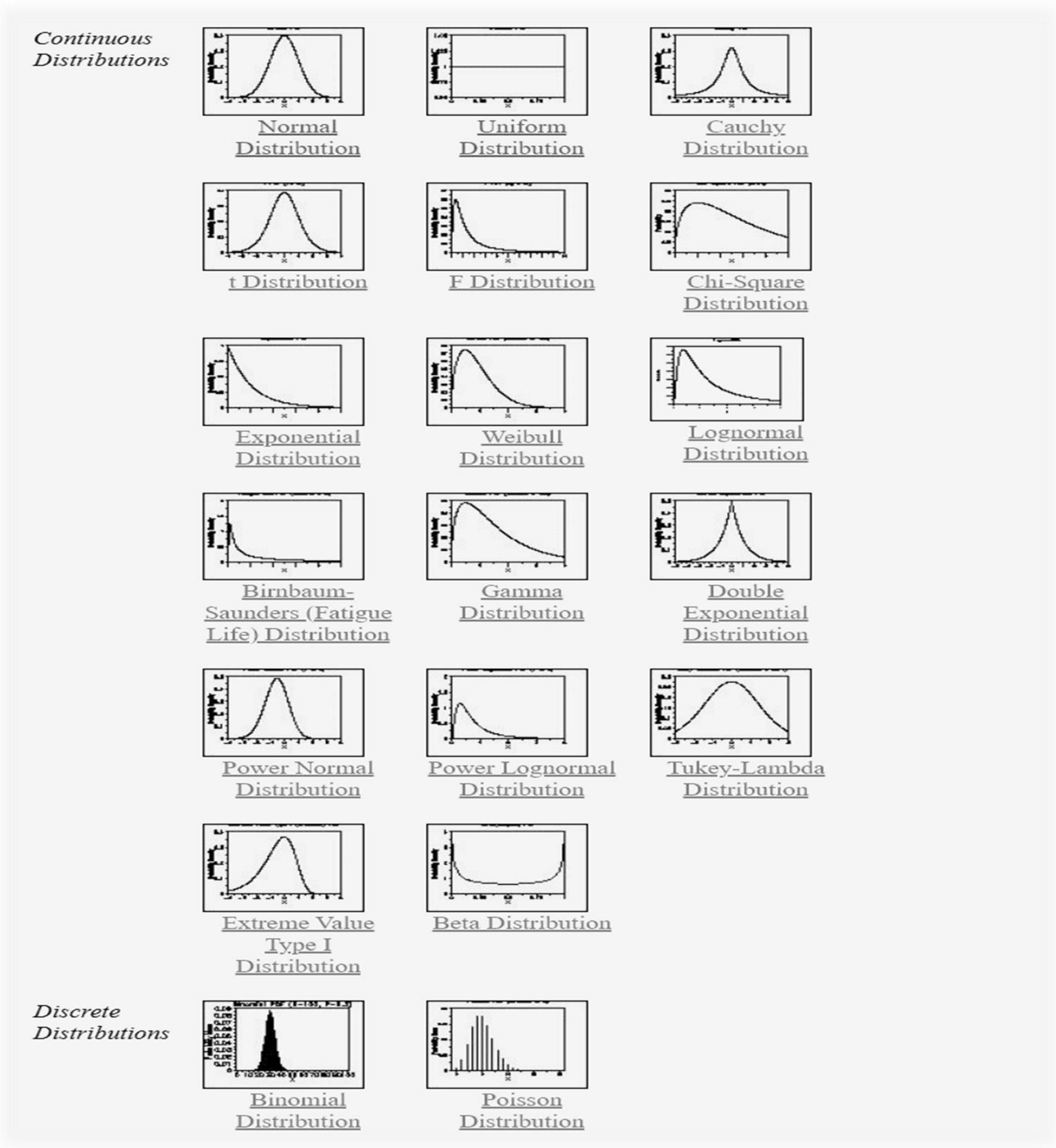
Types of Distributions:



Continuous Distributions

Normal Distribution

Uniform Distribution

Cauchy Distribution

t Distribution

F Distribution

Chi-Square Distribution

Exponential Distribution

Weibull Distribution

Lognormal Distribution

Birnbaum-Saunders (Fatigue Life) Distribution

Gamma Distribution

Double Exponential Distribution

Power Normal Distribution

Power Lognormal Distribution

Tukey-Lambda Distribution

Extreme Value Type I Distribution

Beta Distribution

Discrete Distributions

Binomial Distribution

Poisson Distribution

What is a Design Pattern?

