# Efficiency Comparisons of Normality Test Using Statistical Packages

**Umaporn Chantasorn**
Department of Statistics, Faculty of Science,
King Mongkut's Institute of Technology Lat Krabang, Lat Krabang, 10720
Bangkok,Thailand.

## Abstract

The purpose of this study was to compare the efficiency of commercial statistical packages in testing normality. The six (6) tests being studies were Kolmogorov, Lilliefors and Shapiro-Wilk available in SPSS14 and Kolmogorov, Anderson-Darling and Ryan-Joiner available in MINITAB 14. The data for this study was obtained from simulation, by the method of Monte Carlo, under conditions of a normal distribution and slightly different from the normal distribution, by using Calc-Random Data menu of MINITAB 14. In each situation, 500 iterations were carried out with different sample size: 10, 20, 30, 50 and 100. Comparison of type I error rate and empirical power among the six test statistics were also made. It was found that Ryan-Joiner Test in MINITAB 14 had the highest empirical power in all cases and sample sizes, at a significance level of 0.10. The sample size of 100, in particular, showed the empirical power at almost 1. It also had an ability to control probability of Type I error in almost all situations under the criteria of Cochran. Nonetheless, the K-S test in MINITAB 14, appearing under the appellation of the K-S test was in fact the Lilliefors test.

**Keywords:** Normality test, Kolmogorov test, Lilliefors test, Shapiro-Wilk test, Anderson-Darling, Ryan-Joiner, SPSS, MINITAB.

## 1. Introduction

Theoretically, statistical analysis can be categorized into two board groups which are Parametric Statistics and Non-Parametric Statistics. The former, however, has a limitation of a requirement for a Normal distribution of random variable (1). That is to say a Normal distribution is vital for this kind of analysis. In that case, any researcher that wishes to employ this statistical process must firstly verify the existence of a Normal distribution to ensure that parametric statistics can be carried out. In the case of no Normal distribution, a transformation of data is recommended prior to re-examination to find the Normal distribution. Finally, if the preliminary assumption is not true then Non-Parametric Statistics should be employed. (2)

There are several ways to test the Normal distribution, ranging by degree of easiness, from graphical to statistical tests (3). R.A. Fisher (1923 – 1930), for example, was the pioneer who derived the first statistical test to examine the Normal distribution when mean and variance of population were unknown. The mentioned test was referred to as Standard Third Moment ($\sqrt{b_1}$) and Standard Fourth Moment ($b_2$). Numerous other statistical tests were also developed such as Anderson & Darling

(AD) (4) or Kolmogorov – Smirnov (K-S) (5) or Shapiro – Wilk (S-W) (6), all of which are widely recognized among researchers. Additionally, new statistics have been found such as Chen & Shapiro (1995) and Zhang (7)(1999) as well as the development of a statistical test based on skewness and kurtosis to test out the Normal distribution, known as D'Agostine (1990) and Park (1999).

Available statistical tests found on commercial statistical packages include Kolmogorov-smirnov (K-S), Shapiro-Wilk (S-W), Anderson-Darling (AD), Lilliefors, Cramer-von Mises, and Shapiro-Francia. In Thailand, however, the most widely used statistical packages are SPSS and MINITAB with a variety of Normal distribution statistical tests. SPSS offers 1-sample K-S, Lilliefors, and S-W. These statistics are defined as:

$$D_i = \hat{F}(X_{i-1}) - F_0(X_i)$$
$$\tilde{D}_i = \hat{F}(X_i) - F_0(X_i) \quad i = 1,\ldots N$$

**1K-S** $Z = \sqrt{N} \quad \max_i \left\{ |D_i|, |\tilde{D}_i| \right\}$

**Lilliefors** $D_a = \max\{D_+, D_-\}$

Where $F_0(X_i)$ is the theoretical cumulative distribution function of the normal distribution.

$\hat{F}(X_i)$ is the empirical cumulative distribution function.

$$D_+ = \max_i \left\{ \hat{F}(y_i) - F(y_i) \right\}$$
$$D_- = \max_i \left\{ F(y_i) - \hat{F}(y_{i-1}) \right\}$$

**S-W** $$W = \frac{\left( \sum_{i=1}^{W_S} a_i X_i \right)^2}{\sum_{i=1}^{W_S} (x_i - \bar{x})^2}$$

where

$$\bar{x} = \frac{\sum_{i=1}^{W_S} x_i}{W_S}$$

$$a_1^2 = a_{W_S}^2 =$$
$$\begin{cases} \dfrac{\Gamma(W_S/2)}{\sqrt{2}\Gamma((W_S + 1/2))} & \text{if } 5 \le W_S \le 20 \\ \dfrac{\Gamma(W_S + 1/2)}{\sqrt{2}\Gamma((W_S/2 + 1))} & \text{if } W_S > 20 \end{cases}$$

$$a_1 = -\sqrt{a_1^2}, \quad a_{W_S} = \sqrt{a_{W_S}^2}$$

$$a_i = (2/c)m_i, \quad i = 2,\ldots, W_S - 1$$

While MINITAB offers AD, RJ (Ryan-Joiner which is said to be similar to Shapiro-Wilk) and K-S. These statistics are defined as:

**AD**

$$A^2 = -N - (1/N)\sum (2i-1)(\ln F(Y_i) + \ln(1 - F(Y_{N+1-i})))$$

where
F is the cumulative distribution function of the normal.

**Ryan-Joiner**

$$R_p = \frac{\sum y_i b_i}{\sqrt{s^2 (n-1) \sum b_i^2}}$$

where $y_i =$ are ordered observations

$b_i =$ normal score of your ordered data

$s^2 =$ sample variance

**K-S**

$$D = \max_{1 \le i \le n} \left| F(y_i) - \frac{i}{N} \right|$$

where F is the theoretical cumulative distribution function of the normal distribution.

