Faculty of Graduate Studies
Mathematics Department

Power Comparison of Some Goodness of Fit Tests

By

Ayah Ghayyadah

Supervisor

Dr. Bader Aljawadi

This Thesis is Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Mathematics, Faculty of Graduate Studies, Hebron University, Hebron, Palestine.

2022

Power Comparison of Some Goodness of Fit Tests

By

Ayah Ghayyadah

This thesis was defended successfully on ..... and approved by:

Committee Members:		**Signature**
• Dr. Bader Aljawadi	Supervisor	.....
• Dr. Inad Nawajah	Internal Examiner	.....
• Dr. Tareq Atyani	External Examiner	.....

DECLARATION

I declare and confirm that my thesis, titled "Power Comparison of Some Goodness of Fit Tests", has been written without applying to any assistance inconsistent with scientific ethics and traditions. I declare that all content and ideas drawn directly or indirectly from external sources are indicated in the text and listed in the list of references.

Dedications

I dedicate my thesis to my parents, husband and sisters. The completion of this thesis would not be possible without their support and encouragement.

ACKNOWLEDGEMENT

In The Name of ALLAH, the Most Merciful and Most Beneficent All praise do to Allah, Lord of the universe. Only by His grace and mercy this thesis can be completed. The completion of this thesis would have been impossible if not for the assistance and direct involvement of so many kindhearted individuals. First and foremost, I am very grateful to my supervisor Dr. Bader Aljawadi, for his strong support, guidance, and patience for the very enriching and thought provoking discussions which helped to shape the thesis. I am also indebted to the staff of the Mathematical department at Hebron University for their help and cooperation. I wish to express my deepest gratitude to my parents, brothers and sisters for their prayers, continuous moral support and unending encouragement. Last but not least, I wish especially to acknowledge my beloved husband for his love, support, patience and understanding.

Abstract

There are some common goodness of fit tests that have been studied by researchers over the years such as the Shapiro-Wilk test, Anderson Darling test, Chi-square test and Bickel-Rosenblatt test. Researchers often use the goodness of fit test to decide if an underlying population distribution differs from a specific distribution. The main purpose of this thesis is to compare the power of some common goodness of fit tests, where a comparison of the proposed goodness of tests is conducted using the simulation method of sample data generated from some common distributions; R software was used to generate data by applying Monte Carlo simulation. The power of the tests generally affected by some factors like sample size and the type of distribution being tested in, however, the critical values are used for power comparisons that are obtained based on 10000 simulated samples from different distributions. The power of each test was then obtained by comparing the respective critical values with the goodness of fit test statistics. The main results based on the simulation study indicate that the Anderson Darling test has the highest power in the case of testing symmetric distributions when the data is generated from parametric alternative distributions, while the χ^2 test has the lowest power. Furthermore, the Bickel-Rosenblatt test has the highest power in the case of testing symmetric distributions and the Anderson Darling test has the highest power under other non-parametric alternative distributions. This study also shows that when the Epanechnikov kernel is employed, the Bickel- Rosenblatt test has the highest power compared to the uniform kernel.

Contents

Chapter 1

INTRODUCTION

1.1 Introduction

The goodness of fit test is a statistical hypothesis used to judge whether or not a sample of data fit a specific distribution. In other words, it reveals the discrepancy between the observed value and the expected value in a given model. It's widely used in the analysis of health and medicine related survey data. In this thesis, the main goal is to examine the power of some common goodness of fit tests [7].

The normal distribution is very important since it is being assumed for many statistical procedures such as t-test, linear regression analysis and analysis of variance (ANOVA). When the normality can't be assumed, interpretation and inferences may not be valid or reliable [10][15].

There are three common ways to find out if a random sample of independent observations of size n come from a population that is normality distributed [8]:

1)graphical methods (Histogram, Q-Q plots, and stem and leaf plot: this method still not sufficient to provide conclusive evidence that the normal assumption holds.

2) numerical methods (skewness and kurtosis indices).

3) formal normality tests (Goodness of fit tests).

There are almost 40 tests of normality available in the literature but in this thesis we are interested in five of them: Chi-square test, Kolmogorov-Smirnov test, Anderson darling test, Shapiro-Wilk test and Bickel Rosenblatt test.
The following section will discuss the histogram in general and some of its drawbacks. The third section will talk about kernel density estimation, example and some of its proprieties.

1.2 Histogram

A histogram is a graphical representation of the distribution of the data. It was first introduced by Karl Pearson in 1895. It works by indicating the number of data points that lie within a range of values. These ranges of values are called bins. The frequency of the data that falls in each class is portrayed by the use of a bar. The higher that bar is, the greater the frequency of data values in bins.
Histogram is one of the graphical methods used in assessing whether a sample of n independent observations coming from a normal distribution. However, histogram has drawbacks, it's original sin is data binning which depraves the data of their individual location replacing their location with interval location, also its shape depends mostly on the subjective choice of the number of bins to which the range of a sample is divided and on the choice of the initial point. Therefor, the histogram doesn't introduce decisive evidence that the normal assumption holds and requires another method like the formal normality tests to support [13][6].

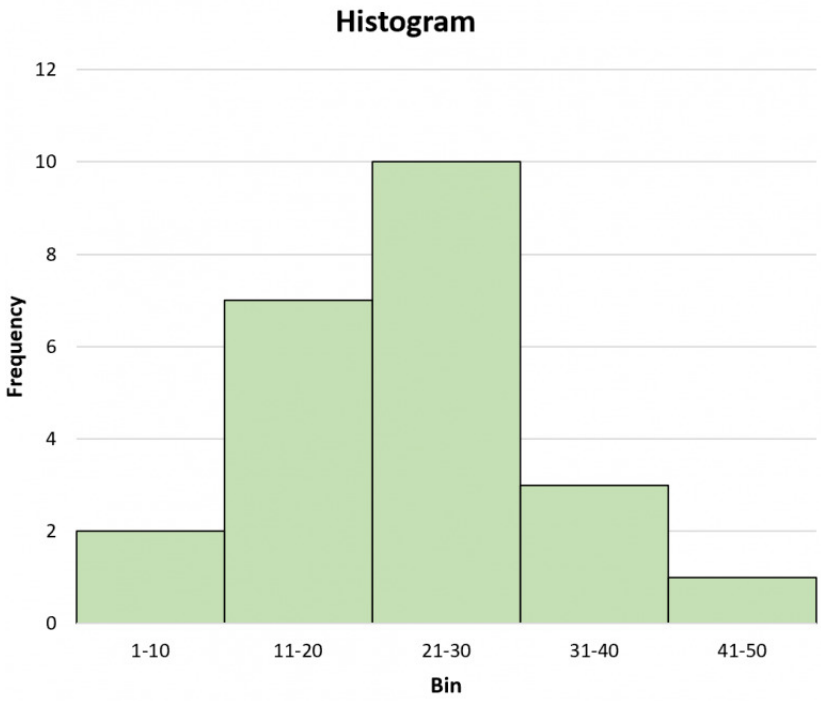

Figure 1.1: Example of Histogram Graph.

1.3 Traditional Goodness of Fit Tests

There are different goodness of fit tests available in the literature. The most common is the Chi-square test which is also known as Pearson's Chi-square test. It's widely used in many cases because it can be applied to any univariate distribution and calculated much easier than other tests. However, when the sample size is small, the performance of the Chi-square test is not satisfactory [7].

The Kolmogorov-Smirnov goodness of fit test used to compare a data with an known distribution. It's a nonparametric test that proposed by Kolmogorov and Smirnov (1933 and 1939). It's a comparison between some theoretical cumulative distribution function and a sample cumulative distribution function such that the sample is randomly selected from a population with unknown distribution function.

Another goodness of fit test is Anderson Darling test which is used to test if a sample of data came from a population with a specific distribution. It's a modification of the K-S test but it gives more weight to the tails than does K-S test and it make use of the specified distribution in calculating the critical values. Anderson darling test has the advantage of allowing a more sensitive test than the K-S test and the disadvantage that critical values must be calculated for each distribution [10][11].

Shapiro Wilk test was proposed by Samuel Sanford and Martin Wilk in 1965. It is used to test the null hypothesis that a sample of data come from a normally distributed population. It compares the observed cumulative frequency curve with the expected cumulative frequency curve and it was originally restricted to a sample size of less than 50 but now it has become the preferred test because of its good power properties. The W test statistic is the ratio of the best estimator of the variance to the usually corrected sum of squares estimator of the variance [8].

Bickel-Rosenblatt test was firstly proposed in 1973 by Bickel and Rosenblatt .it is used to test whether a given sample comes from a population with a hypothesized distribution. Even though the emprical distribution function (edf) tests are more popular than this test but it has more power, especially in directions where edf tests are not effective [5].

1.4 Kernel Density Estimation

Kernel Density estimation is a statistical tool that is used to create a smooth curve given on a set of data and using all sample points in order to estimate a probability density function. It was firstly introduced by Rozenblatt's (1956) and Parzen's (1962). Kernel density estimation is a function defined as the sum of a kernel function on every data point. It's often shorted as KDE and it has some application in fields such as signal processing and

econometric.

Definition 1.1. Let the series $x_1, x_2, ..., x_n$ be an independent and identically distributed (iid) sample of observations taken from a population X with an unknown probability density function $f(x)$. Kernel estimate $\hat{f}(x)$ of original $f(x)$ assigns each i^{th} sample data point x_i with a function $K(x_i, t)$ called a kernel function in the following way [12]:

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K(x_i, t) \tag{1.1}$$

$K(x_i, t)$ is non-negative and bounded for all x and t .

$$0 \leq k(x, t) < \infty, \forall t, x \in \mathbb{R} \tag{1.2}$$

and for all real x ,

$$\int_{-\infty}^{+\infty} k(x, t)dx = 1, \tag{1.3}$$

and that ensure the normalization of kernel density estimate

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{+\infty} k(x, t)dx = 1 \tag{1.4}$$

In other words, the kernel transforms the point location of x_i into an interval centered around x_i.

In most common practical applications, the kernel estimation uses symmetric kernel function, although asymmetric function have recently been increasingly used.

Now, the symmetry property allows writing the kernel function in a form used most frequently [13]:

$$K_s(x,t) = \frac{1}{h}K(\frac{x-t}{h}) \tag{1.5}$$

Where parameter h is called the smoothing parameter or bandwidth and it governs the amount of smoothing applied to the sample.

The kernel function's shape doesn't have much effect on the shape of the estimator even though it sometimes shows some difference in the density estimator. Whereas, the effect of the smoothing parameter h is big : a small value of the smoothing parameter may have too much detail, while too large value of h causes over smoothing of the information contained in the sample.

Examples of kernel function [13] :

There is many type of kernel function can be found in the relevant literature , such as :

Uniform $$K(u) = \begin{cases} \frac{1}{2} & \text{for } |u| \leqslant 1 \\ 0 & \text{for } |u| > 1 \end{cases}$$

Gaussian $$K(u) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}u^2}$$

Epanechnikov $$K(u) = \begin{cases} \frac{3}{4\sqrt{5}}(1 - \frac{1}{5}t^2) & \text{for } |u| < \sqrt{5} \\ 0 & \text{for } |u| \geqslant \sqrt{5} \end{cases}$$

Triangular $$K(u) = \begin{cases} 1 - |u| & \text{for } |u| < 1 \\ 0 & \text{for } |u| \geqslant 1 \end{cases}$$

Biweight $$K(u) = \begin{cases} \frac{15}{16}(1 - u^2)^2 & \text{for } |u| < 1 \\ 0 & \text{for } |u| \geqslant 1 \end{cases}$$

1.4.1 Properties of the Kernel density estimator

In order to know how much the estimator $\hat{f}(x)$ is close to the original function $f(x)$, we are discussing the method of mean square error (MSE) and it's two components, bias and variance. So, we have [16][6]:

$$\begin{aligned} MSE(\hat{f}(x)) &= E(\hat{f}(x) - f(x))^2 \\ &= [E\hat{f}(x) - f(x)]^2 + E[\hat{f}(x) - E\hat{f}(x)]^2 \\ &= Bias^2(\hat{f}(x)) + Var(\hat{f}(x)) \end{aligned} \tag{1.6}$$

1.4.2 The bias, variance and mean squared error of $\hat{f}(x)$

Bias: We first analyze the bias, the bias of KDE is [16]

$$\begin{aligned} E(\hat{f}(x)) - f(x) &= E(\frac{1}{nh}\sum_{i=1}^{n} k(\frac{x - x_i}{h}) - f(x)) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h}\int_{-\infty}^{+\infty} K(\frac{x-t}{h}) f(t)dt - f(x) \\ &= \frac{1}{h}\int_{-\infty}^{+\infty} K(\frac{x-t}{h}) f(t)dt - f(x) \end{aligned} \tag{1.7}$$

where K is a function of a single variable called the kernel and h is a bins width .

Then doing a change of variable $z = \dfrac{x-t}{h}$, i.e. $t = hz + x, |\dfrac{dz}{dt}| = \dfrac{1}{h}$ yields

$$E(\hat{f}(x)) - f(x) = \textstyle\int_{-\infty}^{+\infty} K(z) f(x - hz)dz - f(x)$$

Expanding $f(x - hz)$ in Taylor series yields

$$f(x - hz) = f(x) - hzf'(x) + \tfrac{1}{2}(hz)^2 f''(x) + o(h^2)$$

Where $o(h^2)$ represents terms that converge to zero faster than h^2 as h approaches zero.

Thus

$$\begin{aligned} E(\hat{f}(x)) - f(x) &= \int_{-\infty}^{+\infty} K(z)f(x)dz - \int_{-\infty}^{+\infty} K(z)hzf'(x)dz - f(x) \\ &\quad + \int_{-\infty}^{+\infty} K(z)\frac{(hz)^2}{2}f''(z)dz + o(h^2) - f(x) \\ &= f(x)\int_{-\infty}^{+\infty} K(z)dz - hf'(x)\int_{-\infty}^{+\infty} zK(z)dz - f(x) \\ &\quad + \frac{h^2}{2}f''(x)\int_{-\infty}^{+\infty} z^2K(z) + o(h^2) \\ &= f(x) + \frac{h^2}{2}k_2f''(x) + o(h^2) - f(x) \\ &= \frac{h^2}{2}k_2f''(x) + o(h^2) \end{aligned} \tag{1.8}$$

The bias of KDE is

$$Bias(\hat{f}(x)) = \frac{h^2}{2}k_2f''(x) + o(h^2)$$

Where $k_2 = \int_{-\infty}^{+\infty} z^2K(z)$

This means that the bias$(\hat{f}(x))$ depends on:

1) h; the bias will shrink at a rate $o(h^2)$ when $h \to 0$

2) The second derivative of the density function $f''(x)$; whenever the density function curves more, the bias will be larger.