Design patterns are design level solutions for recurring problems that we software engineers come across often. It’s not code - I repeat, ✕ CODE. It is like a description on how to tackle these problems and design a solution.

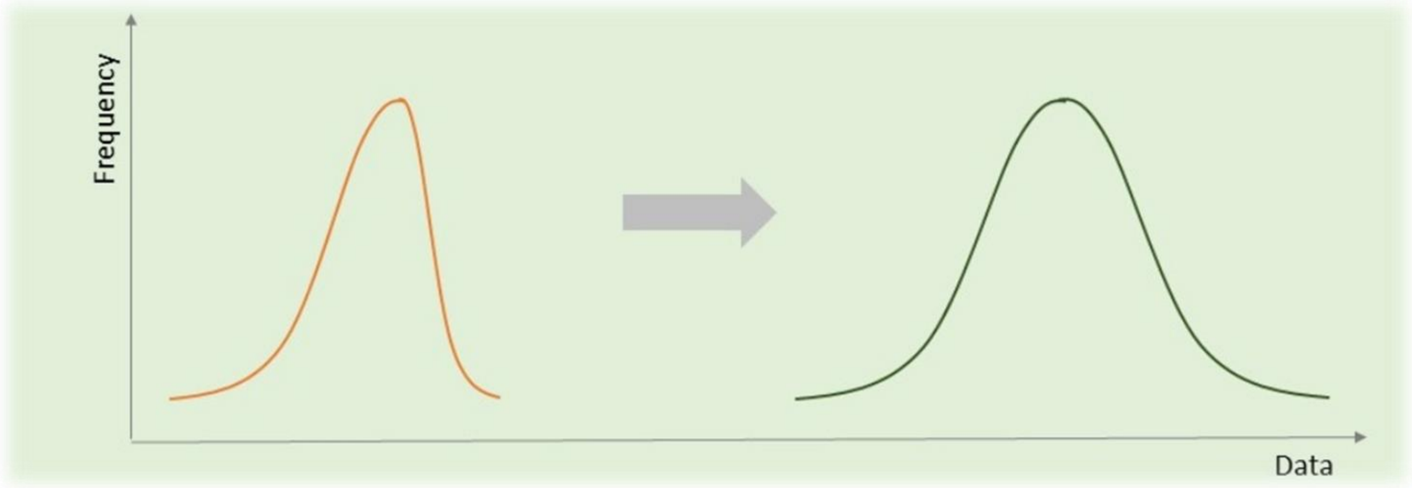
Using these patterns is considered good practice, as the design of the solution is quite tried and tested, resulting in higher readability of the final code. Design patterns are quite often created for and used by OOP Languages, like Java, in which most of the examples from here on will be written.

Types of design patterns:

These 26 can be classified into 3 types:

1. Creational: These patterns are designed for class instantiation. They can be either class-creation patterns or object-creational patterns.
2. Structural: These patterns are designed with regard to a class's structure and composition. The main goal of most of these patterns is to increase the functionality of the class(es) involved, without changing much of its composition.
3. Behavioral: These patterns are designed depending on how one class communicates with others.

How do I transform my data to a normal distribution?