Theoretically, K-S should be used with small sample size to derive critical value from the exact distribution (8),(9) besides,

complete parameter values are required in a null hypothesis (for example, $H_0$ : X with Normal distribution where mean = 100 and variance = 25). However, this statistical test is not recommended for other cases (2). The Nonparametric Test menu in SPSS, however, allows user to omit parameter values. It is noted that the mean and variance values used are estimators. The Help menu also explains that with the aforementioned process, the distribution of K-S statistical test will alter, thus the Explore menu is recommended in lieu. However, users with inadequate statistical knowledge using that method may choose an inappropriate statistical test. In such case, the Nonparametric Test menu should alter its algorithm to allow the user to identify parameters. Otherwise, the Normal distribution test should not be allowed in this menu. Likewise, MINITAB dose not place importance on this issue nor offer further explanation.

Theoretically, the K-S statistical test is appropriate only for Continuous distribution verification (2), yet in SPSS, the user can test a Poisson distribution which is a discrete distribution. The question here is the degree of reliability of the conclusion.

The Lilliefors statistical test has been developed to replace the K-S test. It allows the omission of complete parameter values (i.e. mean and variance) in the null hypothesis. Generally, researchers are not familiar with this statistical test. Cases for distribution of statistical tests also vary, such as with mean value but no variance, or no value at all. In SPSS, however, users can use only one scenario which is “no value”. The question here is accuracy of calculation of p-value from the distribution by SPSS.

There should be correction for the statistical test for both K-S and Lilliefors tests to see differences in the kurtosis of a distribution on two lines of cumulative probability function on the right hand side using the following formula which would induce a true difference:

$$T = \sup_{1 \le i \le r} \left\{ \sup \left[ \left| S(X_i) - F^*(X_i) \right|, \left| S(X_{i-1}) - F^*(X_i) \right| \right] \right\}$$

Both statistical tests have been used in various statistical packages but under names that could cause confusion for users. For instance, despite the name K-S statistics, S-Plus is in fact Lilliefors statistics. While in the R and Matlab packages, both of the tests have been separated so users have to choose (10). SPSS refers to the K-S statistical test as 1-Sample K-S (from menu Analyze-Nonparametric Test) and refers to Lilliefors statistics as K-S with a note to indicate that it is Lilliefors (from menu Analyze-Descriptive Statistics-Explore-Plots-Normality Plots with test). In MINITAB, it is referred to as K-S statistics, yet it is not clear which statistical test it really is (no further explanation in Help menu) when choosing from menu Basic-Statistics-Normality Test.

A number of research studies on the Normality test by K-S statistics showed similar findings, suggesting that K-S statistical test should not be used due to its low efficiency (10), (11) despite the use of large sample size and higher significance level ($\alpha$). This is because K-S statistical test normally accepts that the random variable has normal distribution even though other distributions exist. Also other researches found that among the three statistical tests for Normality available in SPSS, K-S statistical test is the least efficient when compared to S-W and Lilliefors tests (12).

As for S-W, most researches found that it has the highest efficiency and could be used in all scenarios (12), (13), (14). SPSS suggests the use of S-W when the sample size is less than 50 which is contradictory with the research findings which found that if the desired power of the test is closer to 1, the sample size should be closer to 100 (12).

The distribution of the AD statistical test, on the other hand, differs from case to case, ranging from available for both mean and variance values, to only one value available to none. In the last case, there should be correction of statistics (11), yet MINITAB provides only the last case scenario. The parameter value is estimated from the sample and MINITAB sets the AD test as the default. This indicates that the package places preference on this test over the others. Should an inexperienced user further employ statistical analysis derived by the AD statistical test, the power of the test value might be lower than that by others.

Based on the characteristics of various statistical tests mentioned above, it is questionable whether SPSS and MINITAB packages have also been aware of the fact or their algorithms have been developed in accordance with the statistical theories. Both packages feature the K-S statistical test despite much research indicating its low efficiency, which may lead to the use of inaccurate analysis by inexperienced users. Moreover, it is unclear whether the K-S statistical test in MINITAB is in fact Kolmogorov-Smirnov or Lilliefors. Both packages also feature one similar test known as SW (SPSS) and RJ (MINITAB), it is still questionable whether the development of the algorithm in both packages has taken into account the true distribution of the statistical test. A further question is which statistics would provide the highest power of the test under any circumstance, and which package would provide the most reliable conclusion so that users could use and completely trust the conclusion.

Relevant research for this study is Shapiro et al (1968) which was the first to study the power of different statistical tests in a normality test. The nine (9) statistical tests studied were Shapiro-Wilk Statistic (S-W), $\sqrt{b_1}$, $b_2$; Kolmogorov-Smirnov Test (K-S); Cramer-von Mises ($W^2$); Anderson-Darling (AD); Durbin (D); Chi-square Test ($\chi^2$); and Studentized Range Test (U) under 12 distributions with differing parameters leading to a total of 45 distributions. The conclusions were as follows:

1. Shapiro-Wilk Statistics worked well under general tests;
2. Testing using Empirical Distribution Function showed low power;
3. Studentized Range Test (U) had high power of the test when the population distribution was symmetric, short-tailed and low power when the population distribution was asymmetric;
4. $\sqrt{b_1}$ and $b_2$ worked well in the test but showed lower power than S-W.

M.A. Stephens (1974) showed results indicating that the study by Shapiro el al employed a critical value for statistical testing using the Empirical Distribution Function in the normality test inappropriately. This was because the critical value was calculated based on the assumption of known mean and variance of the population. Thus, Stephens recalculated the critical value on the assumption of unknown mean and variance of the population.

Somphis Chotiwittayadharakorn (1988) performed a comparison of power of five statistical tests used to test a Normal distribution, namely Chi-square Test ($\chi^2$); Studentized Range Test (U); Shapiro-Wilk Statistics (S-W); Probability Plot Correlation Coefficient Test (r); and Hanna Oja Statistics ($T_1$ and $T_2$) under two significant distributions –normal and non-normal. It was found that for functionality, the S-W statistical test would be more appropriate as it showed high power most of the time and could control Type I error the best.

