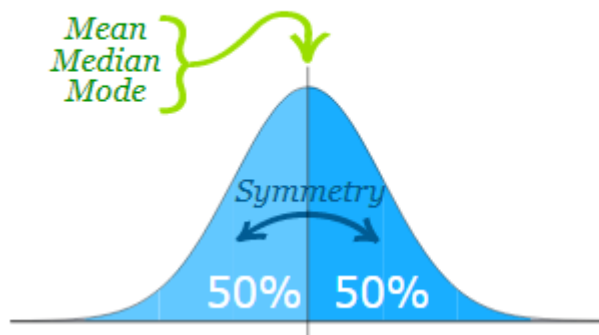
Disadvantage of Variance

One of the most commonly discussed disadvantages of variance is that it gives added weight to numbers that are far from the mean, or outliers. Squaring these numbers can at times result in skewed interpretations of the data set as a whole.

Distribution:

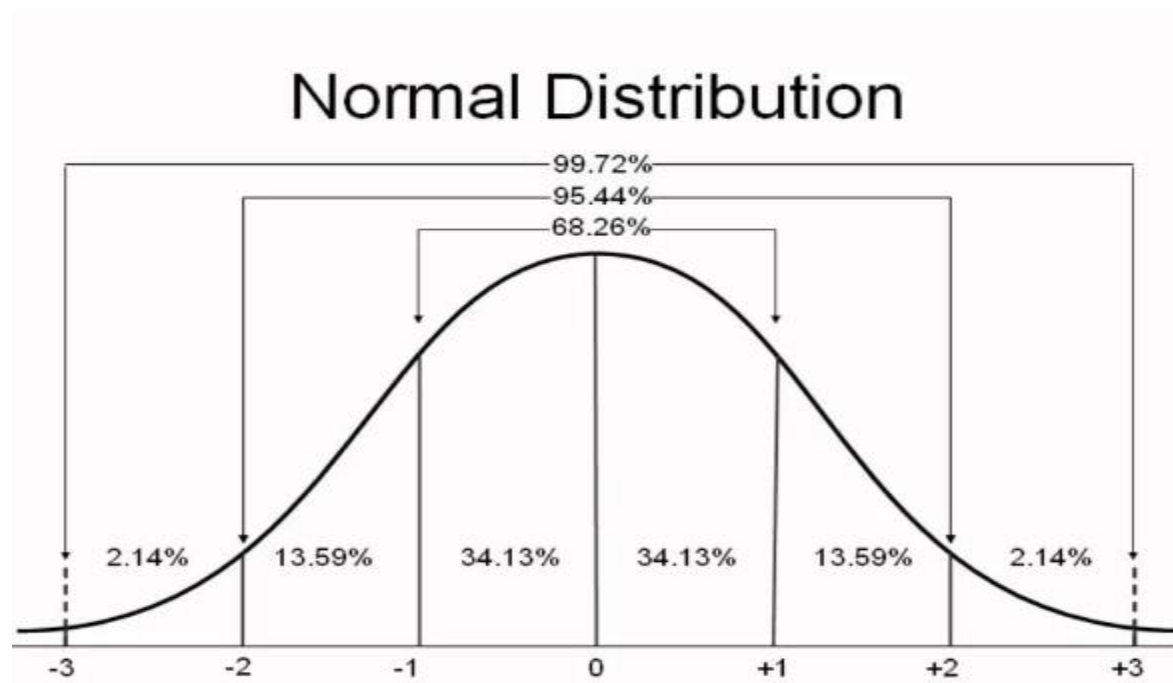
Standard Normal Distribution (Z Distribution): A normal distribution is also known as the Gaussian distribution and it is an arrangement of data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.

We say the data is "normally distributed":



The **Normal Distribution** has:

- mean = median = mode
- symmetry about the center
- 50% of values less than the mean and 50% greater than the mean

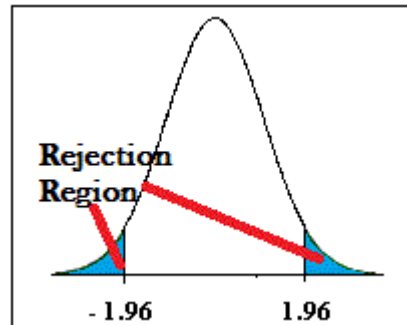


Student's T Distribution (T-Distribution):

The T distribution (also called Student's T Distribution) is a family of distributions that look almost identical to the normal distribution curve, only a bit shorter and fatter. The t distribution is used instead of the normal distribution when you have small samples (for more on this, see: t-score vs. z-score). The larger the sample size, the more the t distribution looks like the normal distribution. In fact, for sample sizes larger than 20 (e.g. more degrees of freedom), the distribution is almost exactly like the normal distribution.

