

**Machine learning approach to test the normality of the data**

**Prepared by:**

Hussein Soboh

1185061

**Committee:**

|  |  |
| --- | --- |
| Dr. Hassan Abu Hassan | Advisor |
| Dr. Tareq Sadeq | Member |
| Dr. Radi Jarrar | Member |

***Submitted in partial fulfillment of the requirements of the “Master Degree in Applied Statistic and Data Science” from the faculty of Graduate Studies at Birzeit University - Palestine***

Table of Content

[Table of Content I](#_Toc42200485)

[List of Tables III](#_Toc42200486)

[List of Figures IV](#_Toc42200487)

[Abstract V](#_Toc42200488)

[ملخص VI](#_Toc42200489)

[Chapter One 1](#_Toc42200490)

[Introduction 1](#_Toc42200491)

[1.1 Background 2](#_Toc42200492)

[1.2 Problem definition 4](#_Toc42200493)

[1.3 Research Objectives 5](#_Toc42200494)

[1.4 Limitations of the study 5](#_Toc42200495)

[Chapter Two 7](#_Toc42200496)

[Literature review 7](#_Toc42200497)

[2.1 Normality tests 7](#_Toc42200499)

[2.1.1 Visual tests 7](#_Toc42200500)

[2.1.2 Statistical tests 7](#_Toc42200501)

[2.2 Previous comparisons 9](#_Toc42200502)

[2.3 Limitations of the statistical tests 11](#_Toc42200503)

[Chapter Three 12](#_Toc42200504)

[Methodology 12](#_Toc42200505)

[3.1 Alternative Distributions 12](#_Toc42200507)

[3.1.1 Beta Distribution 13](#_Toc42200508)

[3.1.2 Student t-distribution 14](#_Toc42200509)

[3.1.3 Chi-squared Distribution 15](#_Toc42200510)

[3.1.4 Log-normal Distribution 15](#_Toc42200511)

[3.1.5 Weibull Distribution: 16](#_Toc42200512)

[3.1.6 Tukey Distribution 17](#_Toc42200513)

[3.1.7 Laplace Distribution 18](#_Toc42200514)

[3.1.8 Uniform (Rectangular) Distribution 19](#_Toc42200515)

[3.1.9 Truncated Normal Distribution 20](#_Toc42200516)

[3.2 Model construction 21](#_Toc42200517)

[3.2.1 Process 21](#_Toc42200518)

[3.2.2 Classification Techniques 22](#_Toc42200519)

[3.2.3 Data set 26](#_Toc42200520)

[3.2.4 Training 27](#_Toc42200521)

[3.2.5 Evaluation 28](#_Toc42200522)

[3.3 Power comparison test 28](#_Toc42200523)

[3.4 Toolbox 29](#_Toc42200524)

[Chapter Four 30](#_Toc42200525)

[Simulation and Results 30](#_Toc42200526)

[4.1 Classification model 30](#_Toc42200528)

[4.1.1 Examining Data 30](#_Toc42200529)

[4.1.2 Features 30](#_Toc42200530)

[4.1.3 Training code (caret, control, code snippet…) 30](#_Toc42200531)

[4.1.4 Results 30](#_Toc42200532)

[4.1.5 Error analysis 30](#_Toc42200533)

[4.2 Power comparison 30](#_Toc42200534)

[4.2.1 Model thresholds 30](#_Toc42200535)

[4.2.2 Results 30](#_Toc42200536)

[References 31](#_Toc42200537)

List of Tables

[Table 1: Alternative distributions used in the research 12](#_Toc39144847)

List of Figures

[Figure 1: PDF of Normal distribution (Normal distribution, 2020) 2](#_Toc42201437)

[Figure 2: Probability density function for Beta variate **β**: v, ω 13](#_Toc42201438)

[Figure 3: Probability density function for Student’s t variate, **t:**  14](#_Toc42201439)

[Figure 4: Probability density function for the Chi-Squared variate 15](#_Toc42201440)

[Figure 5: Probability density function for the Log-normal variate **L**: 16](#_Toc42201441)

[Figure 6: Probability density function for Weibull variate 17](#_Toc42201442)

[Figure 7: PDF of Tukey lambda distribution (Tukey lambda distribution, 2019) 18](#_Toc42201443)

[Figure 8: Probability density function for the Laplace variate 19](#_Toc42201444)

[Figure 9: Probability density function for the rectangular variate 20](#_Toc42201445)

[Figure 10: Probability density function for the truncated normal variate 21](#_Toc42201446)

[Figure 11 Support vectors and margin representation 24](#_Toc42201447)

[Figure 12 Snippet from training data 26](#_Toc42201448)

Abstract

Normality tests are very important in statistical inference; their purpose is to know if the data is sampled from Normal population. The normality of the data is a prerequisite for several parametric statistics. There are many normality tests exist where can only be applied under certain conditions. The main goal of this research is to use the machine leaning techniques to create a model based test that could be of a good quality compared to the existing tests. This research tries to build a classification model using features of the data such as the moments, and it compares the quality of the model with the other tests according to the power. The data will be simulated from alternative distributions using Monte Carlo simulation.

ملخص

تعد اختبارات التوزيع الطبيعي (Normal distributions tests)مهمة للغاية في الاستدلال الاحصائي، والغرض منها هو معرفة ما إذا كانت البيانات مأخوذه من مجتمع توزيعه يتبع للتوزيع الطبيعي. التوزيع الطبيعي للبيانات هو شرط أساسي لعدة إحصاءات حدوديه. توجد العديد من االاختبارات التي تستخدم لهذا الغرض ولكن فعاليتها مشروطة على ظروف عدة للعينة مثل حجم العينة. الهدف الرئيس من هذا البحث هو استخدام تقنيات تعلم الالة لبناء نموذج يمكن أن يكون ذا جوده جيدة مقارنة باالاختبارات الحالية. يحاول هذا البحث إنشاء نموذج تصنيف باستخدام صفات عدة للبيانات مثل (moments) ، ويقارن جودة هذا النموذج مع الاختبارات الاخرى وفقًا لقوة الأختبار (Test power). سيتم خلق البيانات لهذا البحث من توزيعات مختلفة باستخدام محاكاة . MonteCarlo

Chapter One

Introduction

The normal distribution is an underlying assumption of many statistical procedures. Parametric tests such as correlation, regression, t tests, and analysis of variance are based on the assumption that the data follows a normal distribution. When the assumption does not hold, it is hard to draw accurate and reliable conclusions about the data (Ghasemi & Zahediasl, 2012). Visual plots such as P-P plot and statistical tests such as Shapiro-Wilk, Chi-square, D’Agostino-Pearson, Jarque-Bera, and others are the classical methods usually used to detect non-normality (Das & Imon, 2016).