Variance: For the variance we have

$$\begin{aligned} Var(\hat{f}(x)) &= Var[\frac{1}{nh}\textstyle\sum_{i=1}^{n} K(\frac{x-x_i}{h})] \\ &= \frac{1}{n^2h^2}\sum_{i=1}^{n} Var[K(\frac{x-x_i}{h})] \end{aligned}$$

Because x_i, $i = 1, 2, ..., n$,are independently distributed, we have :

$$\begin{aligned} Var[K(\frac{x-x_i}{h})] &= E(K(\frac{x-x_i}{h})^2) - (EK(\frac{x-x_i}{h}))^2 \\ &= \int K(\frac{x-t}{h})^2 f(t)dt - (\int K(\frac{x-t}{h})f(t)dt)^2 \end{aligned}$$

$$\begin{aligned} Var(\hat{f}(x)) &= \frac{1}{n}\int \frac{1}{h^2}K(\frac{x-t}{h})^2 f(t)dt - \frac{1}{n}(\frac{1}{h}\int K(\frac{x-t}{h})f(t)dt)^2 \\ &= \frac{1}{n}\int \frac{1}{h^2}K(\frac{x-t}{h})^2 f(t)dt - \frac{1}{n}(f(x) + Bais(\hat{f}(x)))^2 \end{aligned}$$

Then doing a change of variable $z = \frac{x-t}{h}$, we obtain

$$Var(\hat{f}(x)) = \frac{1}{nh}\int K(z)^2 f(x-hz)dz - \frac{1}{n}(f(x) + o(h^2))^2$$

Applying a Taylor approximation yields

$$Var(\hat{f}(x)) = \frac{1}{nh}\int K(z)^2 (f(x) - hzf'(x) + o(h))dz - \frac{1}{n}(f(x) + o(h^2))^2$$

Note that if n becomes large and h becomes small then the above expression becomes:

$$Var(\hat{f}(x)) = \frac{1}{nh}f(x)\int K^2(z)dz + o(\frac{1}{nh})$$

Therefor, if $n \to \infty$ and $h \to 0$, the variance will shrink at rate of $o(\frac{1}{nh})$

MSE. The MSE of KDE becomes:

$$\begin{aligned} MSE(\hat{f}(x)) &= Bais^2(\hat{f}(x)) + Var(\hat{f}(x)) \\ &= \frac{h^2}{2}k_2 f''(x) + o(h^2) + \frac{1}{nh}f(x)\int K^2(z)dz + o(\frac{1}{nh}) \end{aligned} \tag{1.9}$$

Where $k_2 = \int z^2 K(z)dz$ That is , MSE, is a local measure. It is the sum of the square bias and the variance of $\hat{f}(x)$ at x.

Reducing the bias causes variance to increase and vice versa, so trade-off between these terms is needed.

Now integrating MSE with respect to x gives a global measure of conformity of $\hat{f}(x)$ with $f(x)$, called the mean integrated square error, MISE, which is used to estimate the smoothing parameter h [13].

$$\begin{aligned} MISE(\hat{f}(x)) &= E\int_{-\infty}^{+\infty}[\hat{f}(x) - f(x)]^2 dx \\ &= \int_{-\infty}^{+\infty} MSE(\hat{f}(x))dx \\ &= \int_{-\infty}^{+\infty} Bias^2(\hat{f}(x)dx + \int_{-\infty}^{+\infty} Var(\hat{f}(x)) \end{aligned}$$

Also, AMISE, it is an approximation version of MISE developed by expanding MISE into a Taylor series and taking only the most important parts. ISE, integrated square error, is an intermediate measure between MISE and MSE.

1.4.3 Methods for calculating optimum value of smoothing parameter

The smoothing parameter(h) has played an important role in the equality of KDE but unfortunately how to choose h is still an unsolved problem in statistics so researchers have developed several methods to calculate the smoothing parameter. In the following, we are introducing two frequently used methods of bandwidth selection: Rule of thumb method and Least squares cross validation method (LSCV)[13][6].

1) **Rule of Thumb method** :

The simplest possible plug-in bandwidth which assume that the unknown distribution is normal with parameters μ and σ. This method give a good estimate of h if the true density is normal but if not, it fails Significantly.

Silverman got the bandwidth rule of thumb as follows [13]:

$$h = 1.06 \cdot \sigma \cdot n^{\frac{-1}{5}}$$

Where σ is the sample standard deviation and n is the sample size .

2) **Least Square cross validation method (LSCV)**:

It's a very popular technique. It uses the integrated square error ,ISE. The form of the LSCV criterion function is:

$$LSCV(h) = \int_{-\infty}^{+\infty} \hat{f}(x)dx - \frac{2}{n}\sum_{i} \hat{f}_{-i}(x_i)$$

Where $\hat{f}_{-i}(x_i) = \frac{1}{n-1}\sum_{i \neq j} k(x, x_j)$

The optimal smoothing parameter h_{LSCV} is the value for which the LSCV(h) function achieves the minimum. Unfortunately, the LSCV method has drawbacks: sometimes LSCV has several minimums or doesn't have any minimum at all .The variance of the obtained smoothing parameters calculated for samples drawn from the same distribution is large.

This thesis contains three chapters. Chapter one discusses briefly the goodness of fit test, the kernel density estimation and its properties. Chapter two makes an overall view about the five tests we used in this thesis and example of each test. Chapter three compares between the power of the tests under various sample sizes then making conclusions.

Chapter 2

Some Common Goodness of Fit Tests

2.1 Chi-Square Goodness-of-Fit Test

The chi-square test is popular goodness of fit test. It is a nonparametric test developed by Prof A.R.Fisher in 1870 and improved by karl Pearson in 1900 to its modern form. Pearson's Chi-square, is commonly recommended for the multivariate problem where partially ranked categories exist. It's preferred over other tests because of its good features such that it can be applied to any univariate distribution and it can be calculated much easier than other tests.

Chi-square goodness of fit test depends on three assumptions. First, the sample of data is randomly selected from a population of iid observations. Second, the data must be categorised. In the case of non-binned data, a frequency table is used to bin the data into classes. Finally, the expected frequency of each cell is 5 or greater.

If the Chi-square test is applied to a continuous distribution then this distribution has to categorised by defining a set of bins. There are two methods of binning: bins of equal size and bins with equal probabilities. Most statisticians have discouraged using the Chi-square

test for continuous distribution for a long time because it has less power compared to other tests. However, it has one advantage over most other tests which is dealing with parameter estimation very easily.

Definition 2.1. The Chi-square goodness of fit test utilizes the null hypothesis to determine whether a sample of data $x_1, x_2, ..., x_n$ is coming from a specified population.

Chi-Square test statistic

The value of Chi-square test statistic is :

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \tag{2.1}$$

Where

O_i is the observed value

E_i is the expected value

K is the number of different data cells.

Theorem 2.1. (Chi-Squared goodness-of-fit test for simple hypothesis)

Suppose that we observe an independent and identically distributed sample $X_1, ..., X_n$of random variables that take a finite number of values $B_1, ..., B_k$ with unknown probabilities $p_1, ..., p_k$

If $np_j \geq 5$ for all j, then :

$$T = \sum_{i=1}^{k} \frac{(v_j - np_j^0)^2}{np_j^0} \xrightarrow{d} \chi^2_{k-1} \tag{2.2}$$

Where $v_j = \#\{X_i : X_i = B_j\}$ are the observed counts in each category.

To test the hypothesis:

$H_O : p_j = p_j^0$ for all $j = 1, ..., k$

Versus the alternative hypothesis

H_1: $p_j \neq p_j^0$ for some index

at the level of significance, reject if

$$T = \sum_{i=1}^{k} \frac{(v_j - np_j^0)^2}{np_j^0} \geq \chi^2_{1-\alpha,k-1} \tag{2.3}$$

Computing the Chi-Square Goodness-of-Fit Test

The following example will be employed to understand the use of the Chi-square goodness-of-fit test.

Example 2.1. A new casino game involves rolling 3 dice. The winnings are directly proportional to the total number of sixes rolled. Suppose a gambler plays the game 100 times, with the following observed counts:

Number of Sixes	Number of Rolls
0	48
1	35
2	15
3	3

The casino becomes suspicious of the gambler and wishes to determine whether the dice are fair. What do they conclude?

Solution :

H_0 : *The dice is fair*

H_1 : *The dice is not fair*

Computing the test statistic :

First, we may assume :

a)If the dice is fair, then it means that rolling one dice doesn't affect rolling of the others

b) The number of sixes in three rolls is distributed Binomial (3,1/6). (The probability of rolling a 6 in any dice if it is fair is 1/6).

Now, to test whether the gambler's dice is fair, we will compare his result with the results expected under this distribution. Next, we want to find the expected values for 0,1,2 and 3 sixes under the binomial distribution (3,1/6).

To find them, substitute in the equation:

$p(x) = \binom{n}{x} p^x q^{n-x}, x = 0.1.2.3$

$x = 0, p(0) = \binom{3}{0}(1/6)^0(5/6)^3 = 0.58$

Doing the same for $x = 1, 2, 3$ we get:

$p(1) = 0.345, p(2) = 0.07, p(3) = 0.005$

Since the gambler plays 100 times, the expected counts are the following:

Number of Sixes	Expected counts	observed counts
0	58	48
1	34.5	35
2	7.0	15
3	0.5	3

The two plots shown below provide a visual comparison of the expected and observed values:

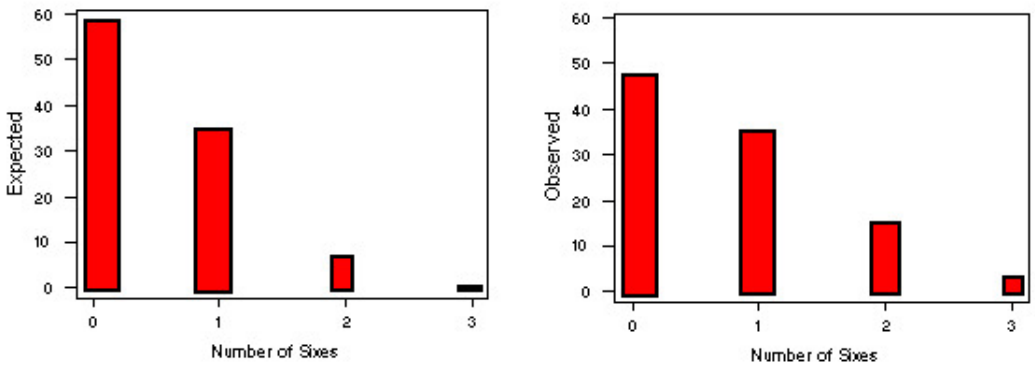

Figure 2.1: Visual comparison of the expected and observed values

From these graphs, it is difficult to distinguish differences between the observed and expected counts.

The chi-square test statistic is

$$\chi^2 = \sum_{i=1}^{3} \frac{(O_i - E_i)^2}{E_i}$$

$$\begin{aligned}\chi^2 &= (48-58)^2/58 + (35-34.5)^2/58 + (15-7)^2/7 + (3-0.5)^2/0.5 \\ &= 1.72 + 0.007 + 9.14 + 12.5 \\ &= 23.367.\end{aligned}$$

Using the table of chi-square with significant level $\alpha = 0.05$ and degrees of freedom $n = 3$ since we have four categories, we get $\chi^2_{0.95} = 7.815$

Now since $\chi^2 > \chi^2_{0.95}$, we reject the null hypothesis. So the dice is not fair.

2.2 Kolmogorov-Smirnov(KS) Test

Kolmogorov-Smirnov test is a well-known nonparametric test which widely used in real life applications. It was firstly proposed by Andrey Kolmogorov (1933) and Nikolai Smirnov (1936) in the $(1930's)$ in papers. Kolmogorov-Smirnov test is based on the maximum distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution. Generally, it's mostly used to decide if a sample comes from a population with a specific distribution.

Previously, KS test were used under some conditions such that $F(x)$ had to be continuous. But there are a lot of real-life applications like physics, engineering and finance that fitted a discrete or mixed distribution and it is important to be able to perform goodness-of-fit tests. So, fortunately, scientists have developed a new fast and accurate method to compute the CDF of the KS statistic when $F(x)$ is discontinuous [11].

Definition 2.2. Kolmogorov-Smirnov test is used to determine whether a sample of data $x_1, x_2, ..., x_n$ is coming from a specified population.

Kolmogorov-Smirnov Test-Statistic

Let $X_1, X_2, ..., X_n$ be observations of a random sample from some distribution F and $X_{(1)}, X_{(2)}, ...X_{(n)}$ be the corresponding order statistic.

Then, the Kolmogorov-Smirnov test statistic is defined as [1]:

$$D_n = \sup_x |F_n(X) - F(X)|, \tag{2.4}$$

Where, $F(x)$ is the cumulative distribution function and F_n is the empirical distribution function of the sample .

It is also can be written in another form :

$$D_n = \frac{1}{n}(\max_i |i - nF(x_i)|); \text{ for } i = 1, 2, ..., n. \tag{2.5}$$

Theorem 2.2. Suppose that we have $X_1, X_2, ..., X_n$ an independent and identically distributed sample from a continuous population such that $X_{(1)}, X_{(2)}, ...X_{(n)}$ is the order statistic with cumulative distribution function $F(x)$.We would like to test the hypothesis that $F(x)$ is equal to a particular distribution $F_0(x)$, where the hypothesis can be written as follows :

$$H_0 : F = F_0$$
$$H_1 : F \neq F_0$$

Where F_0 is some specified distribution function,

Then , reject H_0 if the calculated value is greater than the critical value of the Kolmogrove -Smirnov test.

Computing the Kolmogorov Smirnov Goodness of Fit Test

The following example is an illustration of the previous theorem .

Example 2.2. In a study conducted from different streams of a college of 60 students, with an equal number of students selected from each stream, were interviewed and their intention to join the college drama club was noted.