Uses

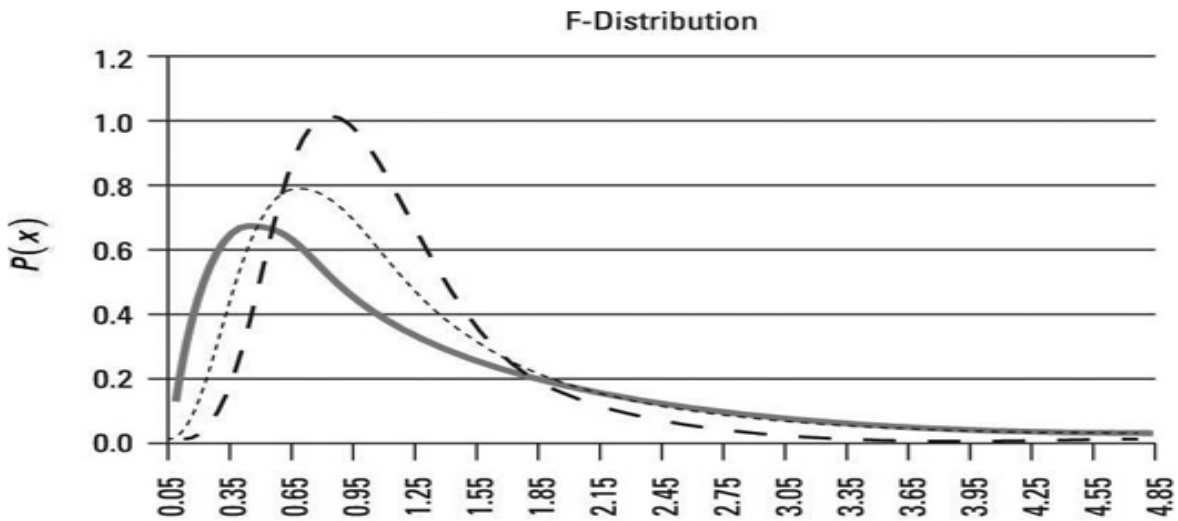
The T Distribution (and the associated t scores), are used in hypothesis testing when you want to figure out if you should accept or reject the null hypothesis.



F Distribution:

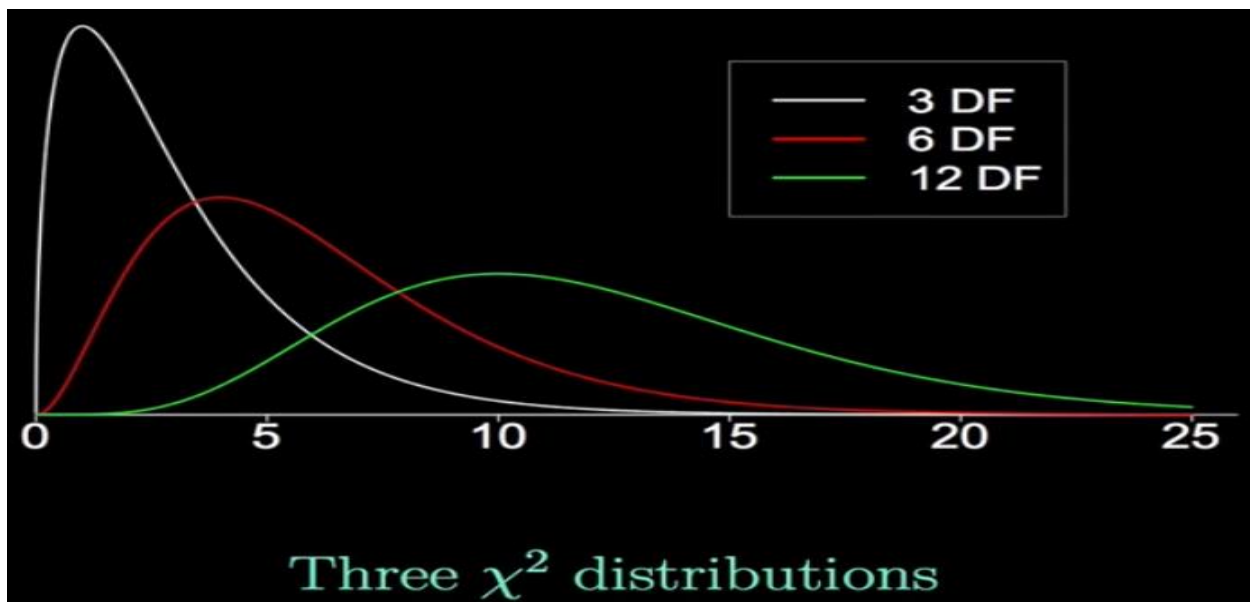
The F-distribution is a *continuous* probability distribution, which means that it is defined for an infinite number of different values. The F-distribution can be used for several types of applications, including testing hypotheses about the equality of two population variances and testing the validity of a multiple regression equation.