Some of the existing normality tests can only be applied under certain conditions. For example, Shapiro-Wilk test has limitation on the size of the sample where it does not perform well on samples with size more than 50 (Shapiro, Wilk, & Chen, 1968). Moreover, different tests of normality often produce different results[[1]](#footnote-1). The contradicting results are misleading and often confuse statisticians.

In this research we try to leverage the power of machine learning techniques to build a new test that could be with comparable performance with the existing tests. The machine learning offers us the ability to build a model that learns from past experience. By providing examples of normal (negative) and non-normal (positive) examples, the model can learn the characteristics of each of these classes to a level that it can classify correctly more examples to the correct normality class.

* 1. Background

Normal distribution, also known as Gaussian distribution is one type of continues probability distributions. It appears as a bell curve (Figure 1) where it is symmetric about its mean, which is identical to its mode and median. 68%, 95%, and 99% of the data falls within 1, 2, and 3 standard deviations respectively (Patel & Read, 1996).

(Forbes, Evans, Hasting, & Peacock, 2011) The normal distributions has the following density function, usually noted as:

Where is the mean, and is the standard deviation. Figure 1 shows the p.d.f of the distribution of multiple examples of and.

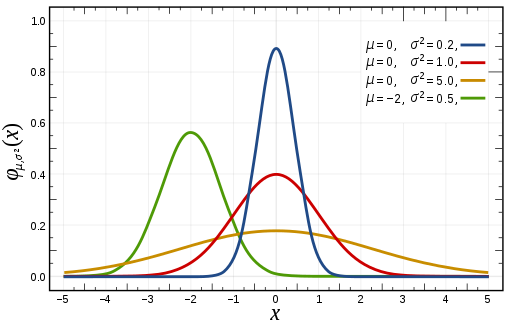


Figure : PDF of Normal distribution (Normal distribution, 2020)

Normal distribution is special as its two parameters ( and) are mutually independent and provide us with complete information of the shape and location of the distribution (Casella & Berger, 2001). The independence of the two parameters characterize the normal distribution from other distributions (Lukas, 1942). Normal distribution is unimodal and it has two inflection points located 1 standard deviation from the mean (Patel & Read, 1996).

If , then the random variable has a N(0,1) distribution, known as standard normal distribution and it is described by p.d.f

This function is symmetric around, where it attains its maximum value and has inflection points at and (Casella & Berger, 2001).

Normal distributions is the most importantly used in natural and social sciences to represent random variables. Quantities such as examination grades, snowflakes sizes, and other phenomena are approximated by of number of normal probability density functions (Lyon, 2014). The importance is mainly due to the central limit theorem, which states that the sum of independent and identically distributed random variables converges to normal distribution as number of samples increases regardless of the type of distribution of the sampled variables. This theorem provide theoretical bases for why so many variables we see in the nature appear to have approximately a normal probability distribution (Hazewinkel, 1994).

Normality tests are used to determine if the data is sampled from Normal distribution. The normality of the data is an assumption need to be verified before applying several parametric statistics such as t-test, linear regression analysis, discernment analysis and analysis of Variance (ANOVA). When the assumption is violated, the accuracy of the conclusions about the data is questionable and not reliable (Ghasemi & Zahediasl, 2012).

The normality test assess the likelihood that a given data set {} comes from normal distribution. The null hypothesis is that the observations are distributed normally versus the alternative that the observations are not distributed normally. There are two set of methods can be used to examine the normality, visual methods and statistical tests methods (Ghasemi & Zahediasl, 2012).

Visual plots such as P-P plot are useful to visualize the distribution of the data but they usually not enough to conclude decisions about the normality of the data. Hence, variety of statistical tests have been developed in this area such as Shapiro-Wilk, Anderson Darling, Kolmogorov-Smirnov tests and others. These tests are parametric tests aim to measure the probability of departure from normality for the data set on different significant levels.

* 1. Problem definition

The departure from normality is very critical in statistical inference. Biased interpretation can be inferred if the normality assumption is violated. Normality tests have traditionally been designed as classical statistical hypothesis testing procedures and, to the best of our knowledge, this has been the only way used so far to find departure from normality.

The long list of tests developed in the literature can make it hard for statisticians to select the appropriate test to use[[2]](#footnote-2). Moreover, these statistical tests are sensitive to the size of the data as shown in the study of (Oztuna, Elhan, & Tuccar, 2006).

In this research, we are proposing a new approach of normality test. In this approach we use the Machine Learning tools to develop a classification model that can classify the sample data to the correct underlying distribution with less sensitivity to the nature of the underlying distribution of the data.

* 1. Research Objectives

In this research, we propose a new approach of testing the normality of the data using Machine Leaning (ML). Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. This approach is known as supervised machine learning. Classification is one type of supervised machine learning where the human provides the algorithm with pairs of inputs and desired outputs, and the algorithm learn a general rule to produce the desired output given an input it has never seen before (Mueller & Guido, S, 2016).

The idea of using the machine learning in testing the normality was not explored in previous literatures we read as of the date of writing this paper. In this research we build a model that learns the properties and the characteristics of both the normal and the alternative distributions by providing examples of both classes. We expect the model to get enough experience to be able to correctly classify the normality of the data regardless the sample size and the underlying distribution. One advantage of this approach comparing to the classical tests is that it can provide us with additional measures to the power of the test. The power measures the ability to detect one type of the classes – the non-normality- while in the classification models, additional quality metrics are available to measure the performance on detecting the two classes, such as Accuracy and F-Measure.

* 1. Limitations of the study

Results and conclusions from a Monte Carlo simulation comparing powers across various distributions are seriously limited in generalizability beyond those distributions. Generalizability of the results depends on the design and how much coverage of different probability distributions is included in the study. In chapter 3, we show a wide range of alternative distributions added to the scope of the research by which we feel this offers a greater potential for generalizing results comparing to the distributions used in previous studies.

Related to that, generalizability of the proposed model could be questionable; results and conclusions of any classification model is limited to the data set it trained with (Cai, et al., 2020). In chapter 3 we try to overcome this limitation by having enough representations of the distributions in the training and by building a model from set of features resilience to the change in the type of the distribution such as skewness and kurtosis.

Another limitation of this study is the choice of the power as the base measure to compare our model against other statistical tests. This comparison is limited to only one of the two sides of the quality of any classification model. The Power which is equivalent to the “Recall” in machine learning terminology, evaluates the performance of detecting the positive class –alternative class in our use case-, and do not evaluate how the model performs in detecting negative class -normal class in our use case. This is because the classical normality tests are statistical tests; if the test does not have an evidence to reject the null hypothesis (sample has normal distribution), it does not mean it accept it. This limitation prevents us from using other quality measures such as Accuracy and F-Measure to evaluate the quality of the classifier on both normal and alternative classes.