Check for Outliers

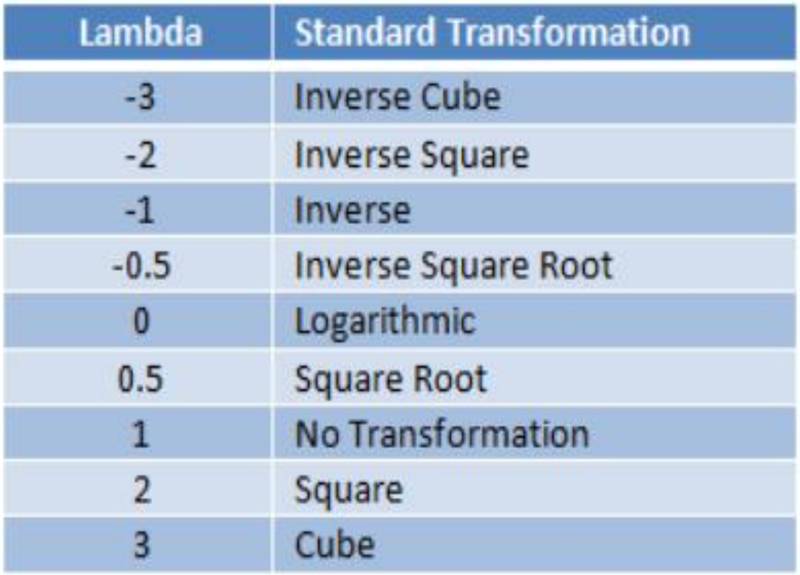
The first thing we need to do check if the data is not normal because of any outliers. A normal data does not have any outliers – hence, if there are outliers in your data, then that may be the reason that the data is not normally distributed. First we need to check if the outliers in the data are because of any data entry errors. If so, we can correct the data and then check if the data is normally distributed. If there are no data entry errors, the next question to ask is if the outliers are because of some special causes which are not going to recur in the future. If so, it may be okay to note the reasons and then delete these outliers. However, if these outliers have a chance of recurring in the future then it would not be appropriate to just blindly delete them from analysis. We need to look for other ways of handling this data.

Box-Cox Transformation

The second approach is to transform the data such that the transformed data is normally distributed. There are some transformations that have been found to make the transformed data normal. For example, if you square the data values, the squared values may be normal. Or, in some cases, the square root of the data or the reciprocal of the data may be normally distributed. In other cases, the logarithm of the data may be normally distributed. Such simple transformations of the data to make the data normal can be grouped together under a transformation called the Box-Cox transformation. The Box-Cox transformation is given by the following general formula:Box-Cox Formula

$$y = x^{\lambda} \; for \; \lambda \neq 0 \; and \; y = \ln(x) \; for \; \lambda = 0$$

Where, x is the raw data and y is the transformed data and lambda is the transformation constant. If lambda = 1, then there is no transformation. If lambda = 2, then it is the square transformation and so on. The following table provides the names of some standard transformations:



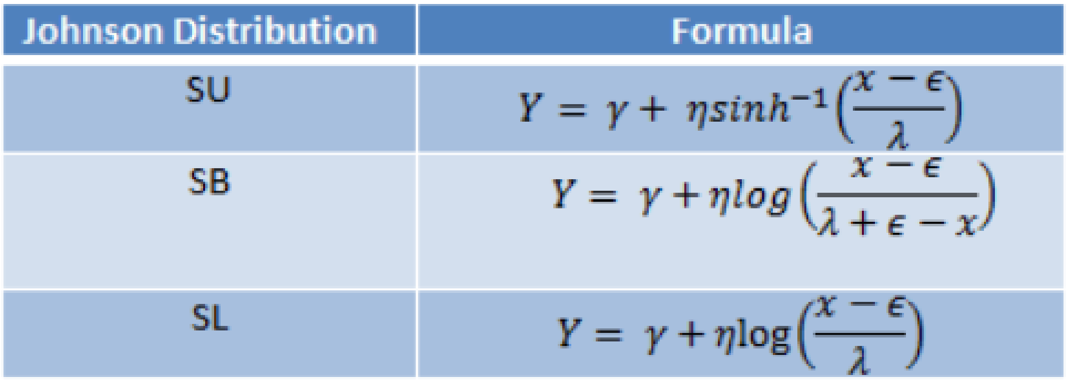
Lambda	Standard Transformation
-3	Inverse Cube
-2	Inverse Square
-1	Inverse
-0.5	Inverse Square Root
0	Logarithmic
0.5	Square Root
1	No Transformation
2	Square
3	Cube