A good example of a positively skewed distribution is household incomes. Suppose that half of the households in a country have incomes below \$50,000 and half have incomes above \$50,000; this indicates that the median household income is \$50,000. Among households with incomes below \$50,000, the smallest possible value is \$0. Among households with incomes above \$50,000, there may be incomes of several million dollars per year. This imbalance between incomes below the median and above the median causes the mean to be substantially higher than the median. Suppose for example that the mean income in this case is \$120,000. This shows that the distribution of household incomes is positively skewed.



Chi Square Distribution:

A chi square distribution is a continuous distribution with degrees of freedom. Another best part of chi square distribution is to describe the distribution of a sum of squared random variables. It is also used to test the goodness of fit of a distribution of data, whether data series are independent, and for estimating confidences surrounding variance and standard deviation for a random variable from a normal distribution.





Statistical Tests:

This is an attempt to mark out the difference between the most common tests, the use of null value hypothesis in these tests and outlining the conditions under which a particular test should be used.

Null Hypothesis and Testing

Before we venture on the difference between different tests, we need to formulate a clear understanding of what a null hypothesis is. A null hypothesis, proposes that no significant difference exists in a set of given observations. For the purpose of these tests in general

- **Null:** Given two sample means are equal
- **Alternate:** Given two sample means are not equal

For rejecting a null hypothesis, a test statistic is calculated. This test-statistic is then compared with a critical value and if it is found to be greater than the critical value the hypothesis is rejected. "In the theoretical underpinnings, hypothesis tests are based on the notion of critical regions: the null hypothesis is rejected if the test statistic falls in the critical region. The critical values are the boundaries of the critical region. If the test is one-sided (like a χ^2 test or a one-sided t-test) then there will be just one critical value, but in other cases (like a two-sided t-test) there will be two"

Critical Value:

A critical value is a point (or points) on the scale of the test statistic beyond which we reject the null hypothesis, and, is derived from the level of significance α of the test. Critical value can tell us, what is the probability of two sample means belonging to the same distribution. Higher, the critical value means lower the probability of two samples belonging to same distribution. The general critical value for a two-tailed test is 1.96, which is based on the fact that 95% of the area of a normal distribution is within 1.96 standard deviations of the mean.


Critical values can be used to do hypothesis testing in following way

- Calculate test statistic
- Calculate critical values based on significance level alpha
- Compare test statistic with critical values.

If the test statistic is lower than the critical value, accept the hypothesis or else reject the hypothesis.

Relationship between p-value, critical value and test statistic

As we know critical value is a point beyond which we reject the null hypothesis. P-value on the other hand is defined as the probability to the right of respective statistic (Z, T or chi). The benefit of using p-value is that it calculates a probability estimate, we can test at any desired level of significance by comparing this probability directly with the significance level.



For e.g., assume Z-value for a particular experiment comes out to be 1.67 which is greater than the critical value at 5% which is 1.64. Now to check for a different significance level of 1% a new critical value is to be calculated.

However, if we calculate p-value for 1.67 it comes to be 0.047. We can use this p-value to reject the hypothesis at 5% significance level since $0.047 < 0.05$. But with a more stringent significance level of 1% the hypothesis will be accepted since $0.047 > 0.01$. Important point to note here is that there is no double calculation required.

Z-test

In a z-test, the sample is assumed to be normally distributed. A z-score is calculated with population parameters such as “population mean” and “population standard deviation” and is used to validate a hypothesis that the sample drawn belongs to the same population.

- **Null:** Sample mean is same as the population mean
- **Alternate:** Sample mean is not same as the population mean

The statistics used for this hypothesis testing is called z-statistic, the score for which is calculated as

$z = (x - \mu) / (\sigma / \sqrt{n})$, where

x = sample mean

μ = population mean

σ / \sqrt{n} = population standard deviation

If the test statistic is lower than the critical value, accept the hypothesis or else reject the hypothesis

T-test

A t-test is used to compare the mean of two given samples. Like a z-test, a t-test also assumes a normal distribution of the sample. A t-test is used when the population parameters (mean and standard deviation) are not known.

There are three versions of t-test

- Independent samples t-test which compares mean for two groups
- Paired sample t-test which compares means from the same group at different times
- One sample t-test which tests the mean of a single group against a known mean.

The statistic for this hypothesis testing is called t-statistic, the score for which is calculated as

$t = (x_1 - x_2) / (\sigma / \sqrt{n_1} + \sigma / \sqrt{n_2})$, where

x_1 = mean of sample 1

x_2 = mean of sample 2

n_1 = size of sample 1

n_2 = size of sample 2

There are multiple variations of t-test which are explained in detail here



ANOVA

ANOVA, also known as analysis of variance, is used to compare multiple (three or more) samples with a single test. There are 2 major flavors of ANOVA

- **One-way ANOVA:** When we compare more than two groups, based on one factor (independent variable), this is called one way ANOVA. For example, it is used if a manufacturing company wants to compare the productivity of three or more employees based on working hours. This is called one way ANOVA.
- **Two-way ANOVA:** When a company wants to compare the employee productivity based on two factors (2 independent variables), then it said to be two way (Factorial) ANOVA. For example, based on the working hours and working conditions, if a company wants to compare employee productivity, it can do that through two way ANOVA. Two-way ANOVA can be used to see the effect of one of the factors after controlling for the other, or it can be used to see the interaction between the two factors. This is a great way to control for extraneous variables as you are able to add them to the design of the study.
- **MANOVA:** MANOVA allows us to test the effect of one or more independent variable on two or more dependent variables. In addition, MANOVA can also detect the difference in co-relation between dependent variables given the groups of independent variables.