Chapter Two

Literature review

1. 1. Normality tests

Large number of methods and tests available to detect departure from normality where each test has its own characteristics and power. We can look for departure from normality using two ways: Visual methods of normal plots or significant tests (Ghasemi & Zahediasl, 2012).

* + 1. Visual tests

The researcher can validate the normality of the data using graphical methods such as P-P plot, Q-Q plot, histogram, box plot, or stem-and-leaf plot. These plots are useful to visualize the distribution of the data but they often do not provide reliable evidence about the normality of the data. The plots are subjective, a plot can be interpreted into different levels of “normality” by different people. Moreover, judging using these visual methods required enough statistical experience of the researcher in order to take a correct decision. These implies to use more formal and reliable tests (Yap & Sim, 2011).

* + 1. Statistical tests

The effort of developing normality tests was initiated by (Pearson, 1895) who used the skewness and kurtosis as indicators of departure from normality. The number of different tests for normality seems to be boundless. The researchers classified the tests by different ways. In this section we present the tests by classifying them into four main groups as following:

* **Empirical Distribution Function (EDF) tests**: These tests involve measuring the discrepancy between the cumulative distribution function of the normal distribution and the empirical distribution function of the sample (D’Agostino & Stephens, 1986). The most popular tests of this type: Kolmogorov-Smirnov (KS) test (1933), Cramer-von Mises (CVM) test, and Anderson-Darling (AD) test. The Anderson-Darling (AD) test(1974)is the recommend one in this family (D’Agostino & Stephens, 1986). KS test is high sensitive to extreme values, and it has low power and it should not be used in testing normality (Throde, 2002).
* **Moments tests**: These tests use the skewness and the kurtosis (the second and the third moments respectively) of the sample to calculate the test statistic (D’Agostino & Stephens, 1986). Popular tests are Jarque-Bera (JB) test (1975) and D’Agostino-Pearson Omnibus test (DP) (1973).
* **Regression and correlation tests**: The tests are based on the correlation between the empirical data and corresponding scores under normality (D’Agostino & Stephens, 1986). Shapiro-Wilk (SW) (1965) test is the popular one in this family. It has good power for sample sizes up to 50. For large samples, the computation of its test statistic is much complicated (Das & Imon, 2016). Other tests in this group are Shapiro-Francia (SF) test and Ryan-Joiner test
* **Chi-Squared test**: It is not recommended for continuous distributions as it computes the number of observations instead of the observations themselves when calculating the test statistic. Chi-Squared test should not be used (D’Agostino & Stephens, 1986).
  1. Previous comparisons

The literature shows many attempts to compare different normality tests trying to find the best performing one. Most of the comparisons are based on comparing the power of the tests on the alternative distributions using Monte Simulation on different alternatives with different sample size and level of significance. The results have a lot of variation.

(Shapiro, Wilk, & Chen, 1968) Indicates that SW (Shapiro and Wilk 1965) has the best power comparing to (statndard third moment), (standard fourth moment), Kolmogorov-Smirnov, Cramer-Von Mises, Weighted CM, Modified KS, chi-squared, and (Studentized range) on alternatives of sample size (10, 15, 20, 35, 50).

In (Muyombya, 2017) study that examined the power of the tests on large sample sizes, Kolmogorov-Smirnov was the most powerful normality test regardless of the nature of the distribution. Followed by Shapiro-Wilk, Shapiro-Francia, Anderson-Darling, Jaque-Bera, and D’Agostino-Pearson.

(Alizadeh & Arghami, 2011) Compared the power of several tests and concluded that Jaque-Bera is the most powerful test for symmetric distributions and Shapiro-Wilk is the most powerful for asymmetric distributions with support. It also reveal Kolmogorov-Smirnov and Shapiro-Wilk have best power for alternatives supported by

A study by (Islam, Normality Testing- A New Direction, 2011) compared tests for the purpose of ensuring the validity of the t-statistic used to assessing the significance of the regressors. It shows that Anderson-Darling is the best option comparing to Jarque-Bera, D'Agostino and Pearson, and Lilliefors (a modification of Kolmogorov-Samirnov test).

(Razali & Wah, 2011) Compared the power of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors, and Anderson-Darling. Shapiro-Wilk was the most powerful test then Anderson-Darling, Lilliefors, and Kolmogorov-Smirnov on both symmetric and asymmetric alternatives. This research also reveals that these tests have low power in small sample size (less than 30).

(Islam, Ranking of Normality Tests: An Appraisal through Skewed Alternative Space, 2019) Evaluated the performance of several tests by using a proposed stringency framework of comparing tests. The research compares Kolmogorov-Smirnov, Anderson-Darling, Jaque-Bera, Shapiro-Wilk, D’Agostino, Coin (COIN), Bonett and Seier test (Tw). And he recommends to use Tw test for slightly skewed, Anderson-Darling and Shapiro-Wilk for moderately skewed, and all except COIN and Tw for highly skewed alternatives.

(Afeez, 2018) Compared several tests on five classes of alternatives: Near Normal, Symmetric long-tailed, Symmetric short-tailed, Asymmetric long-tailed, and Asymmetric short-tailed. SW had good power in a wide range of alternatives comparing to Anderson-Darling, Cramer–von Mises, Jaque-Bera, Chi-Square tests. Jaque-Bera was poor for symmetric short tails, but it is appropriate for symmetric long-tailed distributions.

(Seier, 2002) Claimed that Tests based on skewness and kurtosis are not powerful against symmetric alternative distributions where the kurtosis is close to that of the normal distribution. These tests are more powerful when the alternative is more peaked than normal.

Some of the studies and investigations share similar results. For example Shaipro-Wilk was in a good rank in some of them, but it was not recommended in others. Having a clear answer to the best performing test seems a very complicated task.

* 1. Limitations of the statistical tests

The large number of comparisons with different results confuse the researcher on which normality test to apply where dozens of tests are available to use. Based on what we show from some of the previous literatures, no single test is uniformly more powerful than others.

Comparing the tests based on their power using simulation didn’t succeed having an answer on what is the best test to use, as each test has its area of strengths and weaknesses. The power of the tests depend critically on two factors: The alternative, which can’t be specified when doing the test and as we saw that the same test has different powers when applied on different distributions. The other factor is the sample size, which is critical as well since the normality tests will always reveal non-normality as the sample size grows. (Oztuna, Elhan, & Tuccar, 2006) Show that for small sample size, the normality tests have small power to reject the null hypothesis when it should be rejected. And for large sample size, the normality tests become much sensitive and the test can be significant even in case of a small deviation from normality.

Chapter Three

Methodology

In this research, we propose a new approach of testing normality using state of the art ML techniques. In this chapter, we will start explaining the different steps to be executed in order to build and evaluate the classification model. Then we describe the method that is used in comparing the quality of the “new test” against other popular statistical tests of normality.