How to Fit the Box-Cox Transformation

There are several approaches to determine the value of lambda for the Box-Cox transformation. The most commonly used approach is to use the Most Likely Estimate (MLE) approach. Getting into the details about this approach is out of scope of this article. A simple approach to determine the value of lambda is to vary the value of lambda from -5 to +5 and then determine which value of lambda produces a distribution that is as close to a normal distribution as possible. The value of lambda is selected that provides the smallest value of the standard deviation of the variation between the transformed data and a normally distributed data. Of course, lambda can take any value (say 1.63), but it may be hard to explain what that means to others. Some users like to choose the Box-Cox transformation to the values of lambda shown in the above table so that the transformation can be easily understood by the users. Note that there is no guarantee that a Box-Cox transformation will always result in a normal distribution. It is possible that none of the values of lambda can result in a normally distributed data.

Johnson Transformation

A third approach to transform the data to a normal distribution is to use another type of more complex transformation called the Johnson family of transformations. There are three different families of Johnson distributions:

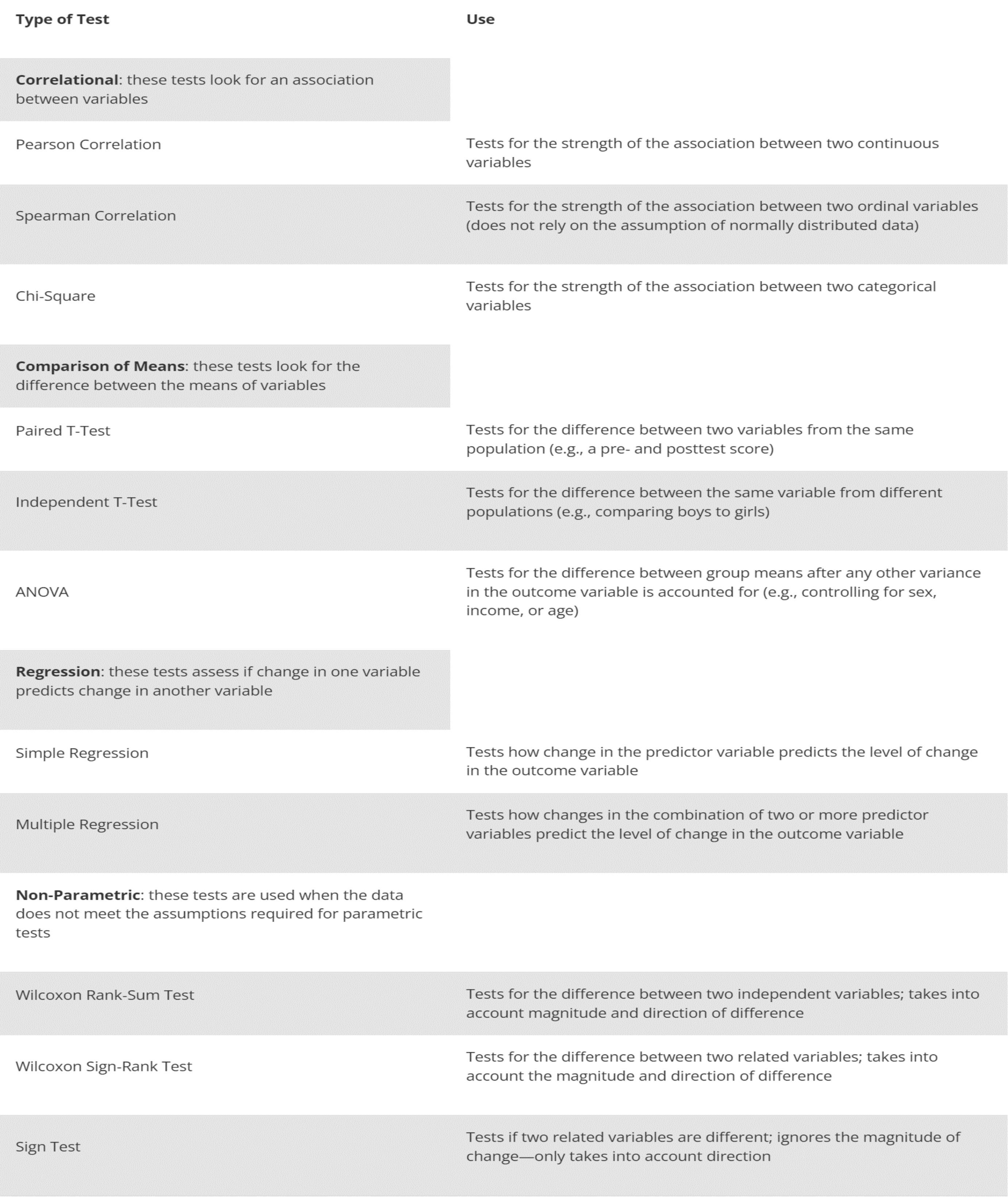


Johnson Distribution	Formula
SU	$Y = \gamma + \eta sinh^{-1}\left(\frac{x-\epsilon}{\lambda}\right)$
SB	$Y = \gamma + \eta log\left(\frac{x-\epsilon}{\lambda+\epsilon-x}\right)$
SL	$Y = \gamma + \eta \log\left(\frac{x-\epsilon}{\lambda}\right)$

Where, Y is the transformed data, X is the raw data, and eta, epsilon, and lambda are the Johnson parameters. Decision rules have been formulated for the selection of the appropriate Johnson family of distributions SU, SB, and SL. There are several algorithms available to fit the Johnson parameters for a given data set. However, due to complex nature of these algorithms, the solutions are not very straightforward and require the use of appropriate software to estimate these parameters. Similar to a Box-Cox transformation, a computer can run through several combinations of these Johnson parameters to determine which set of parameters makes the transformed data as close to normal as possible. Since there are several parameters to fit the Johnson transformation, we usually find that a Johnson transformation does a better job of transforming the data to a normal distribution compared to a Box-Cox transformation. Similar to the Box-Cox transformation, there is no guarantee that a Johnson transformation will be successful in transforming a data to the normal transformation. It should be pointed out that when you transform the raw data using one of these transformations, the specification limits also need to be transformed if you need to calculate the process capability.