	B.SC	B.A	B.Com	M.A	M.Com
No.of each class	5	9	11	16	19

Table 2.1: Data for the Kolmogorov-Smirnov Goodness of Fit Test

It was expected that 12 students from each class would join the Drama Club. Using the K-S test to find if there is any difference among student classes with regard to their intention of joining the Drama Club.

Solution: we want to test the hypothesis

H_0 : There is no difference among students of different streams with respect to their intention of joining the drama club.

H_1 : There is a difference among students of different streams with respect to their intention of joining the drama club.

We develop the cumulative frequencies for observed and theoretical distributions.

Streams	No. of students interested in joining (observed)	(Theoretical)	$F_n(x)$	$F(x)$	$F_n(x) - F(x)$
B.Sc	5	12	5/60	12/60	7/60
B.A	9	12	14/60	24/60	10/60
B.Com	11	12	25/60	36/60	11/60
M.A	16	12	41/60	48/60	7/60
M.Com	19	12	60/60	60/60	0/60
Total	n=60				

Table 2.2: Computations of Kolmogorov-Smirnov Goodness of Fit Test

Test statistic $|D|$ is calculated as:

D=Maximum $|F_n(X) - F(X)| = 11/60 = 0.183$

The table value of D at 0.05 significance level is given by($n = 60 > 50$):

$D_{0.05} = 1.36/\sqrt{n} = 1.36/\sqrt{60} = 0.175$

Reject the null hypothesis since the critical value is less than the calculated value. We conclude that there is a difference among students of different streams in their intention of joining the Club.

2.3 Shapiro-Wilk (SW) Test

Shapiro-Wilk test is one of the most powerful goodness of fit test that proposed by Samuel Sanford Shapiro and Martin Wilk in 1965. It's an improvement of Kolmogorv-Smirnov test but it doesn't affect by ties like Anderson-Darling test. It is used to test if sample of data $x_1, x_2, ..., x_n$ are normally distributed. This test is mostly preferred over other tests because of detecting trivial departures from the null hypothesis.

Definition 2.3. Shapiro-Wilk test (SW) is used to check whether or not a given sample follows a population with normal distribution.

Shapiro-Wilk (SW) Test-Statistic

Suppose $x_1, x_2, ..., x_n$ are the observations of a random sample from a population with normal distribution .Suppose also that $x_{(1)}, x_{(2)}, ...x_{(n)}$ are the corresponding order statistics.
Then, the Shapiro-Wilk (SW) test statistic is defined as [10]:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \tag{2.6}$$

Where

x_i: the i^{th} order statistic and $\bar{x} = (x_1 + x_2 + ... + x_n)/n$, the sample mean

The coefficient a_i are given by:
$(a_1, \ldots, a_n) = \dfrac{m^{\mathsf{T}} V^{-1}}{C}$,
where C is a vector norm :
$C = \|V^{-1}m\| = (m^{\mathsf{T}} V^{-1} V^{-1} m)^{1/2}$
the vector m :
$m = (m_1, \ldots, m_n)^{\mathsf{T}}$ and

$(m_1, m_2, .., m_n) = (EZ_{(1)}, EZ_{(2)}, ..., EZ_{(n)})$ such that $EZ_{(1)} \leqslant EZ_{(2)} \leqslant ... \leqslant EZ_{(n)}$ are the order statistic of independence and identically $Z_1, Z_2, ..., Z_n$ which are normally distributed V is the covariance matrix of those normal order statistic.

Theorem 2.3. Suppose $x_1, x_2, ..., x_n$ are the observations of a random sample from a population with normal distribution. Suppose also that $x_{(1)}, x_{(2)}, ...x_{(n)}$ are the corresponding order statistics. Consider we have to test:

H_0: The sample of data is coming from a normal population, $N(\mu, \sigma^2)$

against

H_1: The sample of data doesn't follow $N(\mu, \sigma^2)$

Then, reject H_0 at a significant level α if the calculated value of the test statistic W is smaller than W_α:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} < W_\alpha \tag{2.7}$$

Remark 2.1. The statistic W is always greater than zero and less than or equal to one $(0 \leq W \leq 1)$.

Computing the Shapiro Wilk Goodness of Fit Test

The following example is an application of the previous theorem:

Example 2.3. A random sample of 12 people is taken from a large population. The ages of the people in the sample are 65,61,63,86,70,55,74,35,72,68,45,58. Is this data normally distributed?

Solution:

Set the null and alternative hypothesis:

H_0: the data is normally distributed.

H_1: the data is not normally distributed.

The W test statistic is:

$$W = b^2/SS$$

Where $b = \sum_{i=1}^{m} a_i(x_{(n-i+1)} - x_{(i)})$

$SS = \sum_{i=1}^{n}(x_i - \overline{x})^2$

The coefficient a_i's are from Shapiro and Wilk (1965)

$m = n/2$ since n is even .

Age	Sorted	$(x_{(n-i+1)} - x_{(i)})$	a_i	$a_i(x_{(n-i+1)} - x_{(i)})$	$(x_i - \overline{x})^2$
65	35				765.462889
61	45	$n = 12$	m=6		312.122889
63	55				58.782889
86	58	51	a_1 =0.5475	27.9225	21.780889
70	61	29	a_2= 0.3325	9.6425	2.778889
55	63	17	a_3= 0.2347	3.9899	0.110889
74	65	12	a_4= 0.1586	1.9032	5.442889
35	68	7	a_5 = 0.0922	0.6454	28.440889
72	70	2	a_6=0.0303	0.0606	53.772889
68	72				87.104889
45	74				128.436889
58	86				544.428889
	$\bar{x} = 62.67$			$\sum_{i=1}^{m} a_i(x_{(n-i+1)} -$ $x_{(i)})$=44.1641	$\sum_{i=1}^{n}(x_i - \overline{x})^2$=2008.667

Table 2.3: Computations of Shapiro Wilk Goodness of Fit Test

$b = \sum_{i=1}^{m} a_i(x_{(n-i+1)} - x_{(i)}) = 44.1641$

$SS = \sum_{i=1}^{n}(x_i - \overline{x})^2 = 2008.667.$

Therefor $W = b^2/SS = 44.1641^2/2008.667 = .971026$.

Now from the table of Shapiro Wilk test for 0.971026 and when $n = 12$ we find that the p-value lies between 0.5 and 0.9. The Wvalue for 0.5 is 0.943 and the Wvalue for 0.9 is 0.973.

Interpolating 0.971026 between these values by using linear interpolation, we get the p-value:

$$P = 0.9 - \frac{(0.9 - 0.5)(0.973 - 0.971026)}{(0.973 - 0.943)} = 0.8736$$

Since p-value=0.8736 $> 0.05 = \alpha$, we can retain the null hypothesis that the data is normally distributed.

2.4 Anderson-Darling(AD) Test

Anderson-Darling is one of the most important and powerful goodness of fit tests which is a modification of K-S test. It was proposed in 1954 by Theodore Anderson and Donald Darling and was mainly used for engineering purposes. Anderson Darling test is used to test if a sample of data come from a population with a specified distribution.

Anderson Darling test is more sensitive than KS test since the critical values are calculated relying on the specific distribution we are testing. Also it is sensitive to the shape and scale of distribution. It can be applied to a small sample and it gives more weightage to the tails of the distribution. Like KS test, it's based on the cumulative probability distribution of data [10].

Definition 2.4. Anderson-Darling test (AD) is used to check whether or not a given sample come from a certain specified population.

Anderson-Darling(AD) Test-Statistic

Suppose $x_1, x_2, ..., x_n$ are the observations of a random sample from a population distribution with a cumulative distribution function $F(x)$. Suppose also that $x_{(1)}, x_{(2)}, ...x_{(n)}$ are the corresponding order statistics. Then, the Anderson Darling test statistic is defined as:

$$A = -n - S \tag{2.8}$$

Where,

$$S = \sum_{i=1}^{n} \frac{2i-1}{n} [\ln F_0(Z_i) + \ln(1 - F_0(z_{n+1-i}))]$$

F is the cumulative distribution function of the specified distribution and Z_i are the ordered data [8].

Theorem 2.4. Suppose $x_1, x_2, ..., x_n$ are the observations of a random sample from a population distribution with a cumulative distribution function $F(x)$. Suppose also that $x_{(1)}, x_{(2)}, ...x_{(n)}$ are the corresponding order statistics. Consider the problem of testing:

$$H_0 : F = F_0$$
$$H_1 : F \neq F_0$$

Where F_0 is some specified distribution function,
Then, reject H_0 if the calculated value is greater than the critical value of the Anderson-Darling test.

Computing the Anderson Darling Goodness-of-Fit Test

The following example is an application of the previous theory:

Example 2.4. Test the normality for the data contains a small sample of six batches, drawn at random from the same population.The six observations are :338.7, 308.5, 317.7, 313.1, 322.7, 249.2 [9].

Solution:

To assess the normality of the sample, we first obtain the point estimations of the assumed normal distribution parameters: sample mean and standard deviation.

Variable	N	Mean	Median
Data Set	6	315.82	315.40

Table 2.4: Descriptive Statistics of the Data

Then we calculate the Anderson Darling statistic using the data in the table (2.4)

i	x	$F_0(Z)$	$N+1-i$	$F_0(z_{N+1-i})$	$1-(F_0(z_{N+1-i}))$	$ln(1-(F_0(z_{N+1-i})))$	$lnF(Z)$
1	249.2	0.072711	6	0.938310	0.061690	-2.78563	-2.62126
2	308.5	0.311031	5	0.678425	0.321575	-1.13453	-1.16786
3	313.1	0.427334	4	0.550371	0.449629	-0.79933	-0.85019
4	317.7	0.550371	3	0.427334	0.572666	-0.55745	-0.59716
5	322.7	0.678425	2	0.311031	0.688969	-0.37256	-0.38798
6	338.7	0.938310	1	0.072711	0.927289	-0.07549	-0.06367

Table 2.5: Computations of Anderson Darling Goodness of Fit Test

$$A = \frac{1}{6}(-2.62126 + -2.78563) + \frac{3}{6}(-1.16786 + -1.13453) + \frac{5}{6}(-0.85019 + -0.79933)$$
$$+ \frac{7}{6}(-0.59716 + -0.55745) + \frac{9}{6}(-0.38798 + -0.37256) + \frac{11}{6}(-0.06367 + -0.07549) - 6$$
$$= -0.901148 + -1.151195 + -1.3746 + -1.347045 + -1.14081 + -0.2551266 - 6$$
$$= 6.16993 - 6$$
$$= 0.16993$$

The AD statistic is 0.1699

The critical value for a normal distribution is calculated as:

$$CV = \frac{0.752}{1 + 0.75/n + 2.25/n^2} = \frac{0.752}{1 + 0.752/6 + 2.25/36} = 0.6333.$$

Since $AD = 0.1699 < CV = 0.6333$, Anderson Darling test doesn't reject the null hypothesis: the sample may have been drawn from a normal population.

2.5 Bickel-Rosenblatt Test

Peter Bickel and Murray Rosenblatt proposed this test in 1973 which measure the L_2 distance between a kernel density estimate for the underlying density and its expected value under the null hypothesis. Then, many authors have been studied it to improve its power like Fan(1998), Neuhaus (1987), Konakov, Lauter and Liero (1995) who extended the Bickel-Rosenblatt test to a more general null hypothesis such that the underlying density lies in a parametric class. The asymptotic null distribution of this test depend on an integrated square distance between a nonparametric density estimate and its expected value under the null hypothesis.

The value of test statistic of Bickel-Rosenblatt test depends on the bandwidth parameter, sample size (n), selected kernel function, the weighted function $a(x)$, and on the specified distribution with probability density function that is being tested [12].

Definition 2.5. Bickle-Rosenblatt (BR) test is employed to test whether a sample of data $x_1, x_2, ..., x_n$ is coming from a specified distribution.

Bickle-Rosenblatt test statistic

Let $X_1, X_2, ..., X_n$ be independent identically distributed random variables with a specified continuous probability density function $f(x)$. Bickel-Rosenblatt (BR) test statistic is defined as follows:

$$T_n = nb_n \int [f_n(x) - E_0 f_n(x)]^2 a(x) dx \tag{2.9}$$

Where f_n is the kernel estimator,

$$f_n(x) = \frac{1}{nb_n} \sum_{i=1}^{n} K\big[\frac{x - X_i}{b_n}\big] \tag{2.10}$$

a is a suitable chosen function on R and $E_0 f_n(x)$ denotes the expectation of f_n under f_0.

Theorem 2.5. (Bickel-Rosenblatt Goodness-of-Fit Test)

Let $X_1, X_2, ..., X_n$ be independent identically distributed random variables with a specified continuous probability density function $f(x)$,Consider we have to test [12]:

$$H_0 : f = f_0$$

Against

$$H_1 : f \neq f_0$$

where f_0 is completely specified at a specified significance level α.

Reject H_0 if

$T_n \geq \mu(K, a) + z_\alpha b_n^{1/2} \sigma(K, a)$ Where

$$\mu(K,a) = I(K)\int f_0(x)a(x)dx,$$
$$\sigma^2(K,a) = 2J(K)\int f_0^2(x)a^2(x)dx$$

and

$$I(K) = \int K^2(x)dx$$
$$J(K) = \int [\int K(x+y)k(x)dx]^2 dy$$

also $\Phi(z_\alpha)$ defined by:

$$\Phi(z_\alpha) = 1-\alpha, \Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} exp(-x^2/2)dx$$

Chapter 3

Simulation and Results

In this chapter, simulation studies will be employed to verify the power of the proposed goodness of fit tests in testing whether a random sampled compose of independent observations comes from a population that has a specified distribution.

However, the main hypothesis in such a case can be written in the following manner:

H_0: The data come from a specified distribution

H_a: The data is not related to the specified distribution.

Testing the hypothesis will be conducted under various sample sizes and level of significance $\alpha = 0.05$ to investigate the effect of the sample size and level of significance on the power of the goodness of fit tests since the significance level and sample size have a role in determining the power of the goodness of fit tests [14]. However, the proposed sample sizes in this simulation are:

$n = 20, 40, 80, 100, 120, 150, 180, 200, 300, 500, 1000$ and 2000.

The critical values for each test were obtained from the nonparametric alternatives based on 10000 simulated samples from T-distribution, standard normal distribution and uniform distribution.