When the factor comparison is taken, then it said to be n-way ANOVA. For example, in productivity measurement if a company takes all the factors for productivity measurement, then it is said to be n-way ANOVA

More on Anova: <https://www.linkedin.com/pulse/working-example-analysis-varianceanova-r-prerna-sahay/>

The hypothesis being tested in ANOVA is

- **Null:** All pairs of samples are same i.e. all sample means are equal
- **Alternate:** At least one pair of samples is significantly different

The statistics used to measure the significance, in this case, is called F-statistics. The F value is calculated using the formula

$F = ((SSE1 - SSE2)/m) / SSE2/n-k$, where

SSE = residual sum of squares

m = number of restrictions

k = number of independent variables

There are multiple tools available such as SPSS, R packages, Excel etc. to carry out ANOVA on a given sample.

Chi-Square Test

Chi-square test is used to compare categorical variables. There are two type of chi-square test

- Goodness of fit test, which determines if a sample matches the population.
- A chi-square fit test for two independent variables is used to compare two variables in a contingency table to check if the data fits.
- A small chi-square value means that data fits
- A high chi-square value means that data doesn't fit.

The hypothesis being tested for chi-square is

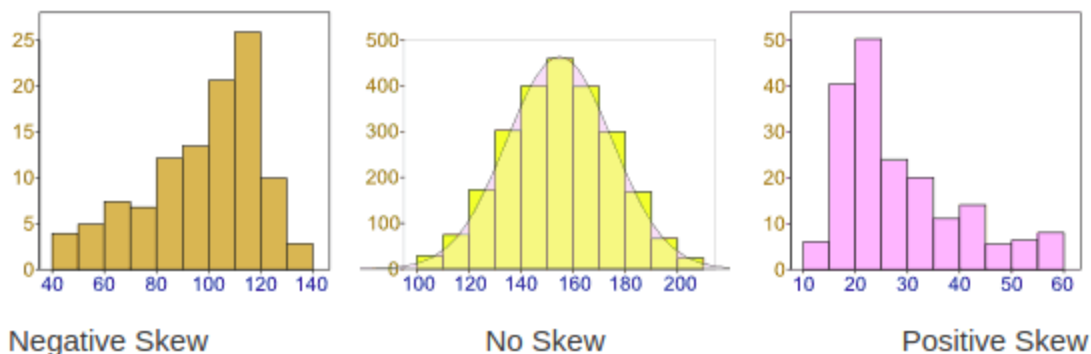
- **Null:** Variable A and Variable B are independent
- **Alternate:** Variable A and Variable B are not independent.

The statistic used to measure significance, in this case, is called chi-square statistic.

Skewness

Data can be “skewed”, meaning it tends to have a long tail on one side or the other.

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean.”



- **Negative Skew:** The long tail is on the “negative” (left) side of the peak. Generally, the **mean** less than **median** less than **mode** in this case.
- **No Skew:** There is no observed tail on any side of the peak. It is always symmetrical. **Mean**, **median** and **mode** are at the center of the peak i.e. $\text{mean} = \text{median} = \text{mode}$
- **Positive Skew:** The long tail is on the “positive” (right) side of the peak. Generally, **mode** < **median** < **mean** in this case.

We want our data to be **normally distributed**. The second assumption looks for skewness in our data. The null hypothesis states that our data is normally distributed. In our case, since the p-value for this is >0.05 , we can safely conclude that our null hypothesis holds hence our data is normally distributed.

Common transformations of this data include square root, cube root, and log.