1. 1. Alternative Distributions

Alternative distributions can be classified into five major families based on the distribution skewness and kurtosis: asymmetric long-tailed, asymmetric short-tailed, symmetric long-tailed, symmetric short-tailed, and close to normal (Shapiro, S. & Wilk, B. & Chen, J. 1968). The alternative distributions used in this study were selected from these families on different levels of parameters in order to cover a wide range of the data. Five instances from each family are chosen as shown on Table 1, and an overview of the corresponding probability distributions is provided later in this section. The alternatives will be used in the proposed model as positive examples, and also used in the later phase of comparing the power of the new test against other statistical tests.

Table : Alternative distributions used in the research

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Family** | **Alternatives** | | | | |
| **Asym. long-tailed** | Weibull(0.5, 1) | Weibull(2, 1) | LogNormal  (0, 1) |  |  |
| **Asym. short-tailed** | Beta(2, 1) | Beta(3, 2) | LogNormal  (0, 0.15) | LogNormal  (0, 0.25) | LogNormal  (0, 0.35) |
| **Sym. long-tailed** | t(1) | t(2) | t(4) | t(7) | Tukey (10) |
| **Sym. short-tailed** | Uniform(0,1) | Beta(1.3, 1.3) | Beta(1.5, 1.5) | Tukey(1.5) | Truncated normal (-2, 2) |
| **Close to normal** | Tukey (0.1) | Tukey (0.2) | Tukey (5) | t (10) | Laplace(0, 10) |

* + 1. Beta Distribution

(Walck, 2007) (Forbes, Evans, Hasting, & Peacock, 2011) The Beta distribution denoted by is a continuous distribution given by:

Where the quality is the Beta function defined in terms of Gamma function as:

For v = = 1, the Beta distribution simply becomes a uniform distribution between zero and one. The mean and the variance of the Beta distribution given by

Figure 2 shows the Beta distribution on different levels of and .

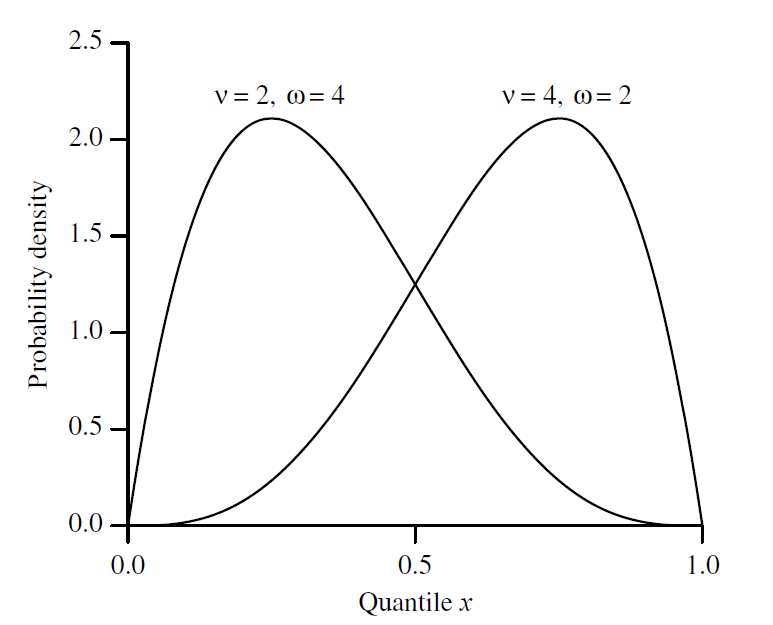


Figure : Probability density function for Beta variate **β**: v, ω

* + 1. Student t-distribution

(Walck, 2007) (Forbes, Evans, Hasting, & Peacock, 2011) The Student’s t-distribution (or simply the t-distribution) denoted by is given by

Where is the degrees of freedom and t is a real number. The functions Γ and β are the usual Gamma and Beta functions. The mean of t-distribution is 0 for , otherwise undefined. The variance is given by

Figure 3 shows the t distribution on different values of .

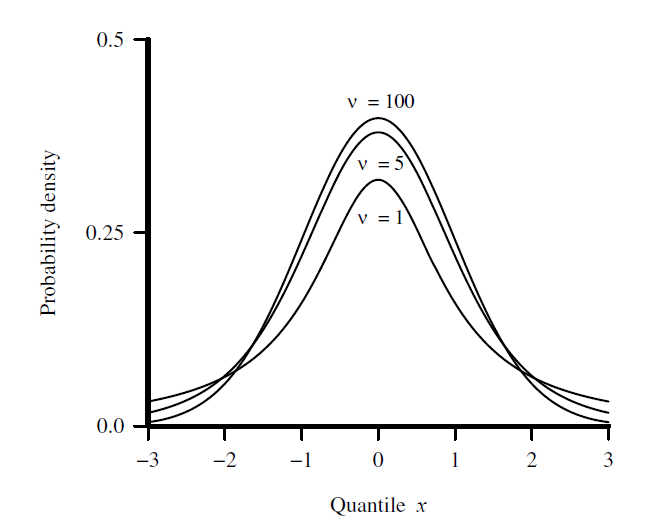


Figure : Probability density function for Student’s t variate, **t:**

* + 1. Chi-squared Distribution

(Walck, 2007) (Forbes, Evans, Hasting, & Peacock, 2011) The chi-squared distribution denoted by with degrees of freedom is the distribution of a sum of the squares of independent standard normal random variables. Where a set of data is represented by a theoretical model, the chi-squared distribution can be used to test the goodness of fit between the observed data points and the values predicted by the model, subject to the differences being normally distributed. It is given by

The mean of chi-squared distribution is equal to the degrees of freedom, and the variance is double the mean =. Figure 4 shows the distribution on different values of .

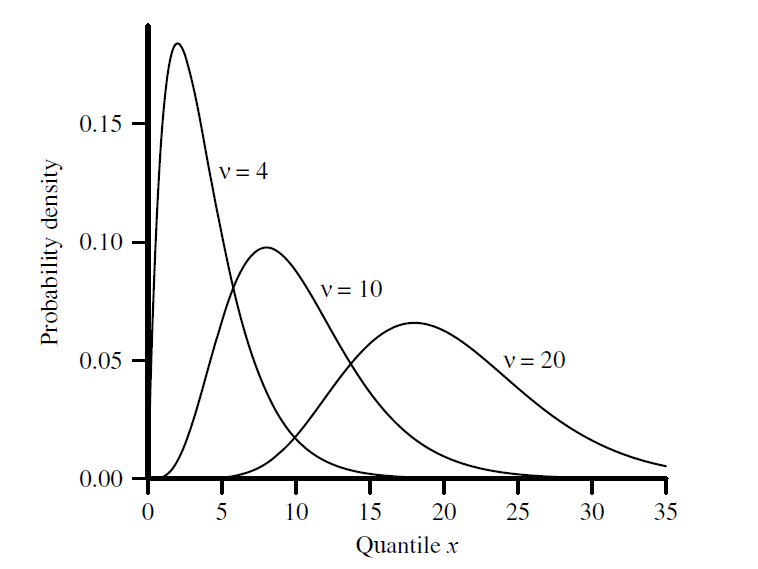


Figure : Probability density function for the Chi-Squared variate

* + 1. Log-normal Distribution

(Walck, 2007) (Forbes, Evans, Hasting, & Peacock, 2011) The Log-normal distribution denoted by is a continuous distribution of a random variable whose logarithm is normally distributed. The lognormal distribution is applicable to random variables that are constrained by zero but have a few very large values. The resulting distribution is asymmetrical and positively skewed. It is given by