Seirr (2004) (15) conducted a study on 10 statistical tests for normality by simulating data in sizes ranging from 20 to 100 and 1,000 iterations from differing distribution characteristics such as Bimodal,

Short-tailed, slightly skewed, highly skewed and with kurtosis. In each situation, the empirical, alpha, and power values were studied. It was found that the statistical test following Regression test criteria namely D'Agostino (1972), Shapiro, Royston (aka Shapiro Corrected), Chen and Shapiro and Zhang were best i.e. with highest power in almost all of the distributions. Nonetheless, if the objective was to see whether the distribution had highly symmetrical kurtosis, the statistical test with skewness and kurtosis test criteria should be used, such as D'Agostino (1990) and G-kurtosis, consisting of the first statistical $G^2_w$ and the second statistical $G^{2*}_w$ which could give the highest power. It was also suggested that commercial statistical packages should provide the aforementioned statistical tests.

The statistical analysis of commercial statistical packages found some errors or misleading features in several areas. Kamon, for example, (16,17) (2004), studied the calculation of standard error of estimated mean in factorial design in the case of the mixed effect model of SPSS and found that the result was inaccurate, while SAS gave an accurate result. Kamon also studied the p-value reporting by Chi-square test from the two way contingency table of SPSS and found that the value was one sided probability rather than two sided as reported by the programme. Thus, when further employed, the user should not divide it by two as the Chi-square test was one tailed.

Bergmann, Ludbrook and Spooren (18) studied the outcomes from the Wilcoxon-Mann Whitney statistical test in 11 commercial statistical packages such as SPSS 8.0; StatXact 4.0; SigmaStat 2.03; S-Plus 2000; and SAS 6.12. It was found that the outcomes were varied in several areas such as ties correction, continuity correction, and large sample size correction, as well as having inadequate description of algorithms.

Knüsel (19) studied accuracy of probability of Binominal distribution, Poisson, Hypergeometric, Gamma and Inverse Beta from Excel 2003. It was found that probability of binominal distribution and Poisson values were accurate only when the value of a random variable was in the mid range of the distribution. When the random variable was in the extreme lower tail, Excel 2003 would round it up to zero (0) while Excel 97 provided an accurate answer. It therefore showed that the algorithm development in Excel 2003 still had a flaw even though it was developed on Excel 97. The study by McCullough and Wilson (20) found three statistical techniques, namely Regression Analysis (both Linear and Non-Linear), Random Generator, and all Distributions in Excel 97 were not accurate. It recommended that those statistics should not be analyzed by Excel. The last case was a study by Kusaya (21) which compared the efficiency of data generation by SAS and MINITAB. It was found that both programmes produced similar outcomes for a small sample size. Yet for a large sample size, SAS was recommended.

## 2. Objectives

2.1 To compare the empirical power of three statistical tests in MINITAB and SPSS to find out which test provides the highest empirical power in each situation (only for statistics with ability to control Type I error rate);

2.2 To compare the empirical power of similar statistical tests featured in MINITAB and SPSS i.e. K-S (in MINITAB) and 1-sample K-S aka Lilliefors (in SPSS); and RJ (in MINITAB) and S-W (in SPSS);

2.3 To find out whether K-S statistics in MINITAB is actually K-S or Lilliefors (based on comparison of Type I error rate and empirical power from 1-sample K-S and Lilliefors in SPSS).

## 3. Study Protocol

Simulation of Symmetric Population Distribution in accordance of the characteristics of each situation as follows:

- Normal distribution
- Long tailed distribution
- Distributions with high kurtosis
- Distribution with kurtosis slightly higher than the Normal
- Short tailed distribution

The above mentioned data was generated by MINITAB14 using Calc-Random Data menu based on the finding of Kusaya's research (21), sample size (n) were 10, 20, 30, 50, 100; Iterations = 500 in each situation. Some graphs of a studied distribution when compared to the normal distribution can be seen in Figure 1.

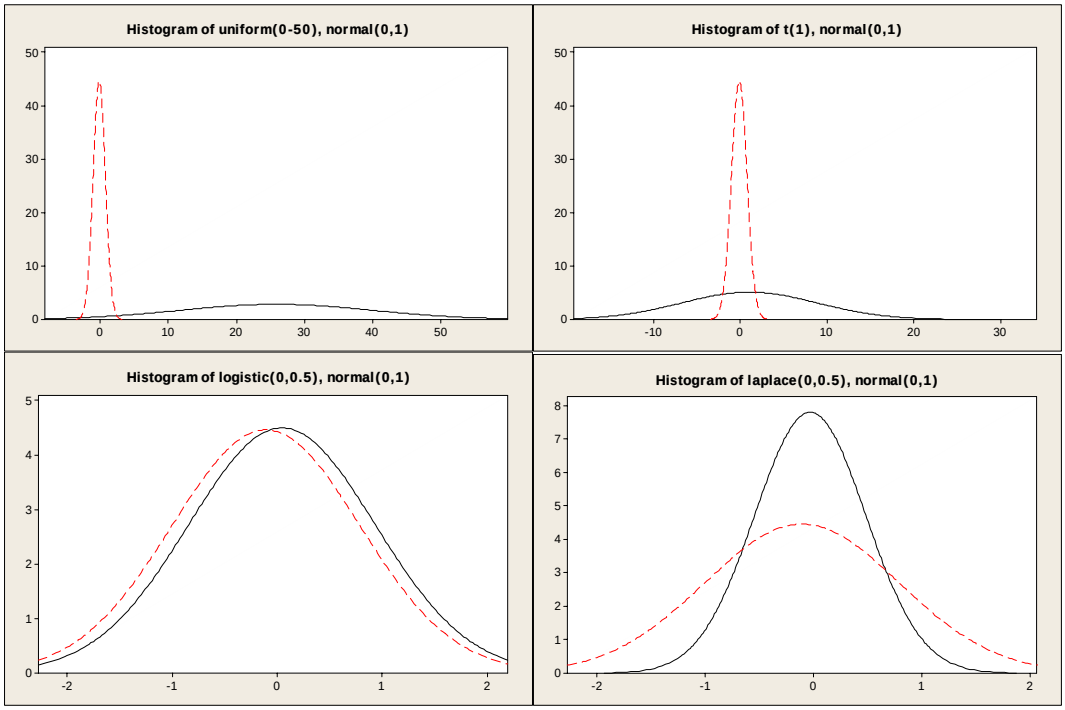


**Figure 1** Studied distribution when compared to the normal distribution

Analyze data obtained by **SPSS** programme using three statistical tests under the following menus:

Analyze

- Descriptive Statistics
- Explore
- Plot
- Normality Plots with test (with Lilliefors and S-W Statistics)

**and**

Analyze

- Nonparametric Test
- 1-sample K-S (with K-S statistics)

The results were statistical test and p-value.