The critical values of Kolmogorov Smirnov and Shapiro Wilk tests depend on both the sig-

nificance level and sample size, where these tests are right tailed tests so the critical values can be obtained by getting the $100(1-\alpha)^{th}$ percentile of the empirical distribution of the test statistics. The Anderson Darling test and χ^2 test are the same as Kolmogorov Smirnov test and Shapiro Wilk test but the critical value of the Anderson Darling test depends on the distribution being tested while the χ^2 test depends on the number of intervals for the grouped data set.

The power of statistical test can be defined as the complement of type II error (β); in other words it is the probability of rejecting the null hypothesis when a hypothesis testing is conducted. The power test provides the ability to determine whether the study has a sensible chance of obtaining statistically significant results. The power of statistical tests should be accounted to avoid the risk of type II error. However, the power of statistical test is defined as the test which has the higher percentage of rejecting null hypothesis. Therefore, higher power of statistical test implies higher ability of rejecting the null hypothesis [3].

However, to verify the simulated power of the normality tests at the proposed significance levels for each sample size, 10000 samples can be generated from symmetric parametric alternative distributions such as: Normal$(0,1)$, Uniform$(0,1)$, and T(6), While from the skewed parametric distributions the data can be generated from: Exponential(0.8), Exponential(4), $\chi^2(5)$ and weibull(10,2). After data generation based on the proposed algorithm, the test statistic for each goodness fit test is calculated under the null hypothesis. Then a comparison between the values of the test statistics with the obtained critical values **where the power of the test is determined by determining the number of test statistics that exceed the given critical values divided by the number of generated samples**, and hence the power of the goodness of fit tests represent probabilities.

In this chapter, the focus is on the construction of a simulation study and gets a comparison between the powers of the proposed goodness of fit tests discussed in the previous chapter based on the empirical distribution function and the nonparametric alternatives. The results will be shown in interchangeable manner, where the symmetric parametric distributions in the null hypothesis will be discussed versus symmetric parametric distributions such as: $N(0,1), U(0,1)$ Exponential (0.8), Exponential (4), $\chi^2(4)$ and student-t(6 (alternative hypothesis) and versus skewed parametric distributions(alternative hypothesis).

Case 1:

The goodness of fit tests between symmetric distributions and symmetric parametric distributions where the results are shown in the following tables. The data were generated from the distributions determined in the alternative hypothesis for different samples sizes and different number of intervals (K) and different kernel functions (Uniform, Epanechnikov) for 10000 simulations.

Table A1: The critical values under normal distribution at different sample size (n) and at $\alpha = 0.05$ for the tests: Chi-square (χ^2), Kolmogorov-Smirnov (KS), Anderson Darling (AD), Shapiro-Wilk (SW) and Bickel-Rosenblatt (BR).

n	χ^2	KS	AD	SW	BR	
	K=5				Uniform	Epanechnikov
20	17.02	0.273	2.61	0.216	1.6120	0.93461
40	17.02	0.261	2.61	0.217	1.5723	0.91032
80	17.02	0.223	2.61	0.219	1.5132	0.88731
100	17.02	0.208	2.61	0.219	1.4704	0.83213
120	17.02	0.189	2.61	0.219	1.4119	0.80112
150	17.02	0.152	2.61	0.219	1.3832	0.78615
180	17.02	0.115	2.61	0.219	1.3113	0.77212
200	17.02	0.084	2.61	0.219	1.2976	0.76312
300	17.02	0.071	2.61	0.219	1.2643	0.75219
500	17.02	0.061	2.61	0.219	1.2522	0.74312

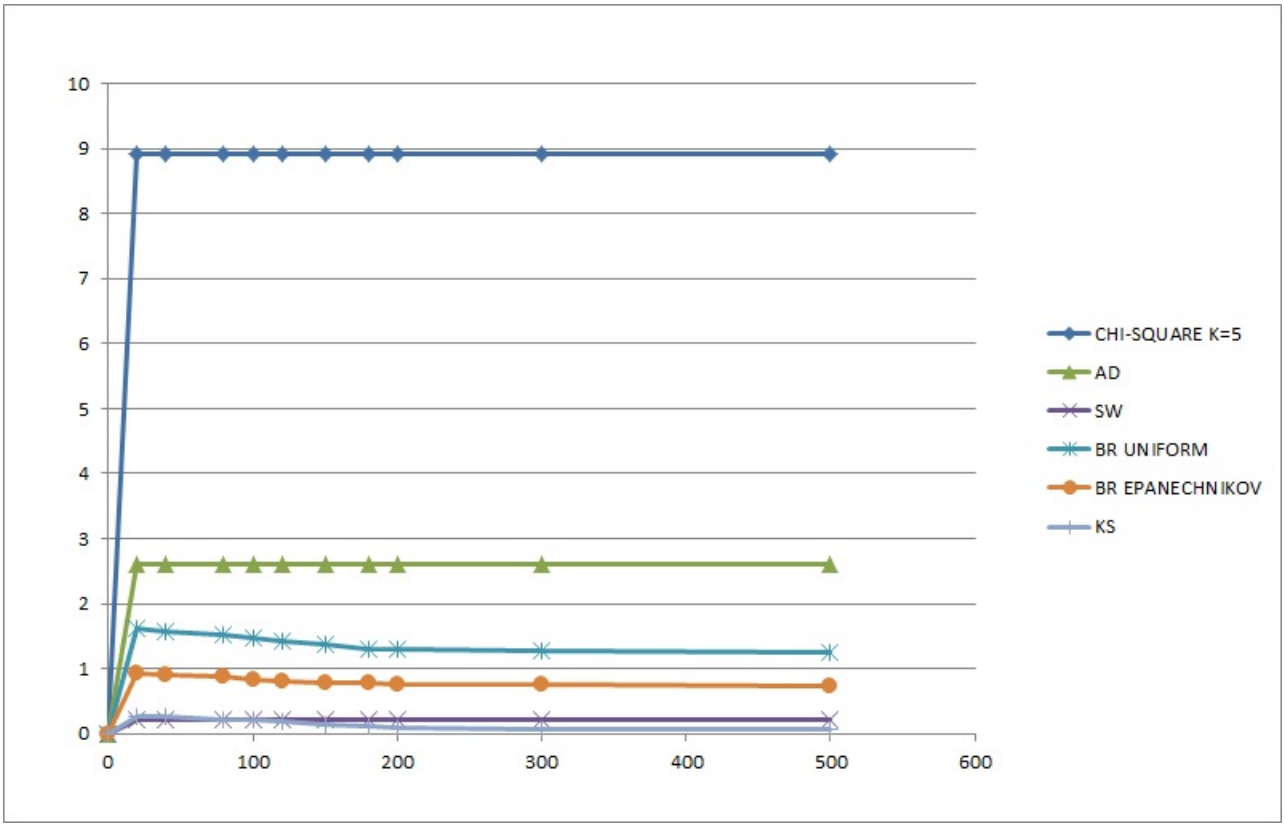

Figure 3.1: Critical value for some goodness of fit tests for different sample sizes.

Table A2: The power of goodness of fit tests :χ^2, KS, AD, SW, and BR for various sample size at $\alpha = 0.05$ under the hypothesis:

H_0: Normal $(0.5, 1)$

H_a: Normal $(0, 1)$

n	χ^2	KS	AD	SW	BR	
	K=5				Uniform	Epanechnikov
20	0.0043	0.7091	0.5878	0.8762	0.0672	0.1772
40	0.7761	0.7419	0.7489	0.9145	0.0910	0.2112
80	0.9532	0.7991	0.7765	0.9712	0.1320	0.3712
100	1.00	0.8398	0.8403	1.00	0.2130	0.5876
120	1.00	0.8891	0.8904	1.00	0.4561	0.7659
150	1.00	0.9612	0.9554	1.00	0.6321	0.5603
180	1.00	0.9910	0.9951	1.00	0.7551	0.9231
200	1.00	1.00	1.00	1.00	0.8702	0.9712
300	1.00	1.00	1.00	1.00	0.9703	0.9930
500	1.00	1.00	1.00	1.00	0.9931	1.00

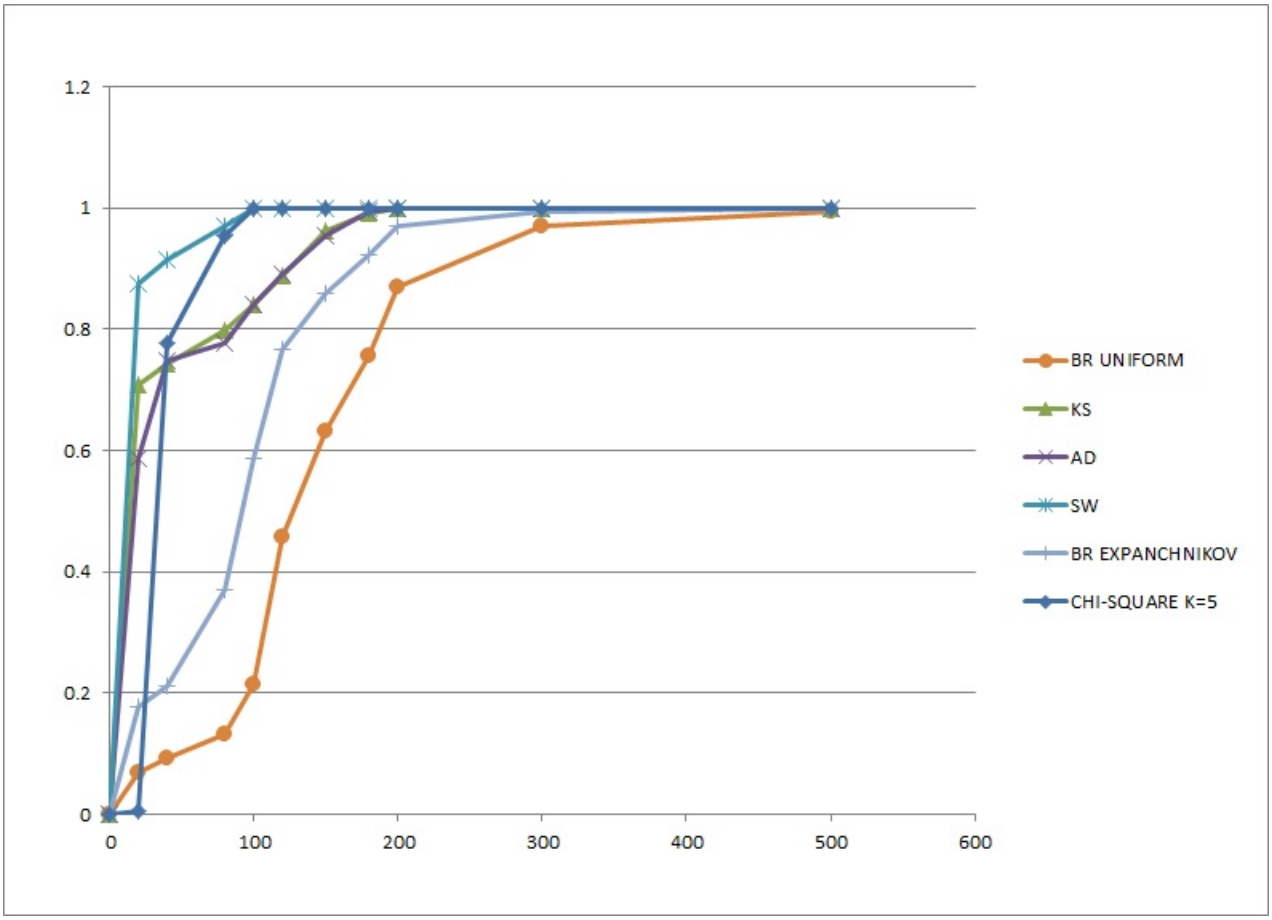

Figure 3.2: The power of the selected goodness of fit tests.

Table B1: The critical values under normal distribution at different sample size (n) and at $\alpha = 0.05$ for the tests: Chi-square (χ^2), Kolmogorov-Smirnov (KS), Anderson Darling (AD), Shapiro-Wilk (SW) and Bickel-Rosenblatt (BR).

n	χ^2	KS	AD	SW	BR	
	K=5				Uniform	Epanechnikov
10	8.912	0.418	0.761	0.221	2.184	1.4023
30	8.912	0.210	0.761	0.226	1.958	1.2420
40	8.912	0.189	0.761	0.231	1.887	1.1161
50	8.912	0.137	0.761	0.231	1.852	1.1091
100	8.912	0.114	0.761	0.231	1.731	1.0311
200	8.912	0.062	0.761	0.231	1.621	0.9432
300	8.912	0.051	0.761	0.231	1.512	0.9123
500	8.912	0.044	0.761	0.231	1.443	0.8971
1000	8.912	0.032	0.761	0.231	1.387	0.8541
2000	8.912	0.022	0.761	0.231	1.332	0.8134