An Alternative parameter of scale is where The mean and the variance of the Log-normal distribution are given by

Figure 5 shows the Log-normal distribution on different values and .

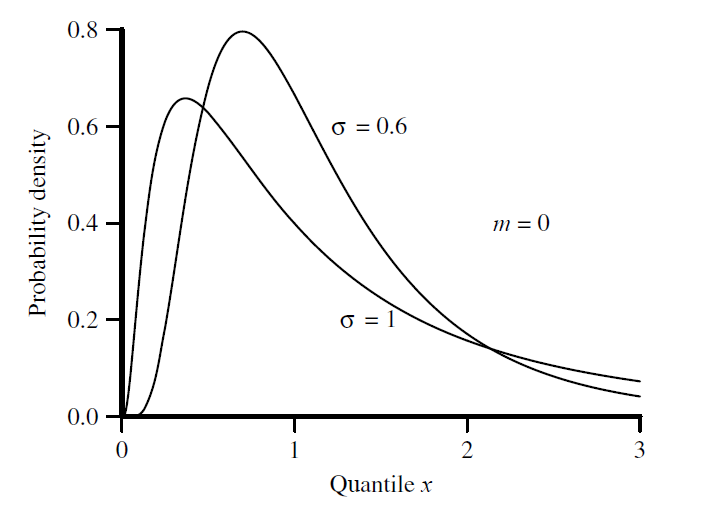


Figure : Probability density function for the Log-normal variate **L**:

* + 1. Weibull Distribution:

(Walck, 2007) (Forbes, Evans, Hasting, & Peacock, 2011) The Weibull distribution denoted by is named after the Swedish physicist Waloddi Weibull (1887-1979) who described it in details in 1951. Weibull variate is commonly used as a lifetime distribution in reliability applications. The two-parameter Weibull distribution can represent decreasing, constant, or increasing failure rates. The parameter is the shape parameter, and is simply a scale parameter and the variable has distribution

The Weibull distribution is given by

The mean and the variance of the Weibull distribution are given by

Figure 6 shows the Weibull distribution on different levels of and

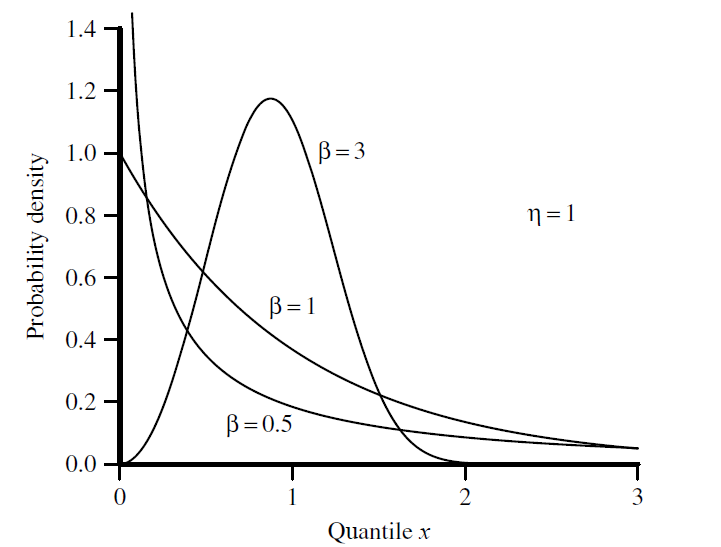


Figure : Probability density function for Weibull variate

* + 1. Tukey Distribution

(Stephanie, 2015) (Joiner & Rosenblatt, 1971) Tukey lambda distribution denoted by is a continuous symmetric probability distributed defined in terms of its quantile function, named after the American mathematician John Wilder Tukey (1915-2000). Unlike most other probability distributions, there isn’t a “one size fits all” formula for probability density function. It is defined in terms of quantiles where the quantile function (i.e. the inverse of the cumulative distribution function) and the quantile density function () are

Figure 7 shows the Tukey lambda distribution on different levels parameters

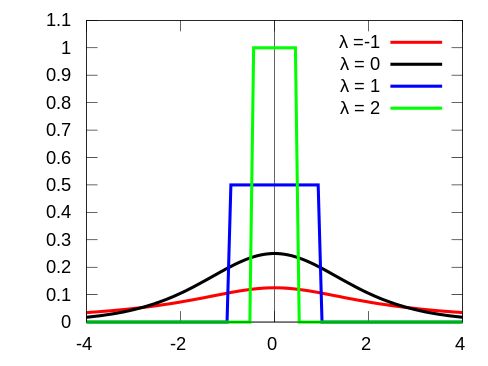


Figure : PDF of Tukey lambda distribution (Tukey lambda distribution, 2019)

* + 1. Laplace Distribution

(Walck, 2007) (Forbes, Evans, Hasting, & Peacock, 2011) Laplace distribution sometimes called double exponential distribution is a continuous probability distribution named after Pierre-Simon Laplace (1749-1827). It is a symmetric distribution whose tails fall off less sharply than the Gaussian distribution but faster than the Cauchy distribution. The distribution has an interesting feature as the best estimator for the mean µ is the median and not the sample mean. The distribution is given by

Where is the location parameter, and b > 0 is the scale parameter. The variance of the distribution is Figure 8 shows the Laplace distribution on different levels of and b