Analyze data obtained (i.e. same data set as above) by **MINITAB** using the menus:

- Basic Statistics

- Normality Test

i.e. AD, RJ and K-S statistics wherein AD was set as default.

The specified value of Type I error was = .05 and .10 to see if it could control such a value, under the Cochran and Bradley criteria.

Analyze the data of population with distributions other than the Normal by statistical test in SPSS and MINITAB, to find the empirical power.

## 4. Research Result

The results from SPSS and MINITAB are presented separately and compared in two aspects – Type I error rate and empirical power as shown in Tables 1-4.

**Table 1** Type I Error Rate of Various Statistical Tests used for Normality test in SPSS14

| Distributions | Statistical Tests | Sample Size | α | |
|---|---|---|---|---|
| | | | .05 | .10 |
| $N(100, 10^2)$ | 1 – Sample K-S | 10 | .0000 | .0000 |
| | | 20 | .0000 | .0000 |
| | | 30 | .0000 | .0000 |
| | | 50 | .0000 | .0020 |
| | | 100 | .0000 | .0000 |
| | Lilliefors | 10 | .0520 | .1200 |
| | | 20 | .0540 | .0920 |
| | | 30 | .0620*C | .1160 |
| | | 50 | .0340*C | .0740*C |
| | | 100 | .0400 | .1060 |
| | S-W | 10 | .0580 | .1120 |
| | | 20 | .0540 | .0900 |
| | | 30 | .0680*C | .1120 |
| | | 50 | .0340*C | .0940 |
| | | 100 | .0540 | .1020 |
| $N(100, 50^2)$ | 1 – Sample K-S | 10 | .0000 | .0000 |
| | | 20 | .0000 | .0020 |
| | | 30 | .0020 | .0000 |
| | | 50 | .0000 | .0000 |
| | | 100 | .0000 | .0040 |
| | Lilliefors | 10 | .0540 | .1200 |
| | | 20 | .0400 | .0980 |
| | | 30 | .0340*C | .0800 |
| | | 50 | .0520 | .1000 |
| | | 100 | .0540 | .1140 |
| | S-W | 10 | .0540 | .1080 |
| | | 20 | .0360*C | .0980 |
| | | 30 | .0440 | .0920 |
| | | 50 | .0540 | .1000 |
| | | 100 | .0500 | .0960 |

**Table 1** Type I Error Rate of Various Statistical Tests used for Normality test in SPSS14 (cont')

| Distributions | Statistical Tests | Sample Size | α | |
|---|---|---|---|---|
| | | | .05 | .10 |
| $N(100, 100^2)$ | 1 – Sample K-S | 10 | .0000 | .0000 |
| | | 20 | .0000 | .0000 |
| | | 30 | .0000 | .0000 |
| | | 50 | .0000 | .0020 |
| | | 100 | .0020 | .0000 |
| | Lilliefors | 10 | .0520 | .0900 |
| | | 20 | .0500 | .1020 |
| | | 30 | .0420 | .0880 |
| | | 50 | .0480 | .1060 |
| | | 100 | .0540 | .1000 |
| | S-W | 10 | .0480 | .1000 |
| | | 20 | .0720*C | .1200 |
| | | 30 | .0560 | .0980 |
| | | 50 | .0560 | .1100 |
| | | 100 | .0420 | .0860 |

* C has a value outside the range specified by the Cochran criteria; *B has a value outside the range specified by the Bradley criteria.

From Table 1, it can be seen that the 1-Sample K-S statistical test was able to control the probability of Type I error in almost all cases of Normal distribution and at $\alpha = .05$ or .10. The result of type I error rate was lower than specified error in all cases and almost all were zero (0). In other words, it could be concluded that K-S statistics always accepted that a random variable has a distribution.

The Lilliefors statistical test was able to control the probability of Type I error. Most values were within the controllable range according to the Cochran criteria. Only 13 percent of the values were outside the range. Most values were less than the lower bound of Cochran, i.e. lower than .04 or .08.

The Shapiro-Wilk statistical test gave a similar result as the Lilliefors test.

**Table 2** Type I Error Rate of Various Statistical Tests for normality test in MINITAB14

| Distributions | Statistical Tests | Sample Size | α | |
|---|---|---|---|---|
| | | | .05 | .10 |
| $N(100, 10^2)$ | AD | 10 | .040 | .102 |
| | | 20 | .036*C | .092 |
| | | 30 | .052 | .116 |
| | | 50 | .038*C | .096 |
| | | 100 | .046 | .094 |
| | RJ | 10 | .044 | .096 |
| | | 20 | .042 | .104 |
| | | 30 | .054 | .118 |
| | | 50 | .038*C | .090 |
| | | 100 | .046 | .088 |

**Table 2** Type I Error Rate of Various Statistical Tests for normality test in MINITAB14 (cont')

| Distributions | Statistical Tests | Sample Size | α | |
|---|---|---|---|---|
| | | | .05 | .10 |
| $N(100, 50^2)$ | K-S | 10 | .050 | .118 |
| | | 20 | .034*C | .100 |
| | | 30 | .034*C | .100 |
| | | 50 | .040 | .112 |
| | | 100 | .046 | .114 |
| $N(100, 100^2)$ | AD | 10 | .048 | .120 |
| | | 20 | .034*C | .114 |
| | | 30 | .028*C | .112 |
| | | 50 | .026*C | .114 |
| | | 100 | .030*C | .110 |
| | RJ | 10 | .044 | .126*C |
| | | 20 | .028*C | .120 |
| | | 30 | .024*C*B | .108 |
| | | 50 | .024*C*B | .116 |
| | | 100 | .026*C | .114 |
| | K-S | 10 | .048 | .128*C |
| | | 20 | .040 | .126*C |
| | | 30 | .034*C | .120 |
| | | 50 | .036*C | .118 |
| | | 100 | .032*C | .122*C |

* C has a value outside the range specified by the Cochran criteria; *B has a value outside the range specified by the Bradley criteria.