Types of Statistical Tests

After looking at the distribution of data and perhaps conducting some descriptive statistics to find out the mean, median, or mode, it is time to make some inferences about the data. As mentioned previously, inferential statistics are the set of statistical tests researchers use to make inferences about data. These statistical tests allow researchers to make inferences because they can show whether an observed pattern is due to intervention or chance. There is a wide range of statistical tests. The decision of which statistical test to use depends on the research design, the distribution of the data, and the type of variable. In general, if the data is normally distributed, parametric tests should be used. If the data is non-normal, non-parametric tests should be used. Below is a list of just a few common statistical tests and their uses.



Type of Test	Use
Correlational: these tests look for an association between variables	
Pearson Correlation	Tests for the strength of the association between two continuous variables
Spearman Correlation	Tests for the strength of the association between two ordinal variables (does not rely on the assumption of normally distributed data)
Chi-Square	Tests for the strength of the association between two categorical variables
Comparison of Means: these tests look for the difference between the means of variables	
Paired T-Test	Tests for the difference between two variables from the same population (e.g., a pre- and posttest score)
Independent T-Test	Tests for the difference between the same variable from different populations (e.g., comparing boys to girls)
ANOVA	Tests for the difference between group means after any other variance in the outcome variable is accounted for (e.g., controlling for sex, income, or age)
Regression: these tests assess if change in one variable predicts change in another variable	
Simple Regression	Tests how change in the predictor variable predicts the level of change in the outcome variable
Multiple Regression	Tests how changes in the combination of two or more predictor variables predict the level of change in the outcome variable
Non-Parametric: these tests are used when the data does not meet the assumptions required for parametric tests	
Wilcoxon Rank-Sum Test	Tests for the difference between two independent variables; takes into account magnitude and direction of difference
Wilcoxon Sign-Rank Test	Tests for the difference between two related variables; takes into account the magnitude and direction of difference
Sign Test	Tests if two related variables are different; ignores the magnitude of change—only takes into account direction

