Empirical Power Comparison Of Goodness of Fit Tests for Normality In The Presence of Outliers

This content has been downloaded from IOPscience. Please scroll down to see the full text.
(http://iopscience.iop.org/1742-6596/435/1/012041)

View the table of contents for this issue, or go to the journal homepage for more

Download details:

IP Address: 52.27.131.84
This content was downloaded on 14/07/2015 at 04:02

Please note that terms and conditions apply.
Empirical Power Comparison Of Goodness of Fit Tests for Normality In The Presence of Outliers

Mayette Saculinggan\textsuperscript{1}, Emily Amor Balase\textsuperscript{2}

Zone 1, Cogon El Salvador, Misamis Oriental, 9017 Philippines
\textsuperscript{1}Electronics Engineering Department, Capitol University, Corrales Extension, Cagayan de Oro, Philippines
\textsuperscript{2}Department of Mathematical Sciences, Mindanao University of Science and Technology, Cagayan de Oro City, Philippines

E-mail : maia0812@yahoo.com\textsuperscript{1}, stat_emily@yahoo.com\textsuperscript{2}

Abstract. Most statistical tests such as t-tests, linear regression analysis and Analysis of Variance (ANOVA) require the normality assumptions. When the normality assumption is violated, interpretation and inferences may not be reliable. Therefore it is important to assess such assumption before using any appropriate statistical test. One of the commonly used procedures in determining whether a random sample of size n comes from a normal population are the goodness-of-fit tests for normality. Several studies have already been conducted on the comparison of the different goodness-of-fit tests for normality. This paper compares the power of six formal tests of normality: Kolmogorov-Smirnov test (see [3]), Anderson-Darling test, Shapiro-Wilk test, Lilliefors test, Chi-Square test (see [1]) and D’Agostino-Pearson test. Small, moderate and large sample sizes and various contamination levels were used to obtain the power of each test via Monte Carlo simulation. Ten thousand samples of each sample size and contamination level at a fixed type I error rate were generated from the given alternative distribution. The power of each test was then obtained by comparing the normality test statistics with the respective critical values. Results show that the power of all six tests is low for small sample size (see, for example [2]). But for \( n = 20 \), the Shapiro-Wilk test and Anderson – Darling test have achieved high power. For \( n = 60 \), Shapiro-Wilk test and Liliefors test are most powerful. For large sample size, Shapiro-Wilk test is most powerful (see, for example [5]). However, the test that achieves the highest power under all conditions for large sample size is D’Agostino-Pearson test (see, for example [9]).

1. Introduction

Statistical tests require the normality assumption. Parametric statistical procedures assume a normal distribution of the data therefore it is important to assess such assumption before using any appropriate statistical test.

In any research, the use of different shapes of probability distributions plays a big role in achieving the objective of the study. The actual data must be checked to determine if the observed frequencies in an experiment correspond to the probabilities in a model of the experiment. This is called best fit.

There are different goodness of fit tests available in the literature. The most common is the chi-square goodness of fit test. For Ronneu (2003), this type of test must be used for large sample data. The idea behind the chi-square goodness-of-fit test is to see if the sample comes from the population...
with the claimed distribution. Another way of looking at it is to ask if the frequency distribution fits a specific pattern.

Another method of checking goodness of fit is the Anderson–Darling Test. This is used to test if a sample of data came from a population with a specific distribution. It is a modification of the Kolmogorov-Smirnov (K-S) test and gives more weight to the tails than does the K-S test. The Anderson-Darling test makes use of the specific distribution in calculating critical values. This has the advantage of allowing a more sensitive test and the disadvantage that critical values must be calculated for each distribution.

Another Goodness of Fit Tests is the Shapiro–Wilk Test, and it compares a set of measures against the Normal distribution. It may be used before performing parametric tests to ensure that the data being used follow a Normal distribution. Results on the study of Keskin(2006) showed that Shapiro-Wilk test is the most powerful tests and be used for testing normality of data. The Shapiro-Wilk is an improvement on the more general Kolmogorov-Smirnov curve-fitting algorithm.