From Table 2, the majority of Type I error from AD Statistical test (73 percent) was within the controllable range according to the Cochran criteria. All the values outside the range were lower than the lower bound of Cochran i.e. lower than 0.04 and occurred only when $\alpha = .05$.

RJ statistical test gave similar results to AD, both when $\alpha = .05$ and .10; some values were outside the range specified by the Cochran criteria (six percent) and lower than the lower bound of Bradley, i.e. lower than 0.025.

It was found that the K-S statistical test gave a similar result to Lilliefors in SPSS.. It can also control the probability of Type I error according to the Cochran criteria.

**Table 3** Empirical Power of Various Statistical Tests for Normality Test in SPSS14

| Distributions | Sample Size | Statistical Tests | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1-Sample K-S | | Lilliefors | | S-W | |
| | | α | | α | | α | |
| | | .05 | .10 | .05 | .10 | .05 | .10 |
| Short Tailed Distribution | | | | | | | |
| | 10 | .000 | .000 | .054 | .104 | .066 | .152 |
| | 20 | .000 | .000 | .084 | .170 | .214 | .362 |
| Uniform (0-1) | 30 | .000 | .000 | .124 | .246 | .376 | .596 |
| | 50 | .000 | .006 | .288 | .456 | .778 | .912 |
| | 100 | .016 | .052 | .610 | .750 | .992 | .998 |
| | 10 | .000 | .000 | .064 | .140 | .076 | .182 |
| | 20 | .000 | .000 | .120 | .184 | .210 | .358 |
| Uniform (10-50) | 30 | .000 | .002 | .166 | .276 | .396 | .586 |
| | 50 | .000 | .004 | .248 | .404 | .742 | .880 |
| | 100 | .008 | .044 | .586 | .770 | .998 | 1.000 |
| Long Tailed Distribution | | | | | | | |
| | 10 | .126 | .190 | .546 | .606 | .572 | .620 |
| | 20 | .482 | .500 | .836 | .880 | .856 | .892 |
| t(1) | 30 | .706 | .794 | .934 | .960 | .962 | .966 |
| | 50 | .924 | .952 | .996 | .996 | .996 | .996 |
| | 100 | .998 | .998 | 1.000 | 1.000 | 1.000 | 1.000 |
| | 10 | .000 | .000 | .114 | .164 | .116 | .164 |
| | 20 | .002 | .010 | .148 | .220 | .208 | .258 |
| t(5) | 30 | .004 | .012 | .150 | .228 | .250 | .326 |
| | 50 | .006 | .016 | .200 | .286 | .352 | .446 |
| | 100 | .030 | .054 | .342 | .468 | .576 | .632 |
| With Kurtosis Slightly Higher Than the Normal | | | | | | | |
| | 10 | .000 | .000 | .070 | .134 | .082 | .134 |
| | 20 | .000 | .004 | .096 | .174 | .132 | .200 |
| Logistic (0,0.5) | 30 | .000 | .010 | .108 | .196 | .166 | .234 |
| | 50 | .000 | .008 | .118 | .190 | .214 | .288 |
| | 100 | .006 | .010 | .168 | .252 | .318 | .392 |
| | 10 | .000 | .000 | .080 | .144 | .066 | .138 |
| | 20 | .002 | .006 | .106 | .152 | .122 | .214 |
| Logistic (0,1) | 30 | .000 | .000 | .078 | .148 | .128 | .198 |
| | 50 | .000 | .004 | .092 | .160 | .178 | .266 |
| | 100 | .002 | .008 | .158 | .262 | .284 | .408 |
| With High Kurtosis | | | | | | | |
| | 10 | .000 | .000 | .122 | .218 | .132 | .204 |
| | 20 | .000 | .000 | .210 | .318 | .280 | .370 |
| Laplace (0,0.5) | 30 | .006 | .016 | .308 | .418 | .376 | .464 |
| | 50 | .018 | .054 | .424 | .576 | .520 | .620 |
| | 100 | .082 | .192 | .704 | .796 | .802 | .864 |

**Table 3** Empirical Power of Various Statistical Tests for Normality Test in SPSS14 (cont')

| Distributions | Sample Size | Statistical Tests | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1-Sample K-S | | Lilliefors | | S-W | |
| | | α | | α | | α | |
| | | .05 | .10 | .05 | .10 | .05 | .10 |
| Cauchy (0,0.5) | 10 | .124 | .204 | .568 | .646 | .588 | .660 |
| | 20 | .500 | .606 | .858 | .898 | .888 | .910 |
| | 30 | .682 | .786 | .950 | .964 | .972 | .982 |
| | 50 | .942 | .956 | .998 | 1.000 | 1.000 | 1.000 |
| | 100 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| Cauchy (0,2) | 10 | .104 | .174 | .554 | .614 | .564 | .618 |
| | 20 | .468 | .572 | .814 | .868 | .836 | .872 |
| | 30 | .728 | .784 | .948 | .962 | .970 | .978 |
| | 50 | .928 | .964 | .996 | 1.000 | .998 | 1.000 |
| | 100 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

From Table 3, the 1-Sample K-S statistical test had low empirical power, in fact the lowest in relation to the other two tests. Most values were close to zero (0) even though sample size was as large as 100. There was a slight difference in value when α = .05 and .10, except the Cauchy distribution wherein the empirical power was highest and equaled the Lilliefors and S-W tests.

The Lilliefors statistical test had higher empirical power than the 1-Sample K-S test. The empirical power was also higher with larger sample size. The empirical power was high when α= .10 and higher when α = .05. The empirical power was highest i.e. at 100 percent when the population came from a t-distribution (1), Cauchy (0, .0.5), and Cauchy (0, 2) from a sample of 100.

The S-W statistical test had the highest empirical power. The empirical power was even higher when the sample size was larger. The empirical power at α = 0.10 was higher than at .05 for almost all distributions. The empirical power was also closer to 100 percent when the sample size was 100.