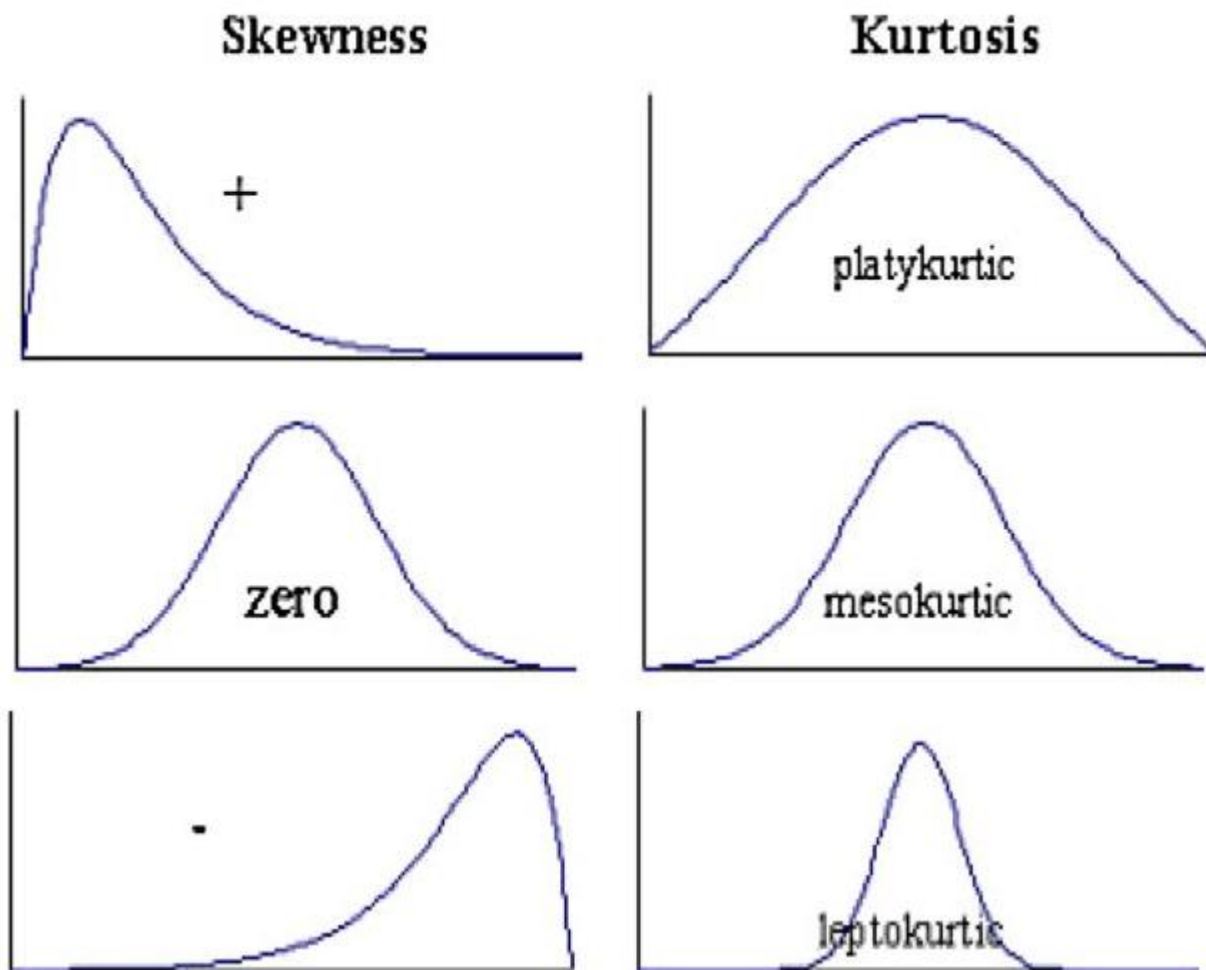
Cube root transformation: The cube root transformation involves converting x to $x^{(1/3)}$. This is a fairly strong transformation with a substantial effect on distribution shape: but is weaker than the logarithm. It can be applied to negative and zero values too. Negatively skewed data.


Square root transformation: Applied to positive values only. Hence, observe the values of column before applying.

Logarithm transformation: The logarithm, x to log base 10 of x , or x to log base e of x ($\ln x$), or x to log base 2 of x , is a strong transformation and can be used to reduce right skewness.

Kurtosis

Kurtosis is all about the tails of the distribution — not the peakedness or flatness. It is used to describe the extreme values in one versus the other tail. It is actually the measure of outliers present in the distribution.





High kurtosis in a data set is an indicator that data has heavy tails or outliers. If there is a high kurtosis, then, we need to investigate why we have so many outliers. It indicates a lot of things, maybe wrong data entry or other things. Investigate!

Low kurtosis in a data set is an indicator that data has light tails or lack of outliers. If we get low kurtosis (too good to be true), then also we need to investigate and trim the dataset of unwanted results.

- **Mesokurtic:** This is generally the ideal scenario where your data has a normal distribution.
- **Platykurtic (negative kurtosis score) (Kurtosis < 3):** A flatter peak is observed in this case because there are fewer data in your dataset which resides in the tail of the distribution i.e. tails are thinner as compared to the normal distribution. It has a shallower peak than normal which means that this distribution has thicker tails and that there are fewer chances of extreme outcomes compared to a normal distribution.
- **Leptokurtic (positive kurtosis score) (Kurtosis > 3):** A sharper peak is observed as compared to normal distribution because there is more data in your dataset which resides in the tail of the distribution as compared to the normal distribution. It has a lesser peak than normal which means that this distribution has fatter tails and that there are more chances of extreme outcomes compared to a normal distribution.