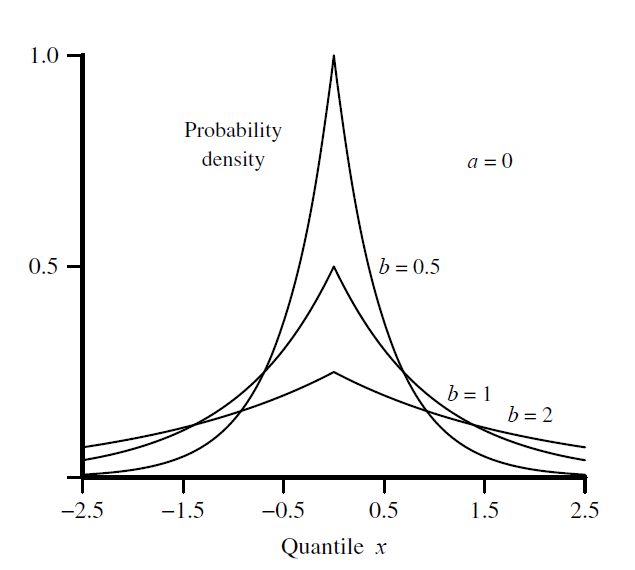


Figure : Probability density function for the Laplace variate

* + 1. Uniform (Rectangular) Distribution

(Walck, 2007) (Forbes, Evans, Hasting, & Peacock, 2011) Uniform distribution denoted by is a symmetric probability distribution defined by two parameters and where the location parameter is and is the scale parameter. It is widely used as the basis for the generation of random numbers for other statistical distributions. Where every value in the range of the distribution is equally likely to occur. This is the distribution taken by uniform random numbers. It is given by

The mean and the variance of the Uniform distribution are given by:

Figure 9 shows the p.d.f of the Uniform distribution.

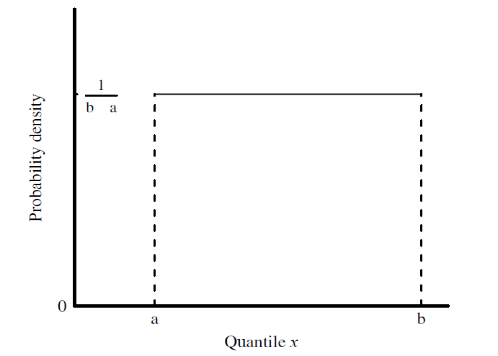


Figure : Probability density function for the rectangular variate

* + 1. Truncated Normal Distribution

(Burkardt, 2014)The truncated normal probability density function is defined in two steps. We choose a general normal PDF by specifying parameters µ and, and then a truncation range (a, b). The p.d.f associated with the general normal distribution is modified by setting values outside the range to zero, and uniformly scaling the values inside the range so that the total integral is 1. Suppose X has a normal distribution with mean and variance and lies within the interval with Then the p.d.f of X truncated on is given by:

Figure 10 shows the p.d.f of the Truncated normal distribution on different levels of and .

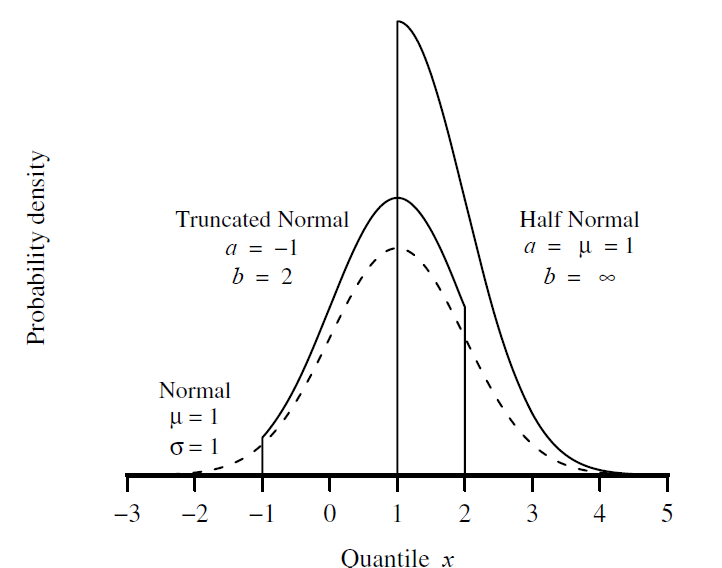


Figure : Probability density function for the truncated normal variate

* 1. Model construction

In this section we describe the process of building the classification model of testing the normality. We start by describing the normality test as classification problem. Then we describe the data that we will use for this purpose and the steps of training and evaluating the model.

* + 1. Process

This problem is a binary classification problem, we predict if the sample data has departure from normality based on some properties such as skewness and kurtosis. The target variable in this classification problem is the type of the distribution where “alternative” represent the positive class and “normal” represent the negative class. We did not choose the positive class to represent the normality – which could make more sense for others– because we need to compare the power of this model with other tests which tries to check if the sample is significantly depart from normality and not vice versa.

In the process of creating the model, we are following the steps of Train-Validate-Test. In the training phase we train a model using training data from positive and negative classes. Where in the validate stage we run the model on the validate data set and tune the model parameters to yield the best quality that can be achieved by the model. In this stage we set in the model the threshold (cut off) points that have optimal quality. The tuned model is then tested on different test sets and the quality of this test represents the final quality of the model.

* + 1. Classification Techniques

In this research we aim to find the best classification technique that has the best performance in our use case, so we tend to build models using several methods and choose the one with best quality. There are plenty of classification algorithms available to use. For example, [caret](http://topepo.github.io/caret/index.html) (Classification And REgression Training) package in R has more than 180 classification techniques from different families. It is hard examine all of the techniques to find the best one that fits our data, and it is not straight forward to select from this long list. So, in this research I refer to previous studies that compared these techniques and evaluated their quality on several data sets.

(Fernandez-Delgado, Cernadas, Barro, & Amorim, 2014) Compared 179 classifiers from 17 families in 121 data sets, and (Wainer, 2016) compared 14 techniques on 115 binary datasets. The two studies show that Random Forest (RF), Gradient Boosting Machines (GBM), and Support Vector Machines (SVM) classifiers are the most performing ones and they are not significantly different from each other. As a result, we will build three models from these classifier types and compare their quality as part of this research. In the sub sections below, we are providing a brief explanation of these classifiers.

* + - 1. Random forest (RF)

(Rebala, Ravi, & Churiwala, 2019) Random forest is an effective model for both classification and regression problems. In classification learning, it is an ensemble classifier constructed from a collection of decision trees that improve the prediction over single decision tree. The data set is split randomly with replacement into different bags – this is called data bagging- each one represents a decision tree. For each bag, different set of features with size or is selected from feature set n and cross validation is used to select which feature set is most appropriate for this specific bag – this is called Feature Bagging. So each data bag will have a different set of features chosen for creating the Decision Tree.

For a new data point, prediction of its class is the aggregation of the predicted classes from all trees. The final result is aggregated using general voting technique as shown in below equation:

Where:

= the final prediction from RF

= number of trees

= the index of decision tree

= the result from decision tree

= the vector of the new data point to predict

Random Forest enables us to see which features are important of the variable to the decision. The intuitive notion in determining the variable importance is that if the variable is important, then rearranging the values of the variable in constructing the trees will not reduce the prediction accuracy.