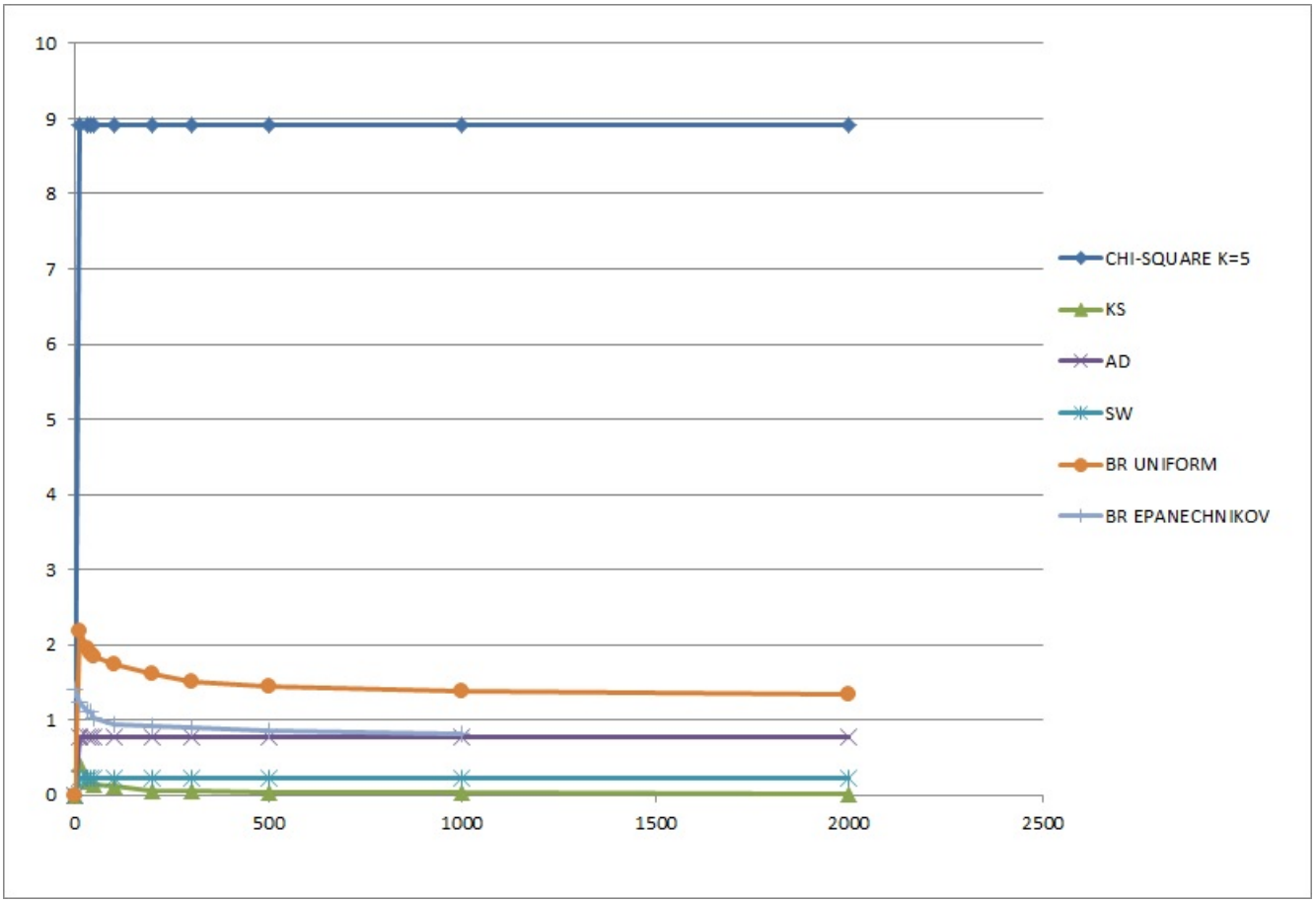

Figure 3.3: Critical value for some goodness of fit tests for different sample sizes.

Table B2: The power of goodness of fit tests: χ^2, KS, AD, SW and BR for various sample size at $\alpha = 0.05$ under the hypothesis:

H_0 : Mean and Variance of a Normal $\equiv$ Mean and Variance of Uniform(0,1)

H_a : Uniform$(0, 1)$

n	χ^2	KS	AD	SW	BR	
	K=5				Uniform	Epanechnikov
10	0.001	0.002	0.079	0.120	0.0087	0.0042
30	0.003	0.009	0.208	0.198	0.0231	0.0175
40	0.213	0.019	0.623	0.234	0.0623	0.2843
50	0.651	0.071	0.928	0.435	0.231	0.7765
100	0.945	0.381	1.00	0.591	0.5642	0.9332
200	0.997	0.683	1.00	0.723	0.9114	1.00
300	1.00	0.850	1.00	0.791	0.9942	1.00
500	1.00	1.00	1.00	0.873	1.00	1.00
1000	1.00	1.00	1.00	0.906	1.00	1.00
2000	1.00	1.00	1.00	1.00	1.00	1.00

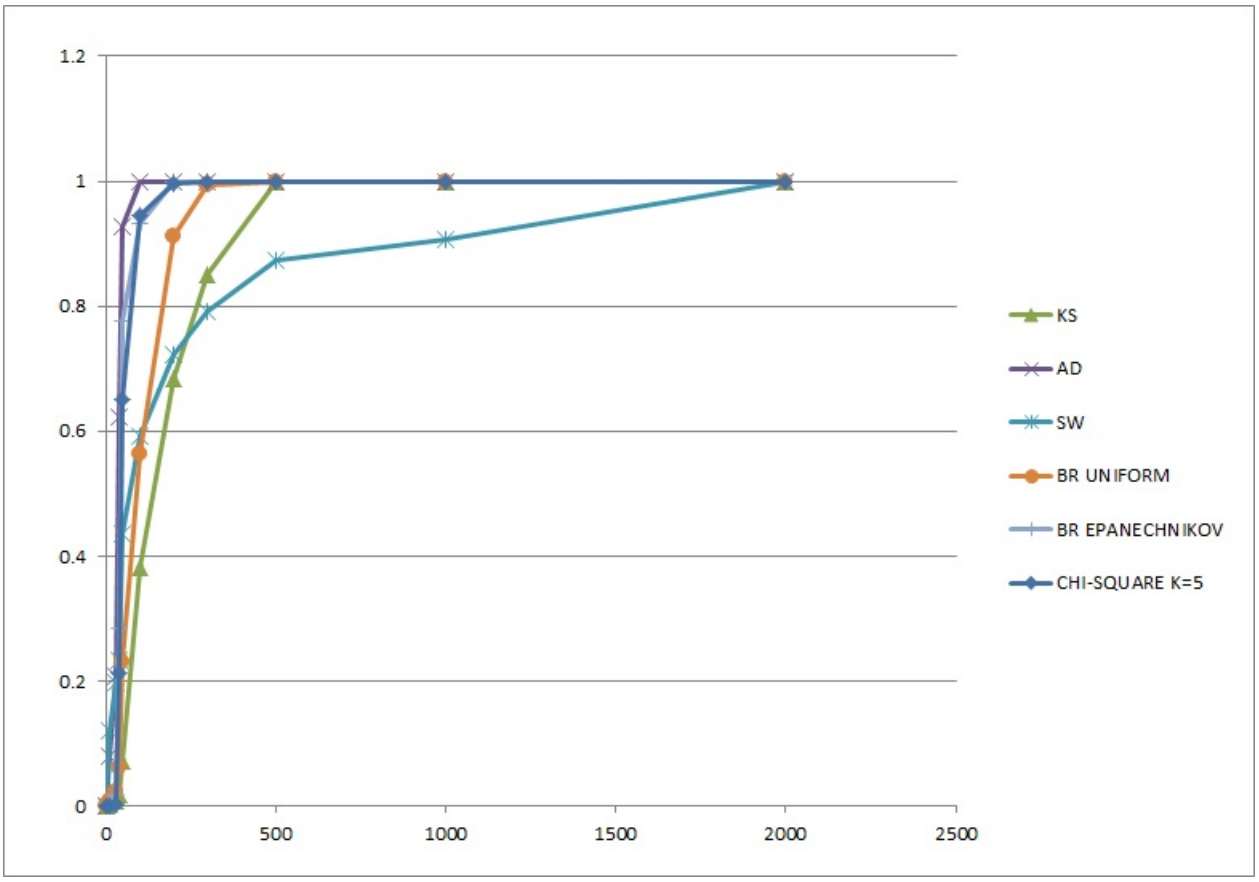

Figure 3.4: The power of the selected goodness of fit tests.

Table C1: The critical values under normal distribution at different sample size (n) and at $\alpha = 0.05$ for the tests: Chi-square (χ^2), Kolmogorov-Smirnov (KS), Anderson Darling (AD), Shapiro-Wilk (SW) and Bickel-Rosenblatt (BR).

n	χ^2	KS	AD	SW	BR	
	K=5				Uniform	Epanechnikov
10	8.912	0.421	0.761	0.205	1.563	0.9834
30	8.912	0.312	0.761	0.211	1.451	0.9351
40	8.912	0.186	0.761	0.219	1.401	0.8942
50	8.912	0.139	0.761	0.233	1.361	0.8612
100	8.912	0.105	0.761	0.241	1.332	0.8122
200	8.912	0.075	0.761	0.241	1.301	0.7834
300	8.912	0.053	0.761	0.241	1.252	0.7530
500	8.912	0.048	0.761	0.241	1.210	0.7239
1000	8.912	0.023	0.761	0.241	1.188	0.7006
2000	8.912	0.021	0.761	0.241	1.151	0.6772

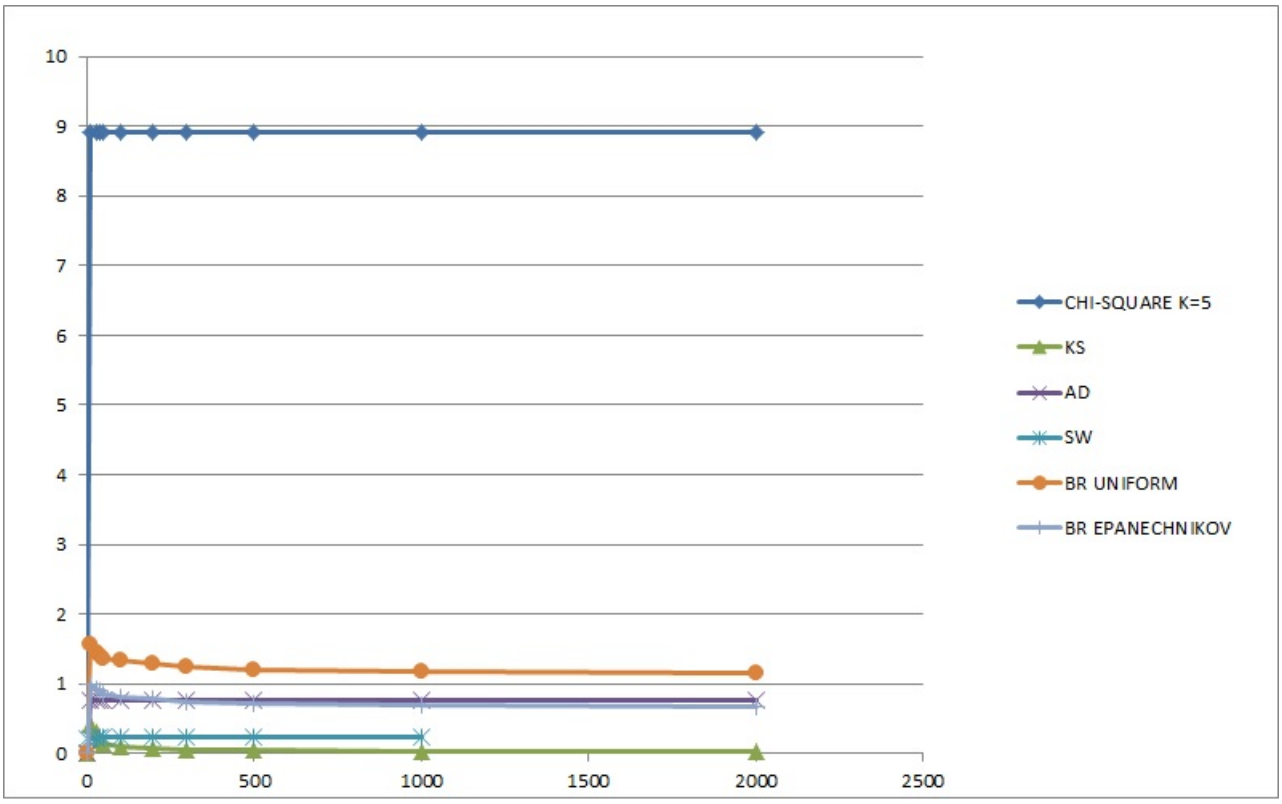

Figure 3.5: Critical value for some goodness of fit tests for different sample sizes

Table C2: The power of goodness of fit tests: χ^2, KS, AD, SW and BR for various sample size at $\alpha = 0.05$ under the hypothesis:

H_0 : Mean and Variance of a Normal $\equiv$ Mean and Variance of t(6)

$H_a : t(6)$

n	χ^2	KS	AD	SW	BR	
	K=5				Uniform	Epanechnikov
10	0.003	0.0401	0.0792	0.0853	0.013	0.0223
30	0.037	0.0441	0.1163	0.1182	0.0182	0.0366
4 40	0.057	0.0463	0.1431	0.1523	0.0232	0.0432
50	0.114	0.0521	0.1772	0.1863	0.0296	0.0632
100	0.232	0.895	0.2832	0.2861	0.0401	0.2776
200	0.361	0.1324	0.5120	0.4723	0.0590	0.5121
300	0.443	0.1667	0.6122	0.6129	0.0834	0.7343
500	0.634	0.2341	0.7621	0.7762	0.1564	0.8023
1000	0.878	0.4981	0.8671	0.8562	0.2951	0.8501
2000	0.979	0.8421	0.9562	0.9892	0.3521	0.8601

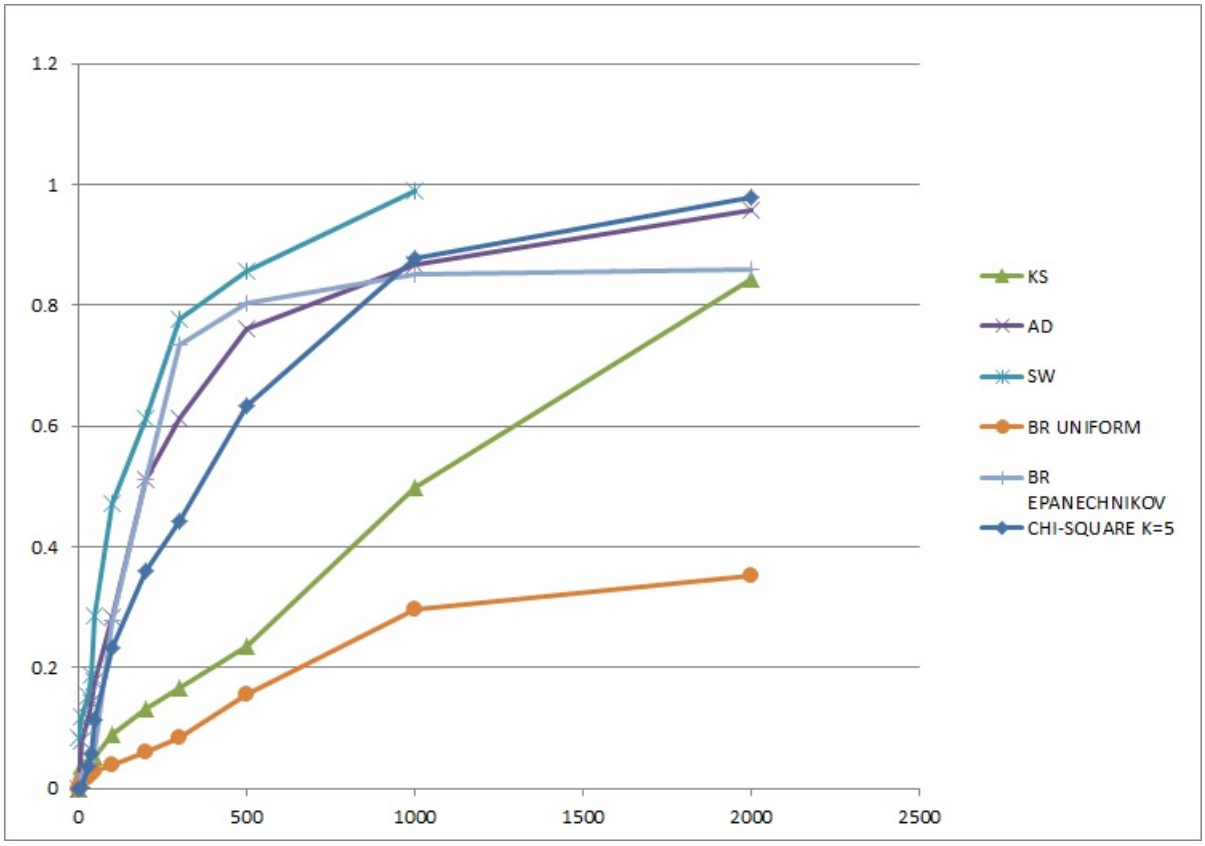

Figure 3.6: The power of the selected goodness of fit tests.