One of the most well-known modifications of the Kolmogorov-Smirnov test for goodness of fit is generally referred to as the Lilliefors test for normality. This test was developed independently by Lilliefors (1967) and by Van Soest (1967). The null hypothesis for this test is that the error is normally distributed; that is there is no difference between the observed distribution of the error and a normal distribution. The alternative hypothesis is that the error is not normally distributed.

Different tests of normality often produce different results. Some tests reject while others fail to reject the null hypothesis of normality. The contradicting results are misleading and often confuse practitioners. Several studies have already been conducted on comparing tests for normality. The study of Razali et al. (2011) compared the power of four formal tests of normality. The power comparison was obtained via Monte Carlo simulation of sample data generated from alternative distributions that would follow symmetric and asymmetric distributions. The power of each test was then obtained by comparing the test of normality statistics with the respective critical values. Results in the study showed that Shapiro – Wilk test was the most powerful normality test, followed by Anderson – Darling test, Lilliefors test and Kolmogorov-Smirnov test. However, the power of all four tests was still low for small sample size.

Another study conducted by Oztuna et al. (2006) aimed to compare the four different normality tests. Results showed that the most powerful test for normal distributions was given by Jarqua-Bera and for non-normal distributions, Shapiro – Wilk test yielded the most powerful results. In that study, Oztuna et al. (2006) concluded that Jarqua-Bera test was superior for normal and standard normal distributions whereas for non normal distributions Shapiro – Wilk achieved sufficient power at the smaller sample size.

An additional study by Mendes et al. (2003) compared Shapiro-Wilks, Lilliefors and Kolmogorov-Smirnov Tests for Type I error and also the power of the three tests. For all different sample sizes and distributions, Shapiro-Wilks revealed the most powerful results, followed by the Lilliefors test. Kolmogorov-smirnov test results were the weakest among all three tests. Finally, results showed that all three tests were the most powerful when ran on data with exponential distribution.

Furthermore, the study of Shahabuddin (2009) has found out that in Simple Random Sampling (SRS), the FKS which is a proposed modification of the Kolmogorov-Smirnov test found to perform almost as powerful as Anderson Darling tests (AD) in SRS. However, in all the cases considered, FKS has always outperformed the original Kolmogorov-Smirnov and Cramer-von-Mises (CV) tests.

Moreover, Keskin (2006) conducted a study where it was found out that Shapiro-Wilk test was the most powerful tests and could be suggested for testing normality of data. It was also concluded that performance of the normality tests was greatly affected by the distribution type and sample size.

Another study by D’Agostino (1990) showed that D’Agostino – Pearson K² statistics are powerful and informative tests. The tests used in the study were based on $\sqrt{b_1}$ and $b_2$ which showed that the two are excellent and powerful tests. The study recommended that for all sample sizes $\sqrt{b_1}$ and $b_2$ should be computed and examined as descriptive statistics. For all sample sizes $n \geq 9$, tests of
hypotheses could be based on these two tests. In particular, for \( n > 50 \) Shapiro-Wilk test should not be used instead D’Agostino-Pearson \( K^2 \) test would be used as the test of choice. This justification is not only because of their fine power but also because of the information supplied on non-normality.

The different studies on goodness-of-fit comparisons have included at most only four tests and have limited the coverage of the research to uncontaminated data with one level of significance.

In this study, the power of the six tests for various degrees of contamination and sample sizes in the presence of outliers is assessed and compared via Monte Carlo simulations. This method is used to determine how contamination of outliers in the tested data affects the sensitivity, performance and reliability of the six tests in identifying if it is normal or non normal.