We want our data to be normally distributed. The third assumption looks for the amount of data present in the tail of the distribution. The null hypothesis states that our data is normally distributed. In our case, since the p-value for this is >0.05 , we can safely conclude that our null hypothesis holds hence our data is normally distributed.

Central Limit Theorem

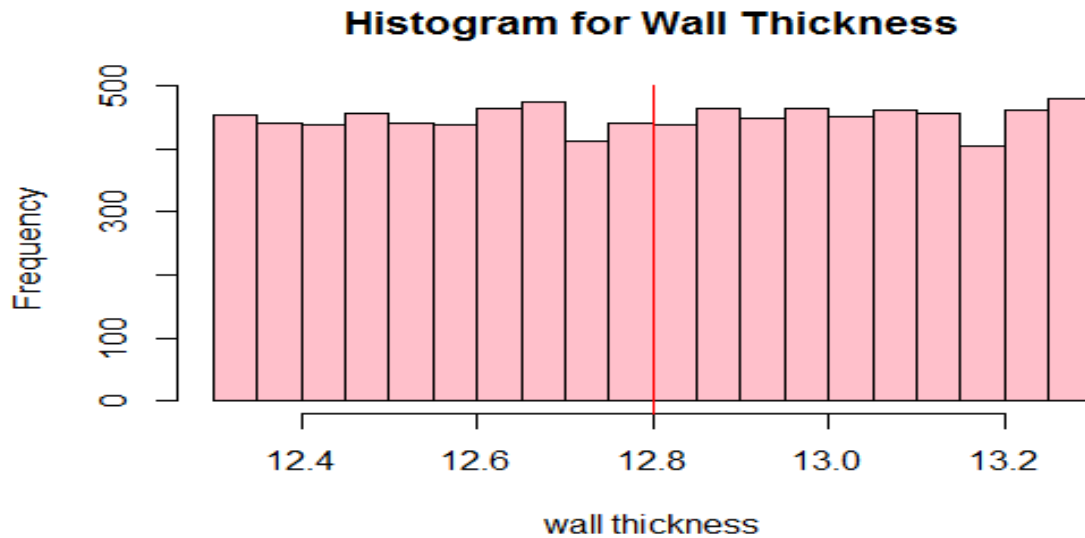
The central limit theorem states that if data is independently drawn from any distribution and the sample size is large enough, the sample mean always appears to be normally distributed.

Or

The **central limit theorem** states that the sampling distribution of the mean of any independent, random variable will be normal or nearly normal, if the sample size is large enough.

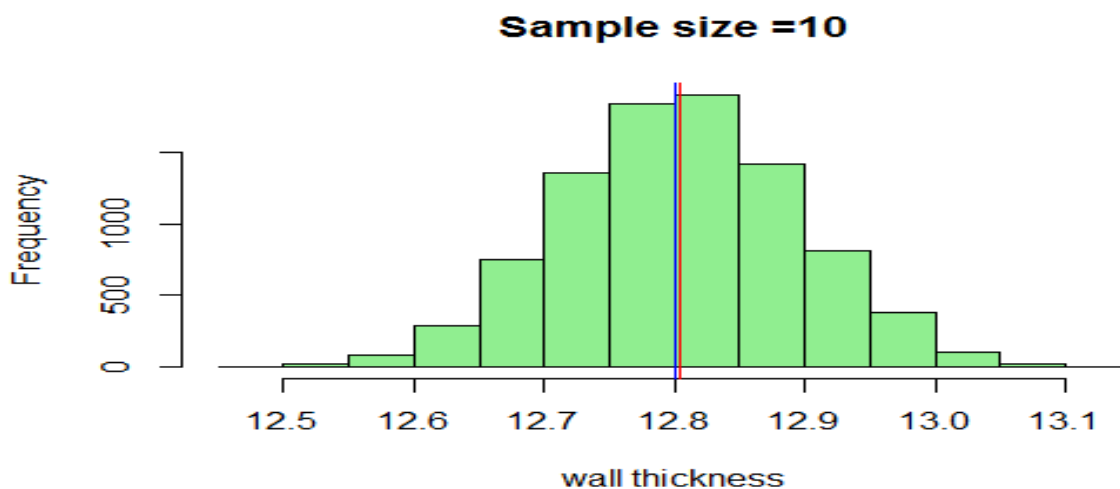
Example:

The central limit theorem will help us get around the problem of this data where the population is not normal. Therefore, we will simulate the central limit theorem on the below example.