**Table 4** Empirical Power of the test of Various Statistical Tests for Normality Test in MINITAB14

| Distributions | Sample Size | Statistical Tests | | | | | |
|---|---|---|---|---|---|---|---|
| | | AD | | RJ | | K-S | |
| | | α | | α | | α | |
| | | .05 | .10 | .05 | .10 | .05 | .10 |
| Short Tailed Distribution | | | | | | | |
| | 10 | .068 | .126 | .042 | .096 | .058 | .110 |
| | 20 | .184 | .287 | .072 | .202 | .082 | .174 |
| Uniform (0-1) | 30 | .306 | .450 | .160 | .330 | .122 | .258 |
| | 50 | .582 | .757 | .463 | .690 | .286 | .457 |
| | 100 | .932 | .972 | .962 | .986 | .620 | .752 |

**Table 4** Empirical Power of the test of Various Statistical Tests for Normality Test in MINITAB14 (cont')

| Distributions | Sample Size | Statistical Tests | | | | | |
|---|---|---|---|---|---|---|---|
| | | AD | | RJ | | K-S | |
| | | α | | α | | α | |
| | | .05 | .10 | .05 | .10 | .05 | .10 |
| | 10 | .076 | .184 | .046 | .118 | .062 | .150 |
| | 20 | .182 | .294 | .090 | .192 | .120 | .188 |
| Uniform (10-50) | 30 | .344 | .464 | .188 | .358 | .170 | .288 |
| | 50 | .560 | .726 | .432 | .648 | .254 | .412 |
| | 100 | .944 | .982 | .962 | .984 | .584 | .766 |
| Long Tailed Distribution | | | | | | | |
| | 10 | .584 | .646 | .586 | .668 | .538 | .610 |
| | 20 | .870 | .898 | .800 | .914 | .836 | .880 |
| t(1) | 30 | .964 | .972 | .960 | .974 | .932 | .960 |
| | 50 | .996 | 1.000 | .998 | 1.000 | .996 | .998 |
| | 100 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | 10 | .122 | .168 | .138 | .178 | .112 | .174 |
| | 20 | .186 | .272 | .240 | .312 | .148 | .224 |
| t(5) | 30 | .224 | .312 | .298 | .376 | .148 | .228 |
| | 50 | .276 | .375 | .409 | .483 | .200 | .288 |
| | 100 | .443 | .586 | .618 | .715 | .350 | .471 |
| With Kurtosis Slightly Higher Than the Normal | | | | | | | |
| | 10 | .077 | .148 | .081 | .152 | .071 | .140 |
| | 20 | .113 | .189 | .163 | .225 | .101 | .179 |
| Logistic (0,0.5) | 30 | .138 | .214 | .192 | .262 | .110 | .200 |
| | 50 | .178 | .250 | .254 | .336 | .118 | .190 |
| | 100 | .242 | .353 | .373 | .464 | .167 | .250 |
| | 10 | .080 | .154 | .096 | .166 | .080 | .162 |
| | 20 | .120 | .187 | .163 | .245 | .108 | .157 |
| Logistic (0,1) | 30 | .106 | .176 | .160 | .234 | .082 | .148 |
| | 50 | .142 | .218 | .240 | .322 | .094 | .166 |
| | 100 | .264 | .351 | .371 | .485 | .158 | .264 |
| With High Kurtosis | | | | | | | |
| | 10 | .146 | .230 | .180 | .244 | .128 | .228 |
| | 20 | .251 | .345 | .295 | .397 | .191 | .289 |
| Laplace (0,0.5) | 30 | .394 | .520 | .420 | .576 | .309 | .464 |
| | 50 | .550 | .615 | .583 | .735 | .427 | .574 |
| | 100 | .826 | .898 | .842 | .922 | .698 | .820 |

**Table 4** Empirical Power of the test of Various Statistical Tests for Normality Test in MINITAB14 (cont')

| Distributions | Sample Size | Statistical Tests | | | | | |
|---|---|---|---|---|---|---|---|
| | | AD | | RJ | | K-S | |
| | | α | | α | | α | |
| | | .05 | .10 | .05 | .10 | .05 | .10 |
| Cauchy (0,0.5) | 10 | .589 | .678 | .606 | .676 | .575 | .666 |
| | 20 | .896 | .914 | .902 | .918 | .858 | .892 |
| | 30 | .992 | .982 | .996 | .984 | .969 | .962 |
| | 50 | 1.000 | 1.000 | 1.000 | 1.000 | .998 | 1.000 |
| | 100 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| Cauchy (0,2) | 10 | .590 | .682 | .602 | .692 | .560 | .656 |
| | 20 | .862 | .892 | .862 | .892 | .820 | .892 |
| | 30 | .972 | .972 | .978 | .986 | .946 | .976 |
| | 50 | .994 | 1.000 | .996 | 1.000 | .990 | 1.000 |
| | 100 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

From Table 4, the RJ statistical test had the highest empirical power when compared with the other two (RJ had lower empirical power to AD in Uniform distribution only, yet when the sample size was 100, the empirical power was higher) in all sample sizes and significance levels. The empirical power was also higher when the sample size was larger, both when α = .05 and .10. In some distributions, when the sample size was 100, the empirical power would reach 100 percent.

The AD statistical test had empirical power second to the RJ test. The trend of empirical power was also similar to RJ in that it depended on sample size and significance level.

The K-S statistical test had the lowest empirical power while its trend was similar to the other two in that the empirical power was higher with larger sample size and α = .10 had higher empirical power than α = .05. It was also found that the empirical power derived by the K-S statistical test was similar to those by Lilliefors in SPSS.

Some graphs of empirical power are shown in Figures 2-5.