2. Methods

In this study, Monte Carlo procedure was utilized to assess power of goodness-of-fit tests for various sample sizes and contamination levels. The following were the steps adopted to achieve the objectives of the study.

a. Generate data from normal distributions with a certain proportion of contaminants called outliers for various sample sizes.

b. Set the contamination level and the level of significance.

c. Evaluate the power of chi-square, Kolmogorov-Smirnov, Anderson-Darling, Shapiro-Wilk, Lilliefors and D’Agostino-Pearson test statistics in testing if a random sample of \( n \) independent data came from a normal distribution with mean \( \mu \) and variance \( \sigma^2 \) in the context of a Tukey’s contaminated normal model as follows:

\[
F(x) = (1 - \varepsilon)N(\mu, \sigma^2) + \varepsilon N(\mu^*, \sigma^{2*}) \quad \text{and} \quad 0 < \varepsilon < 1
\]

Where \( \mu = \sigma^2 = \mu^* = \sigma^{2*} \)

For this study the null and alternative hypotheses were:

- \( H_0 \): The distribution is standard normal
- \( H_1 \): The distribution is not standard normal

d. Determine the best normality tests under different conditions considering the extent of contamination of outliers. In order to obtain the simulated power of the six normality tests, at \( \alpha = 1\%, 5\% \), and 10\% for each sample size, a total of 10,000 samples will be drawn from normal distribution with specified mean \( \mu \) and variance \( \sigma^2 \) and contaminated with outliers fixed at \( \varepsilon \).

2.1. Power Study Process

A Monte Carlo simulation was accomplished with 10000 iterations, considering various sample sizes and different contamination levels. The steps in the process were:

2.1.1. Random data were generated from standard normal distribution as uncontaminated samples. Samples were then contaminated with contaminant given by Normal (5, 100) where \( \varepsilon = .01, .05, .10, .15 \) and .20 for various sample sizes classified as small \( (n=10, 20) \), moderate \( (n=30,60) \) and large \( (n=100, 1000) \).

2.1.2. The statistic of all goodness-of-fit tests considered in this study was calculated for every iteration.

2.1.3. The value of the test statistic was then compared to the critical value at different levels of significance \( \alpha = .10, .05 \) and .01.

2.1.4. The number of times the value of the test statistic exceeds the critical value was counted for at \( \alpha \) level of significance. Exceeding the critical value is equivalent to rejecting \( H_0 \) at the corresponding significance level. The process was repeated at various sample sizes and contamination level.
3. Highlights of findings and discussion/data presentation

The result of the power of Goodness-of-Fit tests was presented according to various sample sizes with different degrees of contamination. The degree of contamination was generated based on Tukey’s contaminated normal model:

\[ F(x) = (1 - \varepsilon)N(0, \sigma^2) + \varepsilon N(\mu^*, \sigma_{\ast}^2) \quad \text{and} \quad 0 < \varepsilon < 1, \]  

(3.1)

Where \( \varepsilon = 0.01, 0.05, 0.10, 0.15 \) and \( 0.20 \) using \( n = 10, 20, 30, 60, 100 \) and \( 1000 \) to represent small, moderate and large sample sizes. Every sample size was tested 1000 times with the applicable GOF tests for each degree of contamination with test of hypotheses carried out at the 5% level of significance. The figure below showed the behavior of each test for every significance level.

The power of each Goodness-of-Fit tests for various sample sizes and degree of contamination were evaluated to identify the behavior of each tests. The KS test as shown in Figure 1, provides low power for all sample sizes. It happened because KS test is a distance test and by research it is considered an insensitive test for power. Although it increased power as sample size \( n \) increases but still it must reached the 80% level to be considered statistically powerful test for normality.

The Anderson-Darling test at a significance level of .01 performed well especially on moderate sample size. The test showed to be statistically powerful at small samples when the data were highly contaminated with outliers. These results also showed that the test would only have its highest power for very large sample size at \( n = 1000 \). The power of Anderson-Darling test increased when sample size and contamination was also increased.

![Figure 1. Power versus Degree of Contamination for Kolmogorov-Smirnov Test.](image-url)
Figure 2. Power versus Degree of Contamination for Anderson-Darling test.

The test was able to show power starting at \( n = 60 \) at 0.05 contamination level while for smaller sample size power was shown only when the data were highly contaminated with outlier starting at 10 percent and higher. At \( n = 1000 \) the test is already statistically powerful even with 0.01 contamination. For small and moderate samples, the Anderson-Darling test showed power when it was contaminated with at least 5% outliers and showed 100% power only when \( n = 1000 \) for all contamination levels.