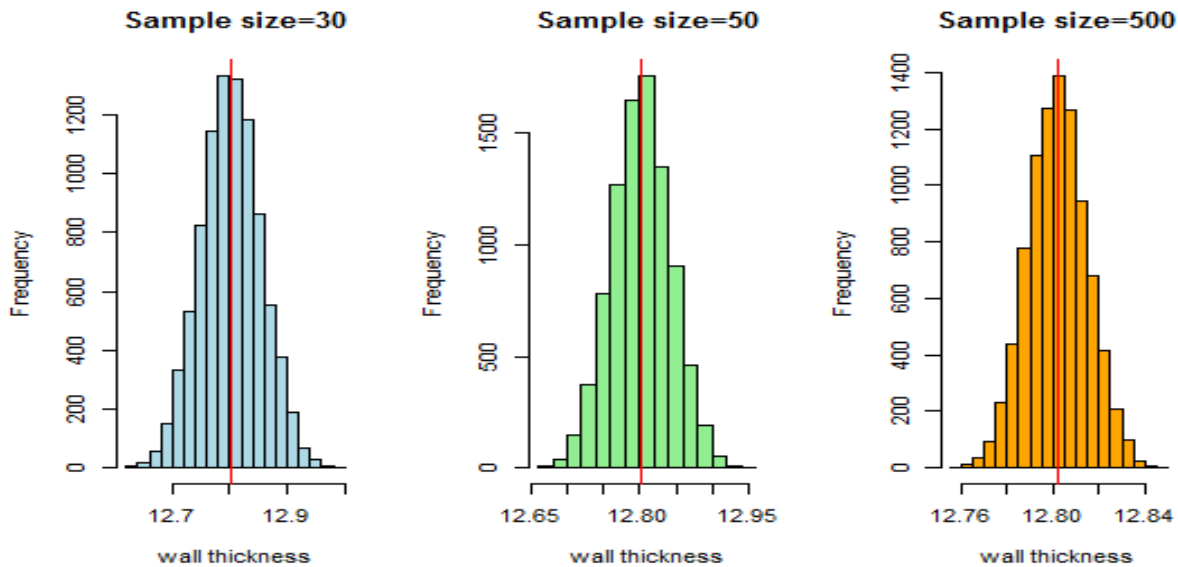


The above figure in red vertical line, that's the population mean. We can also see from the above plot that the population is not normal, right? Therefore, we need to draw sufficient samples of different sizes and compute their means (known as sample means). We will then plot those sample means to get a normal distribution.

In our example, we will draw sufficient samples of size 10, calculate their means, and plot them. I know that the minimum sample size taken should be 30 but let's just see what happens when we draw 10:



Now, we know that we'll get a very nice bell-shaped curve as the sample sizes increase. Let us now increase our sample size and see what we get.



Here, we get a good bell-shaped curve and the sampling distribution approaches normal distribution as the sample sizes increase. Therefore, we can consider the sampling distributions as normal.

Hypothesis Testing

A systematic way to select samples from a group or population with the intent of making a determination about the expected behavior of the entire group.

A hypothesis is similar to a theory

If you believe something might be true but don't yet have definitive proof, it is considered a theory until that proof is provided. Turning theories into accepted statements of fact is the basis of the scientific method, which consists of basic 4 steps:

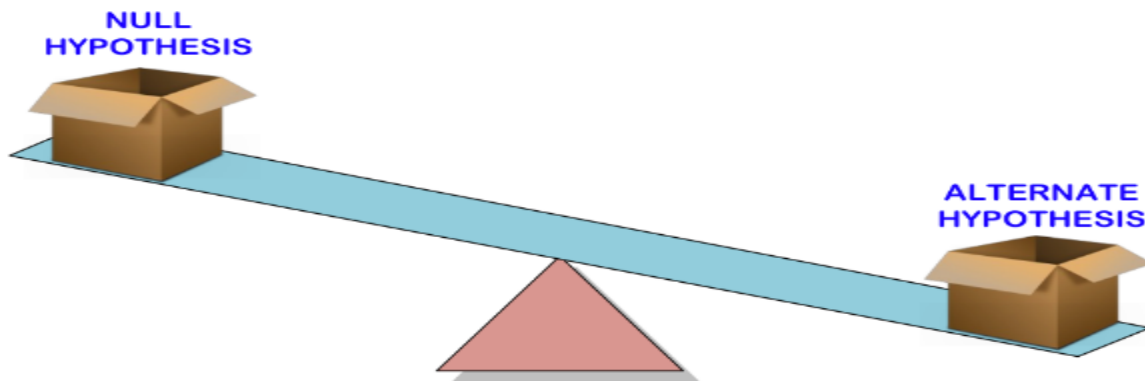
1. Formulate a Hypothesis
2. Collect Data
3. Analyze Data to Test Hypothesis
4. Draw Conclusions

Like many commonly used statistical tools today, A/B testing and multivariate testing are forms of hypothesis testing, so it is important to begin your website testing with a strong **hypothesis statement**.

For example, if you had reason to believe that the color of your landing page might be having a detrimental effect on conversions, your hypothesis statement could be:

"Changing my landing page color from black to blue will have a statistically significant impact on conversions."

