**Analysis of Statistical Tests for Normality**

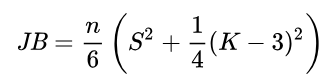
**Note :** For all the statistical test except Anderson Darling Test and Kolmogorov Smirnov Test

H0 : Data follows Normal Distribution

H1: Data does not follow Normal Distribution

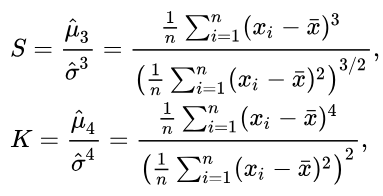
1. **Jarque Bera Test**

Formula for test statistic for Jarque Bera Test :



Where n is no of observation , s is skewness and k is kurtosis.

Formula of Skewness and Kurtosis for Jarque Bera Test :



**Skewness (S):** Skewness measures the asymmetry of a probability distribution about its mean. Positive skewness indicates a distribution with an asymmetric tail extending towards more positive values. Negative skewness indicates a distribution with an asymmetric tail extending towards more negative values. For a normal distribution, the skewness is zero.

**Kurtosis (K):** Kurtosis measures the “tailedness” of the probability distribution. It compares the amount of data close to the mean and those far away in the tails. A normal distribution has a kurtosis of three. Excess kurtosis is therefore K-3, and for a normal distribution, this is zero.

Assumption : Data follows Normal distribution

If the data comes from a normal distribution, the JB statistic asymptotically has a chi-squared distribution with two degrees of freedom. Therefore, the statistic can be used to test the hypothesis that the data are from a normal distribution. The null hypothesis is a joint hypothesis of the skewness being zero and the excess kurtosis being zero. If the JB statistic is significantly different from zero, we reject the null hypothesis and conclude that the data do not have a normal distribution.

**Problem with Jarque Bera Test** :

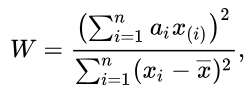
* Even if the data have very small amount outliers ( only 5 to 10) mean value will be affected and it will cause remarkable change in skewness and kurtosis according to its formula. (If either the skewness or kurtosis increases value of JB statistic will increase) Hence p value of JB statistic will reduce drastically and Null hypothesis will be rejected which is a joint hypothesis of skewness being 0 and kurtosis being 0 which actually should not be rejected.
* For small samples, the chi-square approximation is overly sensitive, often rejecting the null hypothesis when it is true. Therefore, for smaller samples, the Jarque-Bera test uses a table derived from Monte Carlo simulations in order to interpolate p-values.

1. **Shapiro Wilk Test**

H0 : Data follows normal distribution

H1 : Data does not follow normal distribution

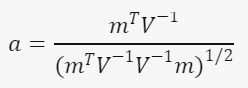
Formula of test statistic of Shapiro Wilk test :



x(i) is the ith order statistic i.e. ith smallest number in the sample

W: This is the test statistic that we calculate. The closer W is to 1, the more likely it is that the sample came from a normally distributed population.

ai​: These are coefficients calculated from the covariance matrix of the order statistics of a standard normal distribution and the expected values of these order statistics. They are calculated as follows:



Assumption : Data follows Normal distribution

Problem with Shapiro Wilk Test :

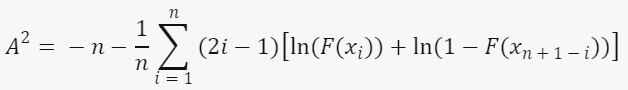
* Does not work well for large size datasets. i.e. For large size this test may reject the null hypothesis even if it is true ( Typically this value is 5000 according to warning given in python)
* Affected by outliers i.e. Data is actually following normal distribution but it is having a few outliers. So that value of test statistic for Shapiro wilk test (W) will deviate significantly from 1 which implies data is not following normal distribution but actually it is.

1. **Anderson Darling Test**

H0 : Data follows specified distribution

H1 : Data does not follow specified normal distribution

Formula of test statistic for Anderson Darling Test :



Where:  
n = the sample size,  
F(x) = CDF (Cumulative Distribution Function) for the specified distribution,  
i = the ith sample, calculated when the data is sorted in ascending order

The test statistic in the Anderson-Darling test (A2) is based on the differences between the EDF and the CDF. It places more weight on the tails of the distribution, making it more sensitive to deviations in those regions.