* + - 1. Gradient boosting machines (GBM)

(Ayyadevara, 2018) Gradient boosting is a machine learning classification technique based on creating an ensemble model from different models built sequential as following. It start by creating an initial model using tree or linear regression that fit the data. The second model is built and its objective is to accurately predicting the cases where the first model performs poorly. The combination of these two models should have higher performance than either model alone. This booting process repeated many times until reaching the minimum prediction error. Gradient refers to the error, or residual, obtained after building a model. Boosting refers to improving. The technique is known as gradient boosting machine, or GBM. Gradient boosting is a way to gradually improve (reduce) error.

* + - 1. Support vector machines with Radial Basis Function Kernel (RBF SVM)

(Rebala, Ravi, & Churiwala, 2019) Support vector machines (SVM) is a binary classifier, it classifies the data points by creating the optimal hyperplane boundary that have the maximum margin for the data points as shown in Figure 11. SVM can handle linear separable data points as shown in the previous figure and can handle data points that are not linear separable by mapping data points into higher dimensional space using “kernel” functions. SVM classifier create a hyperplane of N-1 dimensions for n-dimensional feature vectors to separate the data into two classes. For example, for feature vector of size 2 the hyperplane is a line and can be represented by the following equation:

Where:

= the feature vector

= the weight assigned to feature vector

= the bias term

All values of y greater than the function value are classified as class 1, and all other values are classified as class 2.

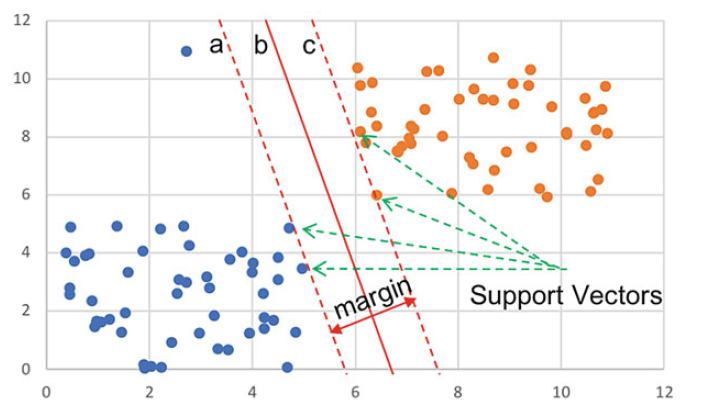


Figure Support vectors and margin representation\

Many kernels exist such as Polynomial and sigmoid functions. It is not obvious which kernel works best. Radial bias kernel (RBF) is one of the commonly used kernel in SVM and can be represented by below equation:

* + 1. Data set

First stage of building a classification model is to prepare a data set for training and testing. We used simulation code to generate samples from alternative and normal distributions, where each generated sample represents a data point in our data set. Several statistics and features calculated on each sample, where the features and the sample underlying distribution represent the data point. The samples were scaled before calculating the features in order to improve the scalability of the model and avoid biasedness toward specific set of sample sizes or distributions.

Data set of 10,000 data points were generated, each data point represent a sample of size n generated from normal or alternative distribution. 50 % of the data points are simulated from alternative distributions and the other 50% are simulated from normal distribution. The data intended to have 1:1 ratio between positives and negatives aiming to a balanced data sets.

The positive labels generated from the alternatives are listed in Table 1. The negative labels generated from the normal distribution on different levels of mean and standard deviation. Both set of labels generated from 200 different sample sizes randomly selected from the range of [5, 2000]. For each size n in the range, 25 samples created from the alternative distributions and another 25 samples created from normal distribution. The negative 25 labels on each size generated as following: Five means were randomly selected from the range [-1000, 1000]. For each mean, five samples generated from normal distribution with coefficient of variation equals to 0.01, 0.1, 0.3, 0.6, and 1.0. Using different levels of variation aims to train the model on representative data set to decrease the biasedness to specific distributions.

We will exclude from the data set data points of one from each of alternative and normal distributions. We call this data set as “Unseen” data. The remaining data points of the other four types of both alternative and normal distributions represent the data set that will be used in training and testing, they will be randomly divided into 60% train, 20 validate, 20% test. The reason of having the unseen data is to test the final model on distributions that the model didn’t see before, to validate the scalability of the model and its ability to generalize to new data. We predict how our model will perform on other distributions not included in this research.

* + 1. Training

Different models will be trained using the three techniques: Random Forest (RF), Gradient Descent Boosting (GBM), and Support vector machines with Radial Basis Function Kernel (RBF SVM). Each model will be evaluated and the best performing one will be considered and used in later stage of comparing against statistical normality tests. The features of the model are calculated from each sample and saved in csv format. The features to be used in the model are properties of the sample data such as mean, median, variance, skewness, kurtosis, sample size, outliers’ ratio. This is an initial set of features we can start with to build a baseline. Other features probably will be added during the time of building the model. Figure 11 shows a snippet of some data points we have in the training data.

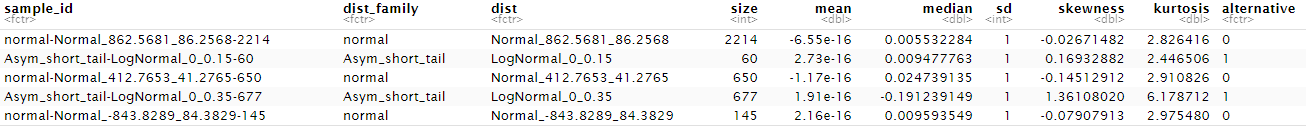


Figure Snippet from training data

Feature selection techniques could be applied on the model to find the most significant features and drop the non-important ones. Techniques such as Feature Importance of Random Forest, Recursive Feature Elimination (RFE), and ANOVA F-test could be used in this study. The goal is to keep the model with minimal set of features that gives the highest possible quality.

* + 1. Evaluation

Several metrics available to use for evaluating the quality of a classification model. We prefer to use the Accuracy measure in this problem more than other measures like F-Measure. The Accuracy represents the combination of Specificity (1- α) and Sensitivity (Power) which are the measures we will use in comparing the quality of the model with other statistical tests.

Validation data set will be used to evaluate quality of the models from several classification techniques. The models will be tuned by applying different model parameters such as number of nodes in decision tree classifier and number of neighbors in KNN classifier. The model with the best performance to be chosen for next steps.

The selected model will be evaluated on the test set and on the unseen data sets. Different quality measures and charts to be used to report and analyze the performance of the proposed test.

* 1. Power comparison test

A power comparison test to be concluded between different normality tests including the new proposed model using Monte Carlo simulation. The alternative distributions considered are the ones listed in Table 1. The comparison will be on two levels of significance α = 0.05 and α=0.10 to investigate the effect of the significance level on the power of the test. Corresponding thresholds of the proposed test on each level of significance can be calculated by choosing the two thresholds that give specificity of 0.95 and 0.90 for 0.05 and 0.10 level of significance respectively. Samples of size n = 10, 20, 30, 50, 100, 200, 500, 1000, 1500, and 2000 will be used in the simulation from each alternative with 1,000 repetitions.