Once this hypothesis is established, you need to run your test to prove (or disprove) it. Including the words “statistically significant” in the hypothesis statement is important, since it means your sample sizes need to be adequate to analyze it as such.



Null Hypothesis: The word “null” comes from the Latin word “nullus”, meaning not any or nothing. Perhaps this definition is helpful in understanding this often confusing term. In hypothesis testing, your null hypothesis is that nothing will change or improve between the two groups of data. Obviously, this is not what you want to prove, but rather what you want to disprove. For example, your null hypothesis might be that your landing page color change will have no impact on conversions.

Or

A statement in which no difference or effect is expected. If the null hypothesis is not rejected, no changes will be made.


Alternate Hypothesis

A statement that some difference or effect is expected. Accepting the alternative hypothesis will lead to changes in opinions or actions. It is the opposite of the null hypothesis.

Or

Also known as the experimental or “research” hypothesis, this is what you are really aiming to prove through testing. The alternate hypothesis is simply the opposite of the null hypothesis. In our landing page color change example, the hypothesis statement is actually the alternate hypothesis. Therefore, you want to disprove your null hypothesis in order to prove the alternate.

An analogy that is often used to describe hypothesis testing is a defendant on trial, since he is presumed innocent until proven guilty. This is equivalent to the null hypothesis being presumed true until proven false. In the courtroom, the jury decides whether or not there is enough evidence to disprove innocence. In an A/B or multivariate test, the tester sets a significance p-



value threshold (such as .05 or 5%) for the test that determines how unlikely the null hypothesis needs to be before we can confidently reject it.

Hypothesis Tests

Statisticians follow a formal process to determine whether to reject a null hypothesis, based on sample data. This process, called **hypothesis testing**, consists of four steps.

- **State the hypotheses:** This involves stating the null and alternative hypotheses. The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false.
- **Formulate an analysis plan:** The analysis plan describes how to use sample data to evaluate the null hypothesis. The evaluation often focuses around a single test statistic.
- **Analyze sample data:** Find the value of the test statistic (mean score, proportion, t statistic, z-score, etc.) described in the analysis plan.
- **Interpret results:** Apply the decision rule described in the analysis plan. If the value of the test statistic is unlikely, based on the null hypothesis, reject the null hypothesis.

Decision Errors


Two types of errors can result from a hypothesis test.

- **Type I error:** A Type I error occurs when the researcher rejects a null hypothesis when it is true. The probability of committing a Type I error is called the **significance level**. This probability is also called **alpha**, and is often denoted by α .
- **Type II error:** A Type II error occurs when the researcher fails to reject a null hypothesis that is false. The probability of committing a Type II error is called **Beta**, and is often denoted by β . The probability of not committing a Type II error is called the **Power** of the test.

One-Tailed and Two-Tailed Tests

A test of a statistical hypothesis, where the region of rejection is on only one side of the sampling distribution, is called a **one-tailed test**. For example, suppose the null hypothesis states that the mean is less than or equal to 10. The alternative hypothesis would be that the mean is greater than 10. The region of rejection would consist of a range of numbers located on the right side of sampling distribution; that is, a set of numbers greater than 10.

A test of a statistical hypothesis, where the region of rejection is on both sides of the sampling distribution, is called a **two-tailed test**. For example, suppose the null hypothesis states that the mean is equal to 10. The alternative hypothesis would be that the mean is less than 10 or greater than 10. The region of rejection would consist of a range of numbers located on both



sides of sampling distribution; that is, the region of rejection would consist partly of numbers that were less than 10 and partly of numbers that were greater than 10.

Power of a Hypothesis Test

The probability of not committing a Type II error is called the **power** of a hypothesis test.

Effect Size

To compute the power of the test, one offers an alternative view about the "true" value of the population parameter, assuming that the null hypothesis is false. The **effect size** is the difference between the true value and the value specified in the null hypothesis.

$$\text{Effect size} = \text{True value} - \text{Hypothesized value}$$

For example, suppose the null hypothesis states that a population mean is equal to 100. A researcher might ask: What is the probability of rejecting the null hypothesis if the true population mean is equal to 90? In this example, the effect size would be $90 - 100$, which equals -10 .

Factors That Affect Power

The power of a hypothesis test is affected by three factors.

- Sample size (n). Other things being equal, the greater the sample size, the greater the power of the test.
- Significance level (α). The lower the significance level, the lower the power of the test. If you reduce the significance level (e.g., from 0.05 to 0.01), the region of acceptance gets bigger. As a result, you are less likely to reject the null hypothesis. This means you are less likely to reject the null hypothesis when it is false, so you are more likely to make a Type II error. In short, the power of the test is reduced when you reduce the significance level; and vice versa.
- The "true" value of the parameter being tested. The greater the difference between the "true" value of a parameter and the value specified in the null hypothesis, the greater the power of the test. That is, the greater the effect size, the greater the power of the test.