The results in Table A2 show that all tests have good power as the sample size close to 300 or above. The chi-square and Shapiro Wilk tests are able to meet the null hypothesis assumption when the sample size becomes greater than 100. Furthermore, the Bickel-Rosenblatt test is more efficient in the case of Epanechnikov kernel than in the case of uniform which means that Bickel-Rosenblatt is sensitive to the kernel function selection. The Shapiro-Wilk test has the best power with respect to the Kolmogorov-Smirnov, Anderson Darling and Bickel-Rosenblatt tests under various sample sizes.

In the further Table (Table B2), the results indicate that all tests have good power when the sample size is greater than 200 except the Kolmogorov-Smirnov test and Shapiro Wilk test. The power of Bickel-Rosenblatt is improved when the sample size ≥ 200 and also the power of this test is high in the case of Epanechnikov kernel rather than the uniform kernel as the sample size becomes greater than 40. Bickel-Rosenblatt test is more powerful than Kolmogorov-Smirnov and Shapiro-Wilk tests for sample sizes of more than 40.

Consequently, Table C2 shows that Bickel-Rosenblatt test has higher power when the Epanechnikov kernel is employed rather than uniform kernel at any sample size. Furthermore, Epanechnikov kernel also has higher power than Kolmogorov-Smirnov and Shapiro-Wilk tests in some cases of high sample sizes.

Case 2:

The goodness of fit tests between symmetric distributions and skewed parametric distributions. The data were generated based on the distributions in the alternative hypothesis for different sample sizes and different number of intervals (K) and the same kernel functions used in the previous case (Uniform, Epanechnikov) for 10000 simulations.

Table D1: The critical values under normal distribution at different sample sizes (n) and at $\alpha = 0.05$ for the tests: Chi-square χ^2, Kolmogorov-Smirnov (KS), Anderson Darling (AD), Shapiro-Wilk (SW) and Bickel Rosenblatt test (BR).

n	χ^2	**KS**	**AD**	**SW**	BR	
	K=5				Uniform	Epanechnikov
20	8.032	0.213	0.773	0.231	2.512	1.471
40	8.032	0.201	0.773	0.238	2.421	1.401
80	8.032	0.187	0.773	0.241	2.313	1.332
100	8.032	0.157	0.773	0.241	2.175	1.251
120	8.032	0.134	0.773	0.241	1.982	1.142
150	8.032	0.118	0.773	0.241	1.853	1.081
180	8.032	0.089	0.773	0.241	1.723	1.016
200	8.032	0.071	0.773	0.241	1.523	0.972
300	8.032	0.0551	0.773	0.241	1.441	0.951
500	8.032	0.044	0.773	0.241	1.412	0.934

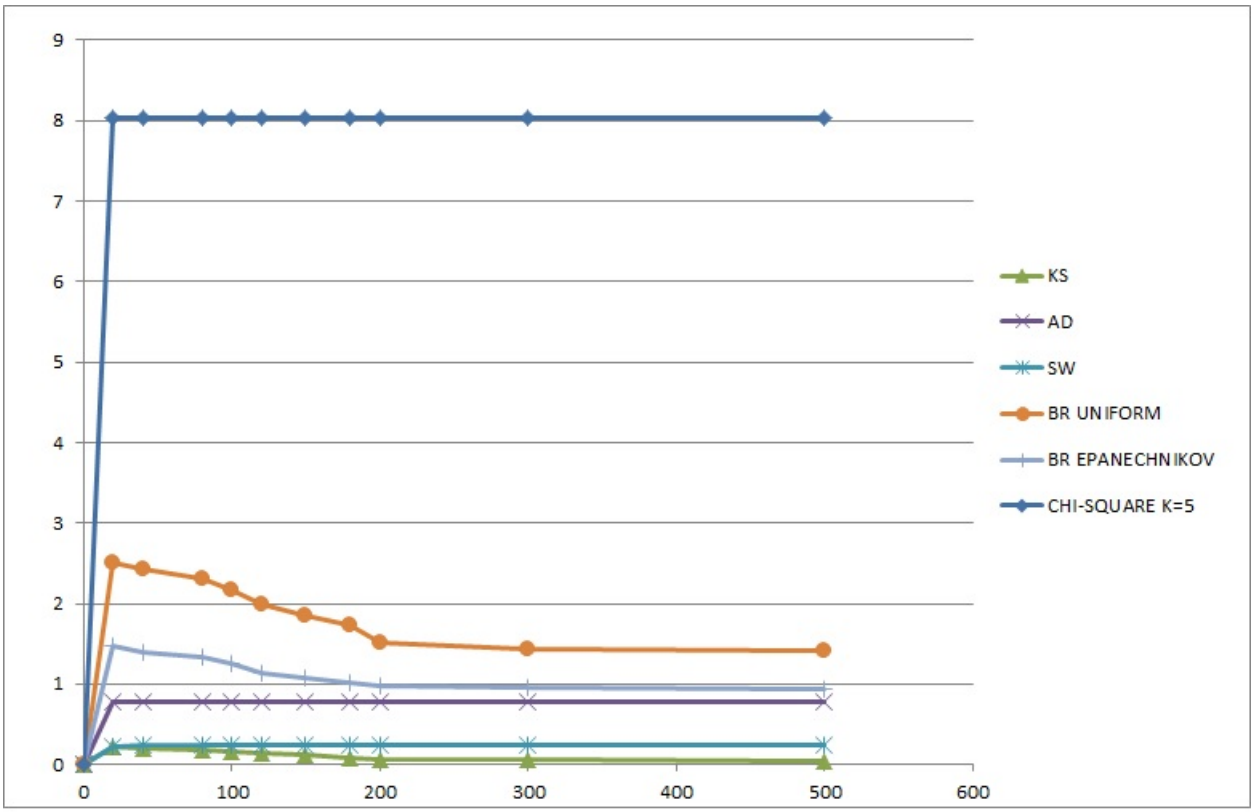

Figure 3.7: Critical value for some goodness of fit tests for different sample sizes.

Table D2: The power of goodness of fit tests: χ^2, KS, AD, SW and BR for various sample size at $\alpha = 0.05$ under the hypothesis:

H_0 : Mean and Variance of a Normal $\equiv$ Mean and Variance of Exponential (4)

H_a : Exponential (4)

n	χ^2	KS	AD	SW	BR	
	K=5				Uniform	Epanechnikov
20	-	0.301	0.910	0.313	0.112	0.0771
40	0.601	0.421	0.944	0.423	0.311	0.3120
80	0.691	0.554	0.982	0.661	0.507	0.7792
100	0.742	0.712	0.961	0.779	0.812	0.8812
120	0.912	0.834	0.989	0.878	0.914	0.9236
150	0.994	0.971	1.00	0.991	0.991	0.9891
180	1.00	1.00	1.00	1.00	1.00	1.00
200	1.00	1.00	1.00	1.00	1.00	1.00
300	1.00	1.00	1.00	1.00	1.00	1.00
500	1.00	1.00	1.00	1.00	1.00	1.00

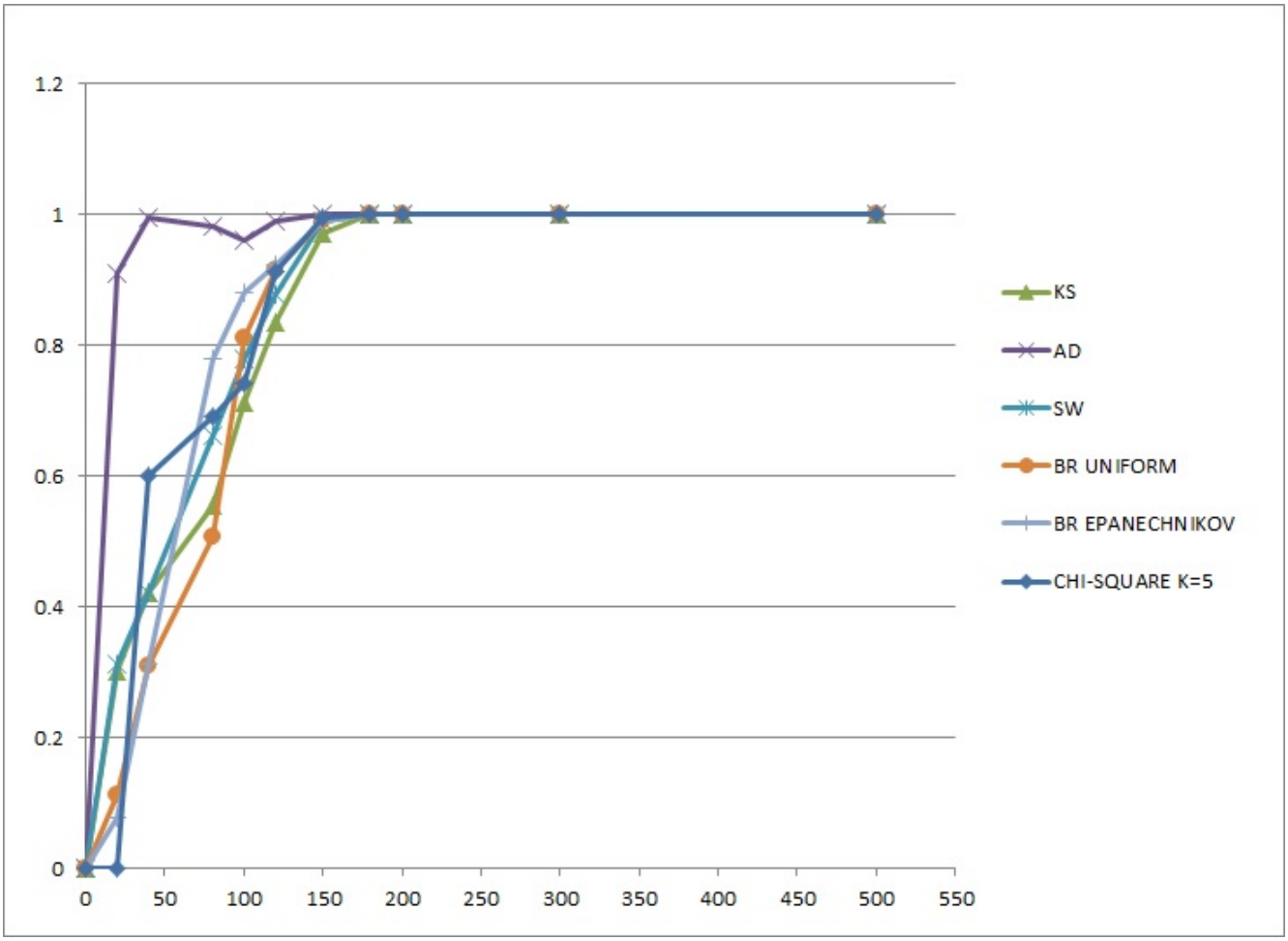

Figure 3.8: The power of the selected goodness of fit tests.