* 1. Toolbox

We will use R as the main programming language in this research. It offers to data scientists and statisticians a vast toolbox and libraries for data loading, modeling, visualization, and analysis. RStudio with R 3.6.2 is used. We use [caret](http://topepo.github.io/caret/index.html) package to build and evaluate the classification models as it provides the data scientists with simple interface for executing many classifiers with automatic parameter tuning for the models. This enables the researcher to use the state of the art classification techniques with minimal knowledge of the underlying algorithms (Kuhn, 2008). We will use caret package to train and tune the models, feature selection, and variable importance estimation. [MonteCarlo](https://cran.r-project.org/web/packages/MonteCarlo/vignettes/MonteCarlo-Vignette.html) library will be used to simulate the power of the model and the statistical normality tests.

Chapter Four

Simulation and Results

1. 1. Classification model
      1. Examining Data
      2. Features
      3. Training code (caret, control, code snippet…)
      4. Results
         1. Test set
         2. Unseen data
         3. Feature importance
      5. Error analysis
   2. Power comparison
      1. Model thresholds
      2. Results

References

Afeez, B. (2018). Selection and Validation of Comparative Study of Normality Test. *American Journal of Mathematics and Statistics, 8*(6), 190-201.

Alizadeh, H., & Arghami, N. (2011). Monte Carlo comparison of seven normality tests. *Journal of Statistical Computation and Simulation, 81*(8), 965-972.

Ayyadevara, V. K. (2018). *Pro Machine Learning Algorithms.* Apress. doi:10.1007/978-1-4842-3564-5

Burkardt, J. (2014, October 17). The Truncated Normal Distribution. Retrieved May 9, 2020, from http://people.sc.fsu.edu/∼jburkardt/presentations/truncated normal.pdf

Cai, X.‐L., Xie, D.‐J., Madsen, K., Wang, Y., Bogemann, S., Cheung, E., . . . Chan, R. (2020). Generalizability of machine learning for classification of schizophrenia based on resting‐state functional MRI data. *Human Brain Mapping, 41*(1), 172-184. doi:https://doi.org/10.1002/hbm.24797

Casella, L., & Berger, R. (2001). *Statistical inference* (2nd ed.). Cengage Learning.

D’Agostino, B., & Stephens, A. (1986). *Goodness-of-fit Techniques.* Marcel Dekker.

Das, R., & Imon, M. (2016). A Brief Review of Tests for Normality. *American Journal of Theoretical and Applied Statistics, 5*(1), 5-12. doi:10.0.45.128/j.ajtas.20160501.12

Fernandez-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we Need Hundreds of Classifiers to Solve Real WorldClassification Problems? *Journal of Machine Learning Research*(15), 3133-3181.

Forbes, C., Evans, M., Hasting, N., & Peacock, B. (2011). *Statistical Distributions* (4th ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.

Ghasemi , A., & Zahediasl, S. (2012). Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. *International Journal of Endocrinology and Metabolism, 10*(2), 486-489. doi:10.5812/ijem.3505

Hazewinkel, M. (1994). *Encyclopedia of Mathematics.* Springer.

Islam, T. (2011). Normality Testing- A New Direction. *International Journal of Business and Social Sciences, 2*, 115-118.

Islam, T. (2019). Ranking of Normality Tests: An Appraisal through Skewed Alternative Space. *Symmetry, 11*(827). doi:10.20944/preprints201905.0190.v1

Joiner, B., & Rosenblatt, J. (1971). Some Properties of the Range in Samples from Tukey's Symmetric Lambda Distributions. *Journal of the American Statistical Association, 66*(334), 394–399.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, 28*(5), 1-26.

Lukas, E. (1942). A characterization of the normal distribution. *Annals of Mathematical statistics, 13*, 91-83.

Lyon, A. (2014). Why are Normal Distributions Normal. *The British Journal for The Philosophy of Science, 65*, 621-649.

Mueller, C., & Guido, S, S. (2016). *Introduction to Machine Learning with Python.* O’Reilly Media.

Muyombya, M. (2017). On empirical power of univariate normality tests under symmetric, asymmetric and scaled distributions. *International Journal of Scientific and Engineering Research, 8*(3).

*Normal distribution*. (2020, April 29). Retrieved May 9, 2020, from Wikipedia: https://en.wikipedia.org/w/index.php?title=Normal\_distribution&oldid=953810074

Oztuna, D., Elhan, A., & Tuccar, E. (2006). Investigation of four different normality tests in terms of type 1 error rate and power under different distributions. *Turkish Journal of Medical Sciences, 36*(3), 171-176.

Patel, K., & Read, B. (1996). *Handbook of the Normal Distribution* (2nd ed.). CRC Press.

Pearson, K. (1895). Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous. *Philosophical Transactions of the Royal Society of London. A, 186*, 343-414. Retrieved from https://www.jstor.org/stable/90649

Razali, N., & Wah, Y. (2011). Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests. *Journal of Statistical Modeling and Analytics, 2*(1), 21-33.

Rebala, G., Ravi, A., & Churiwala, S. (2019). *An introduction to machine learning.* Springer. doi:10.1007/978-3-030-15729-6

Seier, E. (2002). Comparison of tests of univariate normality. *InterStat Statistical Journal, 1*, 1-17.

Shapiro, S., Wilk, B., & Chen, J. (1968). Acomparison study. *Journal of American, 63*(324), 1343-1372.

Stephanie. (2015, August 6). *Tukey Lambda Distribution: Definition*. Retrieved May 9, 2020, from Statistics How To: https://www.statisticshowto.com/tukey-lambda-distribution/

Throde, C. (2002). *Testing For Normality.* Marcel Dekker.

*Tukey lambda distribution*. (2019, June 6). Retrieved May 9, 2020, from Wikipedia: https://en.wikipedia.org/w/index.php?title=Tukey\_lambda\_distribution&oldid=900491544

Wainer, J. (2016). Comparison of 14 different families of classification.

Walck, C. (2007, September 10). Hand-book on statistical distributions for expirements. University of Stockholms. Retrieved May 9, 2020, from http://www.stat.rice.edu/~dobelman/textfiles/DistributionsHandbook.pdf

Yap, W., & Sim, H. (2011). Comparison of various types of normality tests. *Journal of Statistical Computation and Simulation, 81*(12), 2141-2155. doi:10.1080/00949655.2010.520163

1. Previous comparisons between the normality test described in section 2.2 [↑](#footnote-ref-1)
2. The tests and their details are explained in section 2.1 in this document [↑](#footnote-ref-2)