Shapiro-Wilk test was able to detect non-normality of the data even at very low contamination level. The result in the figure below showed that Shapiro-Wilk is a powerful test for moderate and large sample sizes. For small sample size, this test would only show power when the data was contaminated with 5% outliers. The Shapiro-Wilk at \( \alpha = 5\% \), also showed similar behavior as with Anderson-Darling although at smaller samples like when \( n = 20 \) the test was statistically powerful when contaminated with at least 5% of outlier.

Figure 3. Power versus Degree of Contamination for Shapiro-Wilk Test.

As recommended in other studies, Liliefors test was used only for moderate to large sample data. The result presented in the figure below showed that the test has power when used for moderate samples and only when the data is at least contaminated with 5% outliers. However, with \( n = 1000 \), Liliefors could detect non-normality of the data 100% of the time even when it was contaminated with 1 percent outliers.
Power of Liliefors test to detect non-normality of data for moderate and large sample sizes was when the data was contaminated with at least 5% outlier. It also revealed power of 100% when sample size is very large like $n = 1000$ for all contamination levels.

As shown in Figure 5 below, power of Chi-Square test increased as sample size and contamination were also increased. Still, the power of this test was low and could not be considered statistically powerful test for normality. This result just shows that Chi-Square had the lowest performance in terms power for large sample data.

Chi-Square test had low power levels and none of the conditions reached a power of 80%. It can be seen in Figure 5 above that as the sample size and contamination is increased, power of the test also increased. But still, Chi-Square was not powerful enough to detect non-normality in the data tested and cannot be considered as omnibus test for normality.
The D’Agostino which was also tested for large sample size has shown high power even at \( n = 100 \). At 1% significance level, the D’Agostino-Pearson test showed to be a powerful test for large sample data. Starting at 1% contamination up to 20% contamination level for \( n = 100 \) and \( n = 1000 \) the simulation result showed that for almost all conditions this test could detect non-normality 100% of the time.

**Figure 6.** Power versus Degree of Contamination for D’Agostino-Pearson Test.

As shown in the figure above, the D’Agostino-Pearson test has performed well for all contamination levels and at all sample size \( n \). Almost all results showed 100% power except for \( n = 100 \) at 0.01 level of contamination which could detect only 82.44% of the time on non-normality of data. Opposite to Chi-Square, the D’Agostino-Pearson has shown to be statistically powerful test for all conditions which makes it a good choice for large sample data.

The Kolmogorov-Smirnov and Chi-Square test gave poor power to detect non-normality of data which was consistent with the results of the study conducted by Agostino (1986), Mendes et al. (2003) and Shahabuddin et al. (2009). A modification of the Kolmogorov-Smirnov test was studied by Drezner et al. (2008) in order to enhance the behavior of the test. The result of his study showed that the modified KS test was able to detect the difference from normality at \( n = 200 \).

The result of this study showed that, for small and moderate sample sizes, the Shapiro – Wilk gave the most powerful results. Shapiro-Wilk was also most powerful test based on the result of the study conducted by Shapiro, et al. (1986), Mendes et al. (2003), Oztuna et al. (2006), Guner et al. (2009) and Razali et al. (2011). For large sample size data, the D’Agostino-Pearson test outperformed the rest of the test because it was able to give 100% power on almost all sample size and all contamination levels and even at different significance level. This has been similar to the study conducted by D’Agostino (1990), which showed that D’Agostino-Pearson test is an omnibus test for normality.

To be able to show the behavior of each tests per sample, the different GOF tests used for each sample size were compared and discussed below. As shown in figure 7, at \( \alpha = 5\% \) small samples with 1% contaminated data, none of the tests were able to show high power.

For moderate sample size, only Shapiro-Wilk test showed power at \( n = 60 \). D’Agostino-Pearson and Shapiro-Wilk test were able to show power at \( n = 100 \). For \( n = 1000 \), all tests except for Chi-Square were able to detect non-normality of the data 100% of the time.
Figure 7. Comparison of Power of Goodness-of-Fit Tests for various Sample Sizes ($\alpha = 5\%$, $\varepsilon = 1\%$).