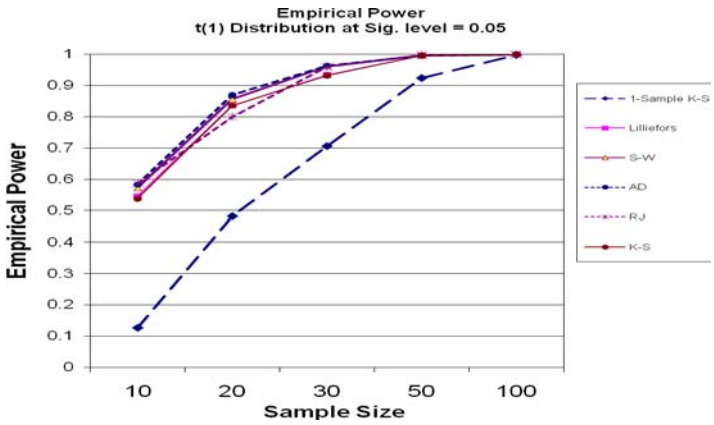


**Figure 2** Empirical Power t (1) Distribution at sig. level 0.05

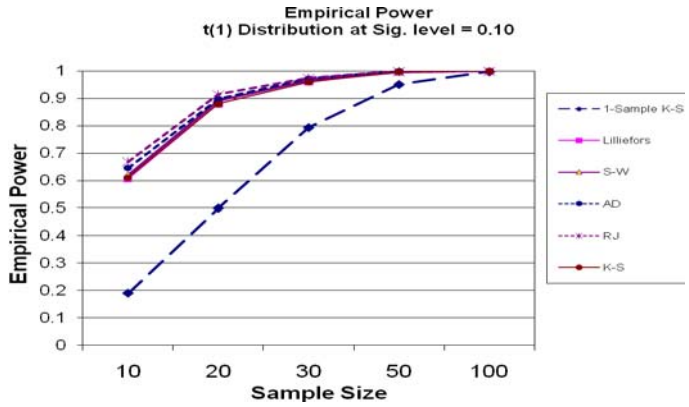


**Figure 3** Empirical Power t (1) Distribution at sig. level 0.10

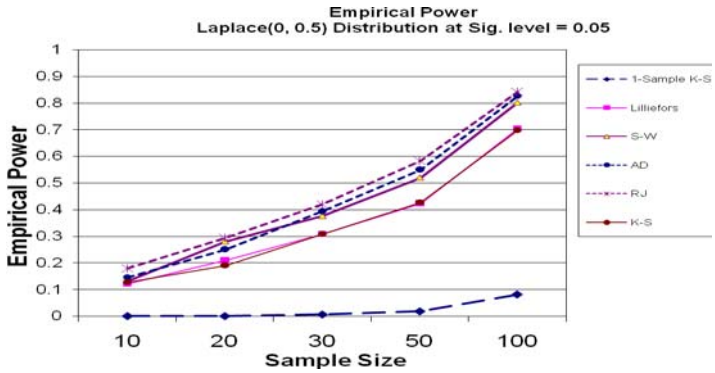


**Figure 4** Empirical Power Laplace (0, 0.5) Distribution at sig. level 0.05

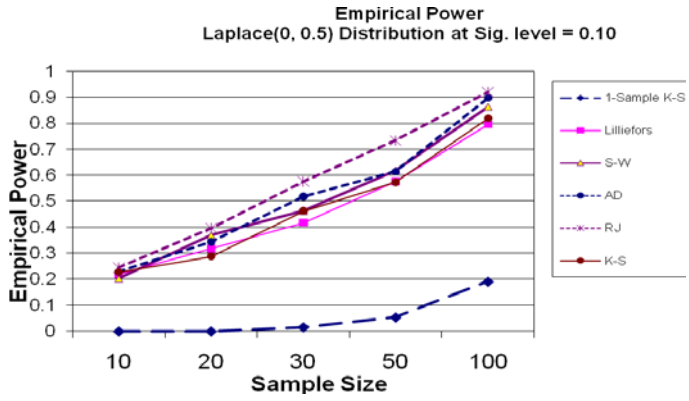


**Figure 5** Empirical Power Laplace (0, 0.5) Distribution at sig. level 0.10

## 5. Research Conclusions

Significant conclusions of this research are as follows:

**Conclusion from SPSS14 Programme**

It was deducible that the below three (3) statistical tests in SPSS were able to control the probability of Type I error.

It was found that the S-W test from SPSS was the most efficient in all distributions. The test by Lilliefors provided the second highest empirical power. Meanwhile, the 1-Sample K-S test provided the least empirical power.

**Conclusion from MINITAB14**

It can be concluded that all three (3) statistical tests in MINITAB were able to control the probability of Type I error.

It was found that the RJ statistical test had the highest empirical power when compared with the other two tests. The AD statistical test had empirical power second to the RJ test. The K-S statistical test had the lowest empirical power.

**Conclusion of Comparison between SPSS and MINITAB**

Comparison of empirical power for similar statistical tests in SPSS and MINITAB.

There were two pairs of similar statistical tests in SPSS and MINITAB:

1. S-W (in SPSS) and RJ (in MINITAB) – it was found that majority of empirical power from RJ was higher than those from S-W;
2. K-S (in MINITAB) and 1-sample K-S or Lilliefors (in SPSS) – the values shown in Tables 1–4 can confirm that the K-S statistical test in MINITAB was in fact Lilliefors. Both tests gave similar empirical power values in all situations of this research.

## 6. Findings Discussion

The result of this study was in accordance with the findings of Gan and Kochler (1990), Edith Seier (2004) and Ketchan (1991) which concluded that S-W had high power. Yet such findings were in contradiction with the SPSS programme manual which suggested the use of S-W when sample size $\leq 50$. This study found that S-W and RJ tests had empirical power closer to 1 when the sample size was closer to 100 in almost all distributions and when $\alpha = .05$ or .10. This meant that if a researcher would like to be highly ensured of the result of a Normal distribution by S-W or RJ, the sample size should be as close to 100 as possible.

Lilliefors (or K-S in MINITAB), on the other hand, gave similar results to S-W, yet most empirical power values were lower. Values of both parameters (i.e. $\mu$ and $\sigma^2$) of the Normal distribution may be unknown or when only one parameter value was known. Yet in SPSS and MINITAB, only the case of unknown $\mu$ and $\sigma^2$ was chosen. So it would be worth considering whether or not the algorithm of SPSS and MINITAB chooses the right distribution, i.e. generating an accurate p-value. Had SPSS and MINITAB included menus allowing users to choose various case situations, the distribution would have been more accurate.

The 1-Sample K-S test had the lowest efficiency. Such findings were in accordance with the conclusion from the study of D'Agostino and Stephens (1986) which suggested that this test should not be used to test a Normal distribution as it has relatively low power when compared to others. Likewise, the study by Steinskog et al. also concluded that this test would normally accept that a random variable had a normal distribution (it is hard to say that it does not have a normal distribution). It also warned that user should be aware that this test featured in various statistical packages,

under this appellation, could possibly be Lilliefors.