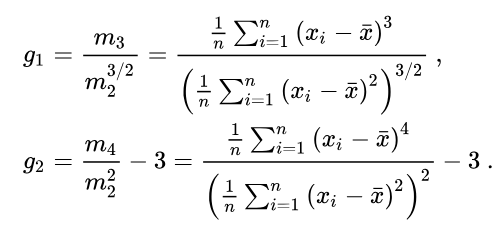
EDF : Empirical Distribution Function of sample data

Assumptions : Continuous distribution, Independence of data points, Identical distribution of all data points

Problem with Anderson Darling Test : High time complexity comparing to other statistical tests for univariate normality

1. **D’Agostino’s K2 Test**

Formula for sample skewness and sample kurtosis for D’Agostino’s K2 Test :



**Problem with D’Agostino’s K2 test** : Normal distribution is decided based on skewness and kurtosis in this test. As mention earlier in Jarque Bera test that when there are a few outliers in data skewness and kurtosis is affected i.e. its value will deviate from zero significantly and if either value of skewness or kurtosis increases test and null hypothesis will be rejected but actually it should not as data is following normal distribution.

For small size data it may not work well as for small size sample test might not assess normality accurately.

1. **Kolmogorov Smirnov Test**

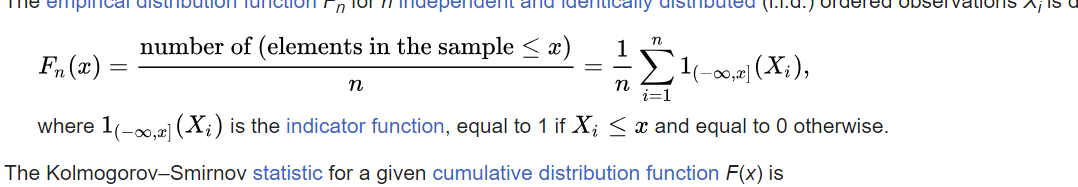
H0: Data comes from a specific distribution

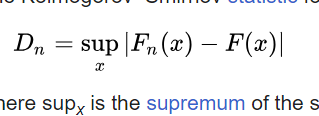
H1: Data does not come from a specific distribution

It is a Non parametric test. There are 2 types of KS test i.e. One Sample KS test and Two sample KS test. In one sample it tests whether the sample data follows a specific distribution or not

One sample KS test assesses similarity between two probability distribution i.e. EDF ( Empirical Distribution Function) of sample data and CDF ( Cumulative Distribution Function) of reference distribution i.e. in our case it is normal distribution

Formula of EDF for One sample KS test:





Where supx is the supremum of set of distances

Characteristic of KS test : Sensitivity is higher at the center of the distribution and lower at the tails

1. **Lilliefors Test**

Extended version of Kolmogorov Smirnov test

In R documentation it is written that “Compared to the Anderson-Darling test and the Cramer-von Mises test it is known to perform worse.”

Characteristics : Unlike Kolmogorov Smirnov test it is sensitive to both tails and peak of data distribution.

It measures the difference between EDF and CDF. In order to calculate first calculate z score for all the samples and then calculate the EDF using these Z scores.

1. **Doornik Hansen Test**

It is based on the skewness and kurtosis of the multivariate data. These statistics are transformed to ensure independence before conducting the test

Mardia’s test is also about skewness and kurtosis

Time complexity : O(n) which much better than Mardia’s test which is O(n2)

Test statistic : DH = z12 + z22

Where z1 represents the transformation of skewness and z2 represents transformation of kurtosis

Approximately test statistics follows chi square distribution with 2 degrees of freedom

* **Final Outcome :**

1. Which statistical test should be used for Univariate Normality and why ?

Ans : Anderson Darling Test

1. Which statistical test should be used for Multivariate Normality and why ?

Ans : Doornik Hansen Test

The Lilliefors test is an adaptation of the Kolmogorov-Smirnov test for univariate normality testing. It's primarily designed for assessing the normality of univariate data.

While the Lilliefors test can be extended to multivariate data by applying the test to each dimension separately or by using a multivariate version of the test, this extension may not always capture the complex dependencies present in multivariate data accurately.

The Doornik-Hansen test is a test for multivariate normality based on the multivariate skewness and kurtosis coefficients.

The test is robust to departures from normality, including skewness and kurtosis.

It can handle both small and large sample sizes efficiently.

while the Lilliefors test can be extended to multivariate data, it may not offer the same robustness, sensitivity, and ease of use as purpose-built tests like the Doornik-Hansen test.