What is Confidence Interval?

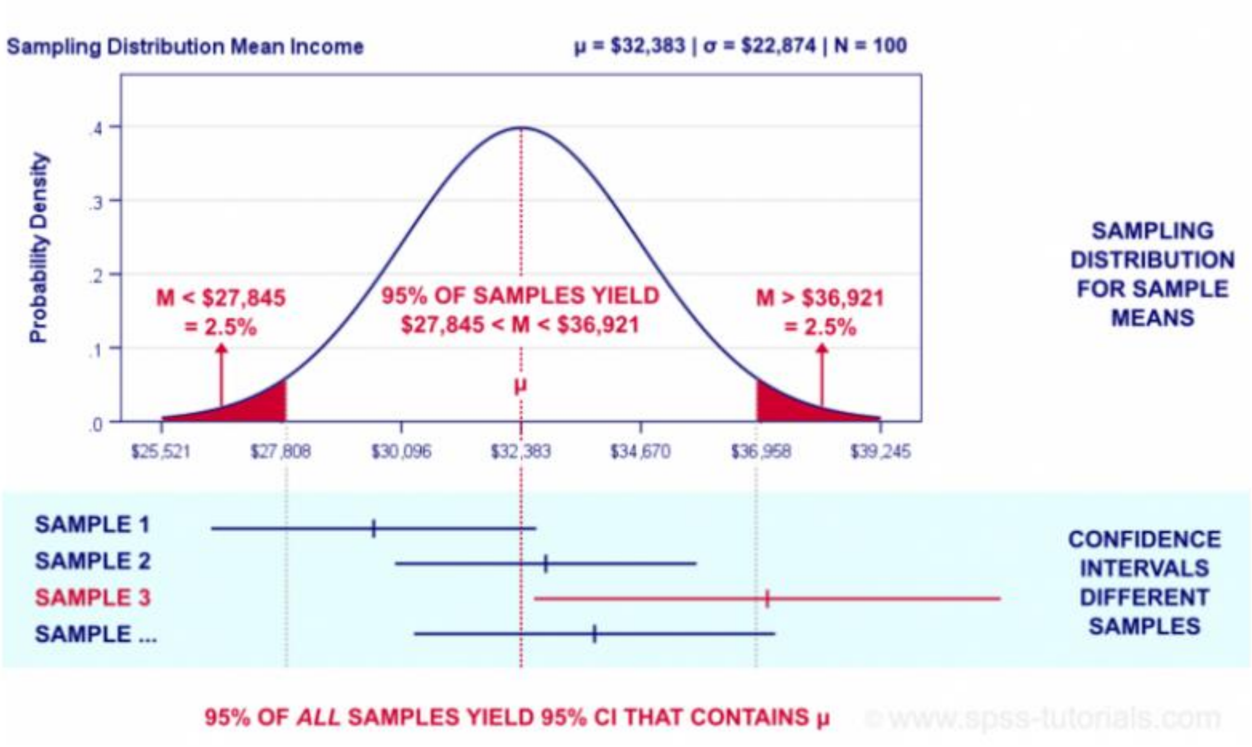
The confidence interval is used to represent the interval or range of values needed to match a confidence level for estimating the parameter of the entire population or population proportion. Recall that Statistics is about estimation. When there is a need to estimate the statistics about the population parameter, it is considered as a good practice to represent the estimate as a confidence interval. The statistics of the population parameter generally represents the mean or median. And, the confidence level is represented using the number such as 98% confidence, 95% confidence etc.