The AD statistical test (in MINITAB) was set as a default test and had high efficiency second to RJ (in MINITAB). However, its efficiency was similar to that of S-W (in SPSS) (i.e. approximately 50 percent of 90 tested powers were lower than those of S-W while the rest were higher). The findings were in contradiction with the MINITAB manual: Home > Support > Answer ID: 1167 (22) which gave an answer to "Which test should be chosen to test a Normal distribution?" Practically, a statistician would firstly choose the AD statistical test if importance was placed on deviation at the tail of a distribution. The manual also said that all three (3) tests had low efficiency in identifying a t-distribution or distinguishing a distribution with non-normal kurtosis from normal. Such a suggestion was in contradiction with this study which tested a t-distribution at t(1) and t(5). It was found that at t(1), all three (3) tests had high empirical power i.e. more than 90 percent when the sample size was only 30 and 100 percent when the sample size was 50. At t(5), AD and K-S had relatively low empirical power (lower than 35 – 58 percent) even when a sample size of 100 was used. Yet RJ gave a relatively high value when the sample size was 100 i.e. at 72 percent ($\alpha$ = .10) and 62 percent ($\alpha$ = .05). As for a distribution with high kurtosis i.e. L(0,0.5), C(0,0.5) and C(0,2), it was found that all three statistical tests had a very high efficiency i.e. 100 percent when the sample size was 100, in almost all cases (lowest value was only 70 percent from K-S at $\alpha$ = .05 when n = 100) and RJ was found to have the highest empirical power. The conclusion of this study supported the answer from the MINITAB manual for a Logistic distribution in which the power was lower than 50 percent even when the sample size was 100, while RJ had the highest value (48.5 percent when n = 100 and $\alpha$ = .10).

## 7. Recommendations

To enable a wider range of findings, there should be further studies on data with huge deviations from a Normal distribution such as distributions with slight skewness, high skewness, bimodal or scale contaminated, and a mixture of Normal distributions. The study should also be made to cover the issue on degree of accuracy from additional distributions – Exponential, Uniform, and Poisson – other than Normal provided in 1-Sample K-S in SPSS. The last issue worth considering would be on the parameters on which a p-value was based in the Lilliefors test to assure a user that the result was based on an accurate distribution

## 8. References

[1] James J. Higgins, Introduction to Modern Nonparametric Statistics, Thomson Brook/Cole, 2004.

[2] P. Sprent, Applied Nonparametric Statistical Methods, 2nd edition, Chapman & Hall, London, 1993.

[3] SPSS-X User's Guide, SPSS Inc., McGraw-Hill, New York, 1983.

[4] Anderson, T. W. and Darling, D. A., A Test of Goodness-of-Fit, Journal of the American Statistical Association, Vol.49, pp.765-769, 1952.

[5] Kolmogorov, A.N., Sulla Determinazione Empirica di una Legge di Distribuzione, Giorna Ist. Attuari., Vol.4, pp.83-91, 1933.

[6] Shapiro, S. and Wilk, M.B., An Analysis of Variance Test for Normality, Biometrika, Vol.52, pp.591-611, 1965.

[7] Zhang, P. ,Omnibus Test of Normality Using the Q Statistic, Journal of Applied Statistics , Vol.26, pp.519-528, 1999.

[8] W.J. Conover, Practical Nonparametric Statistics, John Wiley & Sons Inc., 1971.

[9] Jean Dickinson Gibbons, Nonparametric Methods for Quantitative Analysis, Holt, Inehart and Winston, 1976.

[10] Steinskog, Tjostheim and Kvamsto ,A Cautionary Note on the Use of Kolmogorov-Smirnov Test for Normality, Journal of American Meteorological Society, March, pp.1151-1156, 2007.

[11] Paul H. Kvam and Brani Vidakovic, Nonparametric Statistics with Applications to Science and Engineering, John Wiley & Sons, Inc., 2007.

[12] Umaporn Chantasorn and Manus Paitooncharoenlap, Comparison Results of Normality Test by Various Test Statistics from Statistical Package SPSS, Journal of Science Ladkrabang, Vol.17, No.2, pp. 11-24, 2008.

[13] Gan, F.F. and Kochler, K.J., Goodness of Fit Test Based on P-P Probability Plots, Technometrics ,Vol.32, pp.289-303, 1990.

[14] Ketchan Bhajarinsak, Comparisons of Nonparametric Statistical Tests for Normality Test, Thesis of Statistics Department, Graduate School, Chulalongkorn University, Thailand, 1991.

[15] Edith Seier, Comparison of Test for Univariate Normality, Department of Mathematics, East Tennessee State University, Johnson City, TN 37614, 2004.

[16] Kamon Budsaba, Wrong Standard Error Calculation from SPSS Program for Factorial Experiments: Mixed Model,Thai Science and Technology Journal, Vol.12 , No.1, pp. 77-82**,** 2004**.**

[17] Kamon Budsaba, SPSS Misleading p-value Report for Cross-tabulation Data Analysis with Chi-Square, Thai Science and Technology Journal, Vol.11, No.2, pp.83-86, 2003.

[18] Reinhard Bergmann, John Ludbrook, and Will P.J.M. Spooren , Different Outcomes of the Wilcoxon-Mann-Whitney Test from Different Statistics Packages, Journal of The American Statistician, Vol. 54, No.1, 2000.

[19] Leo Knüsel, On the Accuracy of Statistical Distributions in Microsoft Excel 2003, Computational Statistics & Data Analysis, Vol.48, pp.445 – 449, 2005.

[20] B.D. McCullough, Berry Wilson, On the Accuracy of Statistical Procedures in Microsoft Excel 97, Computational Statistics & Data Analysis, Vol.31, pp.27 -37, 1999.

[21] Kusaya Plungpongpun, A Comparison of the Efficiency of Data Generation Adopted by Statistical Computing Package, Annual Proceeding of Symposium on Statistics and Applied Statistics,Thailand, 2007.

[22] http://www.minitab.com/support/answers/ answer.aspx?ID=1167.