Table E1: The critical values under normal distribution at different sample size (n) and at $\alpha = 0.05$ for the tests: Chi-square (χ^2), Kolmogorov-Smirnov (KS), Anderson Darling (AD), Shapiro-Wilk (SW) and Bickel-Rosenblatt (BR).

n	χ^2	KS	AD	SW	BR	
	K=5				Uniform	Epanechnikov
10	8.032	0.301	0.773	0.231	2.412	1.4206
30	8.032	0.273	0.773	0.238	2.109	1.3219
40	8.032	0.266	0.773	0.241	1.966	1.2215
50	8.032	0.217	0.773	0.241	1.881	1.1063
100	8.032	0.194	0.773	0.241	1.847	1.0432
200	8.032	0.114	0.773	0.241	1.776	0.9871
300	8.032	0.084	0.773	0.241	1.632	0.9653
500	8.032	0.062	0.773	0.241	1.452	0.9442
1000	8.032	0.051	0.773	0.241	1.396	0.9223
2000	8.032	0.044	0.773	0.241	1.334	0.9123

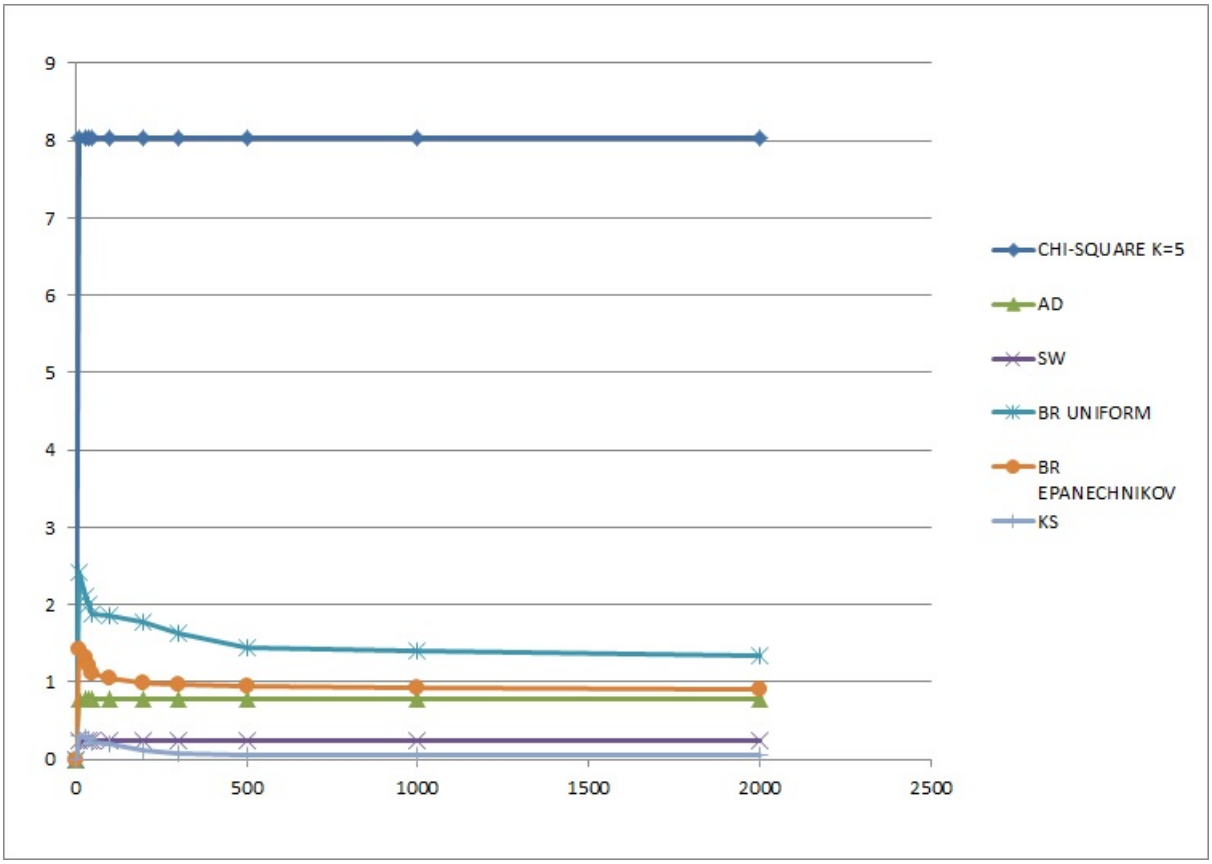

Figure 3.9: Critical value for some goodness of fit tests for different sample sizes.

Table E2: The power of goodness of fit tests: χ^2 , KS, AD, SW and BR for various sample size at $\alpha = 0.05$ under the hypothesis:

H_0 : Mean and Variance of a Normal $\equiv$ Mean and Variance of Weibull(10,2)

H_a : Weibull(10,2)

n	χ^2	KS	AD	SW	BR	
	K=5				Uniform	Epanechnikov
10	-	0.021	0.125	0.013	0.011	0.0082
30	0.109	0.031	0.163	0.021	0.018	0.0098
40	0.131	0.061	0.212	0.053	0.022	0.0311
50	0.187	0.082	0.291	0.123	0.071	0.0921
100	0.213	0.131	0.339	0.402	0.192	0.2123
200	0.271	0.423	0.513	0.692	0.301	0.4132
300	0.341	0.513	0.662	0.779	0.391	0.4861
500	0.532	0.662	0.773	0.865	0.452	0.6122
1000	0.712	0.713	0.845	0.908	0.471	0.6671
2000	0.801	0.873	0.978	0.934	0.501	0.7532

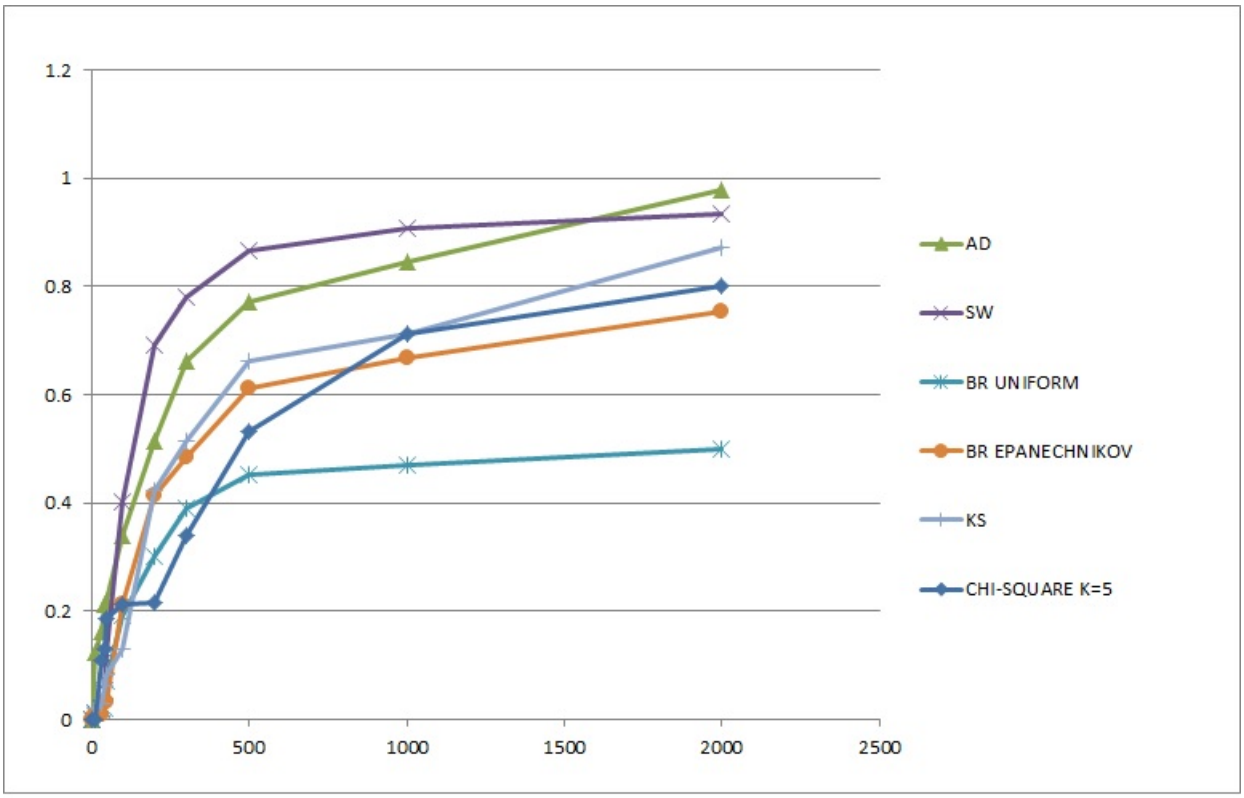

Figure 3.10: The power of the selected goodness of fit tests.

Table F1: The critical values under normal distribution at different sample size (n) and at $\alpha = 0.05$ for the tests: Chi-square (χ^2), Kolmogorov-Smirnov (KS), Anderson Darling (AD), Shapiro-Wilk (SW) and Bickel-Rosenblatt (BR).

n	χ^2	KS	AD	SW	BR	
	K=5				Uniform	Epanechnikov
10	8.032	0.521	0.773	0.231	1.5120	0.8712
30	8.032	0.330	0.773	0.238	1.3612	0.8052
40	8.032	0.189	0.773	0.241	1.3221	0.7791
50	8.032	0.151	0.773	0.241	1.2782	0.7561
100	8.032	0.102	0.773	0.241	1.2241	0.7230
200	8.032	0.078	0.773	0.241	1.1820	0.7092
300	8.032	0.061	0.773	0.241	1.1656	0.6982
500	8.032	0.055	0.773	0.241	1.1442	0.6756
1000	8.032	0.031	0.773	0.241	1.110	0.6691
2000	8.032	0.023	0.773	0.241	1.102	0.6532

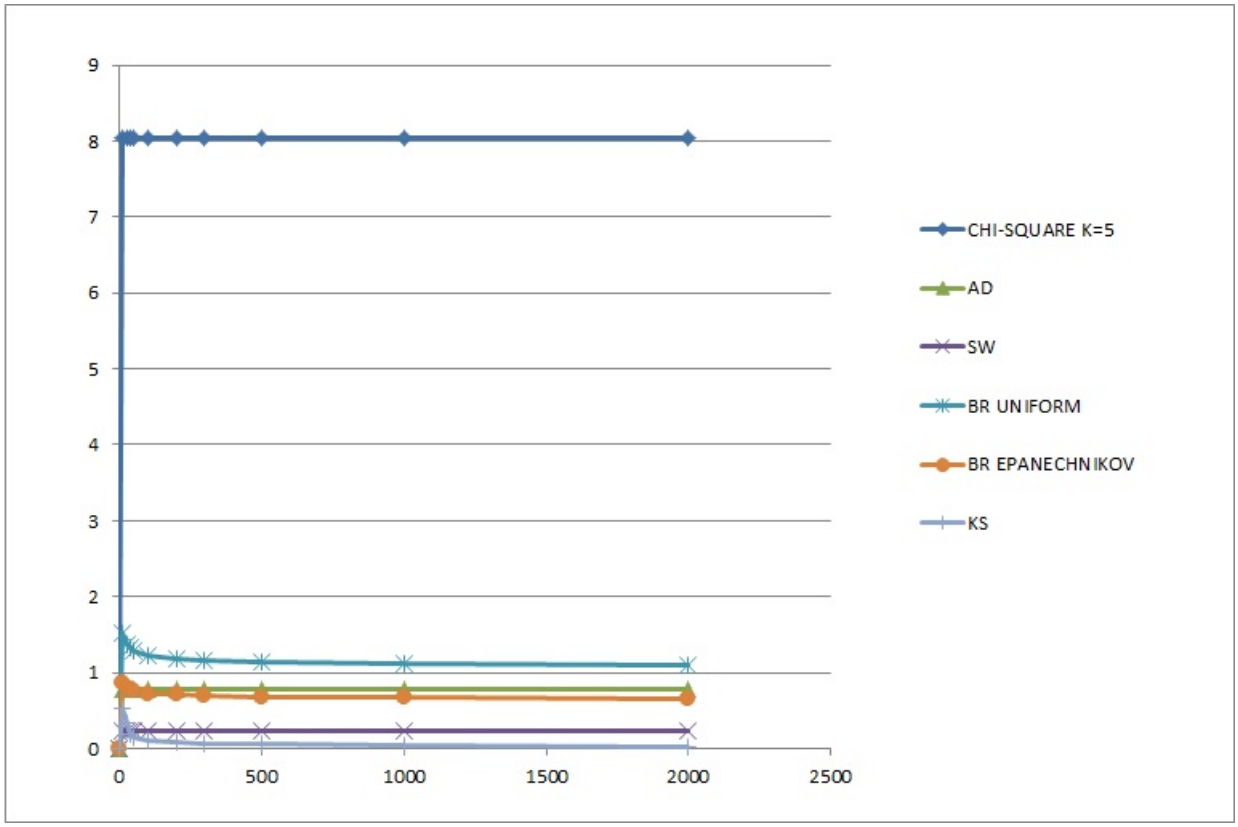

Figure 3.11: Critical value for some goodness of fit tests for different sample sizes.

Table F2: The power of goodness of fit tests: χ^2, KS, AD, SW and BR for various sample size at $\alpha = 0.05$ under the hypothesis:

H_0: Mean and Variance of a Normal $\equiv$ Mean and Variance of $\chi^2(5)$

$$H_a : \chi^2(5)$$

n	χ^2	KS	AD	SW	BR	
	K=5				Uniform	Epanechnikov
10	-	0.0796	0.2312	0.0299	0.0389	0.0564
30	-	0.1321	0.4731	0.1213	0.0532	0.1312
40	0.3362	0.1623	0.7542	0.2764	0.0812	0.1812
50	0.5991	0.4351	0.8452	0.5201	0.1421	0.2721
100	0.8865	0.8435	1.00	0.9711	0.2831	0.6541
200	0.9345	1.00	1.00	1.00	0.5661	0.8561
300	1.00	1.00	1.00	1.00	0.8231	1.00
500	1.00	1.00	1.00	1.00	0.9452	1.00
1000	1.00	1.00	1.00	1.00	1.00	1.00
2000	1.00	1.00	1.00	1.00	1.00	1.00

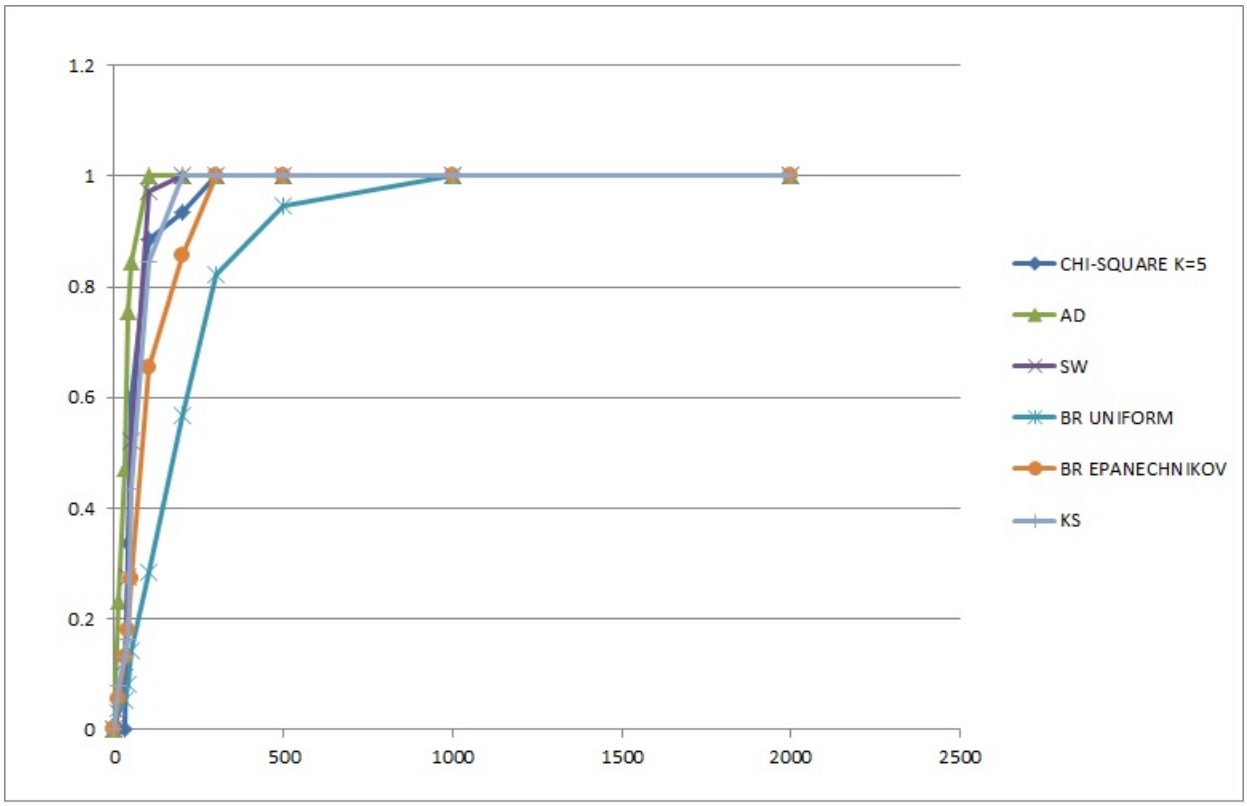

Figure 3.12: The power of the selected goodness of fit tests.