When the data were contaminated with 5\% outlier at $\alpha = 5\%$, only Shapiro-Wilk showed power for small and moderate sampled data at $n = 20$ and $n = 30$ respectively. However, for $n = 60$, three tests showed power except for Kolmogorov-Smirnov test.

All tests showed power and even showed 100\% power for very large data except for Chi-Square test.

Figure 8. Comparison of Power of Goodness-of-Fit Tests for various Sample Sizes ($\alpha = 5\%$, $\varepsilon = 5\%$).

At 10\% contamination of the data, small samples at 5\% significance level have shown that for $n = 10$ only Shapiro-Wilk showed power. When the sample size was increased to $n = 20$, only Kolmogorov-Smirnov showed low power for the tested data out of the three tests. The same also for moderate sampled data, only Kolmogorov-Smirnov did not show power of the test as shown in the
Figure below. It also observed that power of all three tests increased as sample size was increased to 60.

All tests showed power for large sample data excluding Chi-Square test that only showed 14.58 and 21.54 power respectively for each sample size as shown in figure 9.

![Figure 9. Comparison of Power of Goodness-of-Fit Tests for various Sample Sizes ($\alpha = 5\%$, $\varepsilon = 10\%$).](image1)

When the contamination was increased to 15%, at $n = 10$ all tests failed to show high power to be considered statistically powerful. But when $n$ was increased to 20, the result showed that all three test showed power. For moderate sampled data, only Kolmogorov-Smirnov did not show power of the test. All tests have shown to be statistically powerful by providing 100% power at both sample size except for Chi-Square which could detect only 44.17% of the time to detect non-normality of the tested sample.

![Figure 10. Comparison of Power of Goodness-of-Fit Tests for various Sample Sizes ($\alpha = 5\%$, $\varepsilon = 15\%$).](image2)

At $n = 1000$, only Chi-Square test did not show power of the tested data but all four GOF tests showed 100% power.
When the contamination was increased to 20%, almost all tests had high power even at \( n = 10 \) except for Kolmogorov-Smirnov test.

The same with small samples, moderate sample data have shown that three GOF tests showed power of the tests except for Kolmogorov-Smirnov test and would even have 100% power at \( n = 60 \).

Only Chi-Square test showed low power of the test for large sample size at 20% contamination as shown in the figure below.

![Figure 11. Comparison of Power of Goodness-of-Fit Tests for various Sample Sizes (\( \alpha = 5\% \), \( \varepsilon = 20\% \).)](image)

4. Conclusions

Based on the generated results, the different types of Goodness-of-Fit Tests showed different power in various contaminations. For large data samples, the D’Agostino Pearson test has shown to be the appropriate test as it would provide sensitivity for all conditions. For small sample data, Shapiro-Wilk had power at \( n = 20 \) for a contamination of 0.05. When the number of samples is moderate, Shapiro-Wilk or Liliefors test has shown more power as compared to Anderson-Darling and Kolmogorov-Smirnov test. Also as evident in the results, when contamination level and sample size were increased the power had also increased.

In this study, Shapiro-Wilk and D’Agostino-Pearson tests have shown to be a powerful test for normality. There were two tests which have shown poor performance in terms of power. These tests were Kolmogorov-Smirnov test and Chi-Square tests. This is because both tests were distance tests and were considered to be inefficient in identifying non normality in the data. Even with outliers introduced into the tested data, the results of the study were supported by the conclusion also made by Agostino and Razali that D’Agostino-Pearson test has been a powerful test and Kolmogorov-Smirnov test to have poor power for normality (see, for example [5] [9]).

References

[1] Romeu J.L. 2003 Selected Topics in Assurance Related Technologies: The Chi-Square : A Large-Sample Goodness of Fit Test 10 4


[6] Oztuna D. et. al. 2006 Investigation of Four Different Normality Tests in Terms of Type I Error Rate and Power under Different Distributions *TUBITAK Turk J Medical Science*

[7] Mendes M., Pala A. 2003 Type I Error Rate and Power of Three Normality Tests Pakistan *Journal of Information and Technology*