The confidence interval is associated with the confidence level represented using a number, say, N, and termed as an N% confidence interval. N can take values such as 95, 90, etc. An N% confidence interval would mean the following – If an experiment to find the average height of male out of 100 male, is performed for, say, 50 times, the interval in which the average height will fall for 95% of times (45 times) will be between, say, 173 and 179 cm. Thus, a 95% confidence interval for average height will be 173 and 179 cm.

Confidence interval is used to estimate the statistics such as population mean or median and population proportion

The error of the model prediction is a classic example of proportion. The error can be represented as the proportion of misclassification in prediction done by the model. So, the error found in the model trained on sample data is termed as sample error. The objective is to estimate the true error of the model given population data. This can be represented using a confidence interval. Confidence interval can be used to estimate the true error of the model as a function of the sampling error.

Here is a diagram which can be used to understand the confidence interval concepts:



Why is the confidence interval measurement needed?

Simply speaking, confidence interval measurement is needed to find out the range in which the population parameter will fall based on the outcomes from one or more experiments performed on different samples taken from the population. It is used to communicate the accuracy of the estimate of the population parameters. For any outcome to be found using the experiments, you are never going to be 100% confident about the population parameter based on the experiments. Thus, you need confidence intervals to represent the range in which the population parameter will fall. If you’re 95% confident, or 98% confident, that’s usually considered “good enough” in statistics. That percentage of confidence is the confidence interval. For the N% confidence interval, we are saying that given numerous experiments performed, for N times, the population parameter (P) will fall in the range of P + m and P – m. And, the value of m will change with N. Let’s understand with an example. Confidence intervals are usually reported in the context of a margin of error, though they are two unique values.

Let’s say we want to estimate the mean height of the male population in the 20-30 age group in India. Gathering and calculating the height of every individual in the 20-30 age group in India is a real herculean task. Here the statistics of population parameter is mean height. Is there a way in which we can get a fair estimate of this population parameter, mean height? One of the ways is to take a sample of 1000 male individuals from the key cities, gather their heights, and calculate the mean. The objective is to estimate the mean height of the population based on the mean height calculated from the sample. The estimation of the mean height of population is done using confidence interval. Let’s say the procedure is followed 50 times by taking different samples of 1000 male individuals and the following got observed:

- In 48 times, the mean height fell in the range of 175 and 178 cm. Note that 48 is approx 95% of 50. Thus, with 95% confidence level, one could say that the mean height of the population will be in the range of 175 and 178 cm.
- In 45 times, the mean height fell in the range of 173 and 179 cm. Note that 45 is approx 90% of 50. Thus, with 90% confidence level, one could say that the mean height of the population will be in the range of 173 and 179 cm.

What affects the width of the confidence interval?

- Variation: Greater is the variation in the population, larger is the width of the confidence interval and vice versa.

- Sample size: Smaller is the sample size, larger is the width of the confidence interval and vice versa. For the smaller sample, the information contained will be less. Thus, there will be larger confidence interval width.

How to calculate confidence interval?

Confidence interval can be calculated using a normal distribution (Z-distribution) or T-distribution. T-distribution is used if the sample size is smaller (less than 30) or the information about the distribution is not known.

For calculating confidence interval for statistics such as population mean, the following formula can be used. s represents standard deviation and n represents the size of the sample. X bar represents sample mean and t represents t-distribution. T-distribution is used as in most cases the population distribution is not known. In case, the population distribution is known in advance and is found to be the normal distribution, one can use z in place of t.

$$\overline{x} \pm t\left(\frac{s}{\sqrt{n}}\right)$$

The following formula can be used to calculate confidence intervals for estimating the population proportion. For determining the estimate of the population proportion, the normal distribution is used and, thus, z. p represents the mean proportion of the sample. n represents the size of the sample.

$$p \pm z\sqrt{\frac{p(1-p)}{n}}$$