Table G1: The critical values under Uniform distribution at different sample size (n) and at $\alpha = 0.05$ for the tests: Chi-square (χ^2), Kolmogorov-Smirnov (KS), Anderson Darling (AD), Shapiro-Wilk (SW) and Bickel-Rosenblatt (BR).

n	χ^2	KS	AD	SW	BR	
	K=5				Uniform	Epanechnikov
10	9.223	0.311	2.391	0.211	1.712	0.9887
30	9.223	0.234	2.401	0.228	1.641	0.9443
40	9.223	0.194	2.451	0.231	1.572	0.9212
50	9.223	0.166	2.481	0.231	1.508	0.8712
100	9.223	0.112	2.481	0.231	1.423	0.8432
200	9.223	0.093	2.481	0.231	1.391	0.7981
300	9.223	0.077	2.481	0.231	1.344	0.7880
500	9.223	0.612	2.481	0.231	1.331	0.7791
1000	9.223	0.0521	2.481	0.231	1.320	0.7710
2000	9.223	0.0422	2.481	0.231	1.311	0.7692

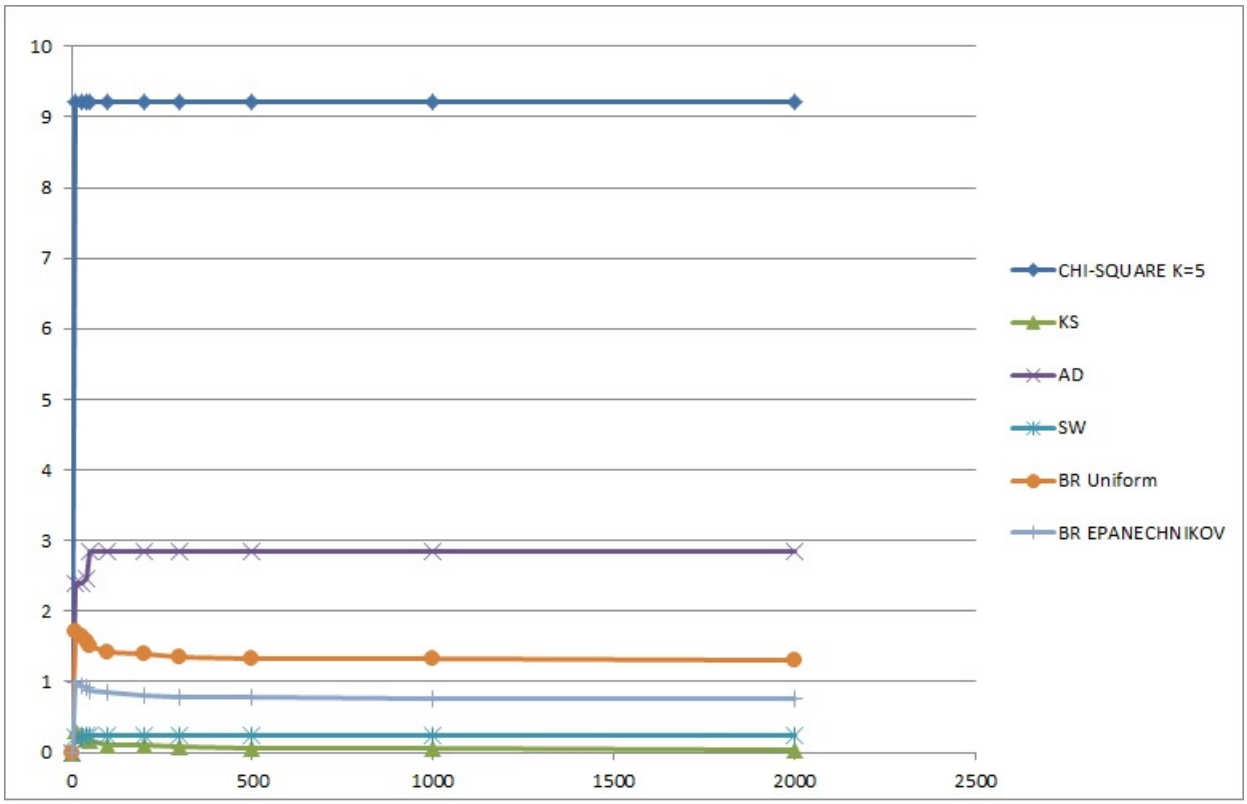

Figure 3.13: Critical value for some goodness of fit tests for different sample sizes.

Table G2: The power of goodness of fit tests: χ^2, KS, AD, SW and BR for various sample size at $\alpha = 0.05$ under the hypothesis:

H_0 : Mean and Variance of a Uniform distribution $\equiv$ Mean and Variance of Exp(0,8)

H_a: Exp(0,8)

n	χ^2	KS	AD	SW	BR	
	K=5				Uniform	Epanechnikov
10	-	0.281	0.901	0.851	0.8534	0.8723
30	0.8234	0.552	0.9232	0.891	0.926	0.9872
40	0.9123	0.723	0.977	0.923	0.983	1.000
50	0.9355	0.832	1.000	0.945	1.000	1.000
100	0.953	0.956	1.000	0.987	1.000	1.000
200	1.000	1.000	1.000	1.000	1.000	1.000
300	1.000	1.000	1.000	1.000	1.000	1.000
500	1.000	1.000	1.000	1.000	1.000	1.000
1000	1.000	1.000	1.000	1.000	1.000	1.000
2000	1.000	1.000	1.000	1.000	1.000	1.000

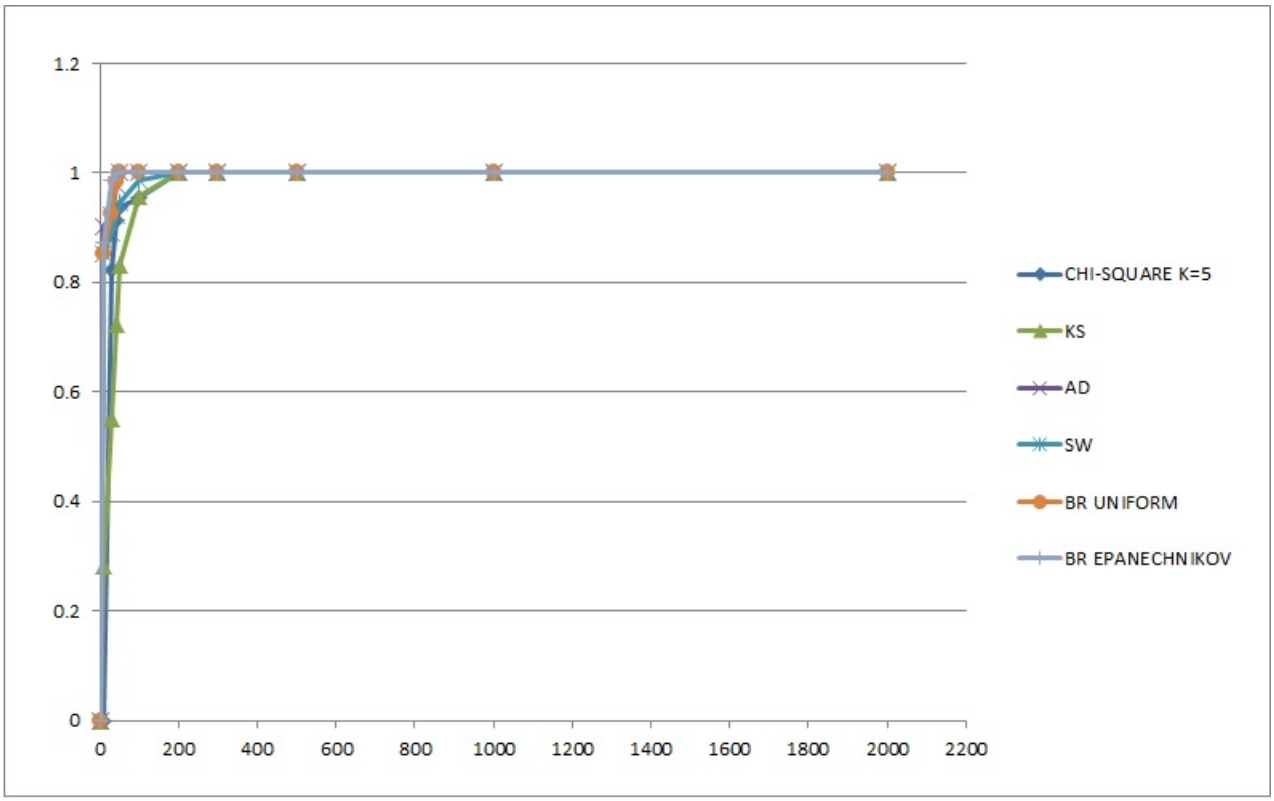

Figure 3.14: The power of the selected goodness of fit tests.

The results shown in Table D2 indicate that the Anderson Darling power test has the best power with respect to all other tests, and this test is able to detect the null hypothesis for samples with small size and the same result can be detected easily in Table E2 where this test detects the null hypothesis for small samples (sample size $\leqslant 100$). Furthermore, all tests in Table D2 show an efficient power to fit the data set for samples with large size (sample size $\geqslant 150$).

In Table E2, the Epanechnikov kernel show higher power than Uniform at sample sizes greater than 40. The results in Table F2 are similar to that in tables D2 and E2 where the Anderson Darling test has the best power compared to all other tests for any sample size. Also, the Epanechnikov kernel choice is better than the Uniform kernel for any sample size can be decided. While in Table G2, results indicated that Bickel-Rosenblatt test has the higher power compared to all other tests and more specifically the power is the highest when Epanechnikov kernel is used rather than Uniform. Furthermore, the Anderson Darling test had higher power compared to Shapiro-Wilk test and Kolmogorov-Smirnov test for any sample size.

Conclusion:

In the real world of statistics, the parameters of any distribution are not observed and should be estimated based on the given data. Therefore, the proposed distributions should fit the data set to facilitate the estimation of the parameters: this can be done using some statistical tests to explore the fitting of the data to the proposed distribution. The estimation process of the unknown parameters may disturb the power of the fit tests. However, it is not easy to assign the most powerful goodness of fit test for all cases. In general, the power of goodness of fit test is highly affected by some common factors such as: the sample size, the type of the distribution being tested and the significance level and also the alternative distribution.

In this simulations study and in general, it is clearly noted that the power of the goodness of fit test increased as the sample size increased. However, for the symmetric distributions against the parametric alternative distribution, the Anderson Darling test has the best power compared to the other selected tests and the Chi-square test has the lowest power even though the Anderson Darling test has low power for small sample size.

Bibliography

[1] Arnold, T. B., Emerson, J. W. (2011). *Nonparametric goodness-of-fit tests for discrete null distributions*. R Journal, 3(2).

[2] Bachmann, D., Dette, H. (2005). *A note on the Bickel–Rosenblatt test in autoregressive time series*. Statistics probability letters, 74(3), 221-234.

[3] Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (revised ed.): New York: Academic Press.

[4] Franke, T. M., Ho, T., Christie, C. A. (2012). *The chi-square test: Often used and more often misinterpreted*. American Journal of Evaluation, 33(3), 448-458.

[5] Huang, L. S. (1997). *Testing goodness-of-fit based on a roughness measure*. Journal of the American Statistical Association, 92(440), 1399-1402.

[6] Kolassa, J. E. (2020). *An introduction to nonparametric statistics*. Chapman and Hall/CRC.

[7] Liu, T. (2016). *Power Comparison of Some Goodness-of-fit Tests.*

[8] Razali, N. M., Wah, Y. B. (2011). *Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests*. Journal of statistical modeling and analytics, 2(1), 21-33.

[9] Romeu, J. L. (2003). *Anderson-Darling: a goodness of fit test for small samples assumptions*. RAC START.

[10] Saculinggan, M., Balase, E. A. (2013, April). *Empirical power comparison of goodness of fit tests for normality in the presence of outliers*. In Journal of Physics: Conference Series (Vol. 435, No. 1, p. 012041). IOP Publishing.

[11] Shi, W. (2014). *An Alternative Goodness-of-fit Test for Normality with Unknown Parameters*.

[12] Tenreiro, C. (2007). *On the finite sample behavior of fixed bandwidth Bickel–Rosenblatt test for univariate and multivariate uniformity*. Communications in Statistics—Simulation and Computation®, 36(4), 827-846.

[13] Weglarczyk, S. (2018). *Kernel density estimation and its application*. In ITM Web of Conferences (Vol. 23, p. 00037). EDP Sciences.

[14] Yazici, B., Yolacan, S. (2007). *A comparison of various tests of normality*. Journal of Statistical Computation and Simulation, 77(2), 175-183.

[15] Zhang, J. (2001). *Powerful goodness-of-fit and multi-sample tests*. Toronto: York University.

[16] Zucchini, W., Berzel, A., Nenadic, O. (2003). *Applied smoothing techniques*. Part I: Kernel Density Estimation, 15, 1-20.