How to test normality distribution for a variable: a real example and a simulation study

Alireza Akbarzadeh Baghban[1], Shima Younespour[2], Sara Jambarsang[3], Maryam Yousefi[4], Farid Zayeri[3,*], Farid Azizi Jalilian[5]

[1]Department of Basic Sciences, School of Rehabilitation, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[2]Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran
[3]Proteomics Research Center, Faculty of Paramedical Sciences, Shahid Beheshti University of medical sciences, Tehran, Iran
[4]Skin Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[5]Faculty of Medicine, Ilam University of medical Scinces, Ilam, Iran

*Corresponding Author: email address: fzayeri@yahoo.com (F. Zayeri)

ABSTRACT

Many commonly used statistical methods require that the population distribution be nearly normal. Unfortunately, in some papers the one-sample Kolmogorov-Smirnov test has been used for testing normality while the assumptions of applying this test are not satisfied. To conduct this test, it is assumed that the population distribution is fully specified. In practical situation where the mean and SD of population distribution is not specified in advance, one can use a modification of the K-S test for checking the normality assumption which is called, Lilliefors test. In this paper, we explain the method of computing this test with some common statistical softwares such as SPSS, S-PLUS, R and StatXact and utilize a dermatology dataset from Skin Research Center of Shohada-e-Tajrish hospital to illustrate how the use of the one-sample K-S (with the mean and SD estimated from the sample) instead of its modification can be misleading in practice. We also use Monte Carlo simulation to compare the approximate power of the one-sample K-S test (with the estimated population mean and SD) with Lilliefors test in some common specified continuous distributions. The result indicates that one should not use the one-sample K-S test for assessing the normality assumption in practical situation.

Keywords: one-sample Kolmogorov-Smirnov test; Lilliefors test; Monte Carlo simulation; testing normality assumption

INTRODUCTION

Many statistical procedures are based on the assumption that the population is approximately normally distributed [1].When this assumption is violated, inference may not be reliable or valid [2]. If a normal distribution is tentatively assumed to be a plausible model, the investigator must still check this assumption once the sample data are obtained [3]. There are two methods of checking the normality assumption, Graphical and numerical methods, which are either descriptive or theory-driven. Graphical methods are used to visualize the distributions of random variables and compare the distribution to a theoretical one using plots. Numerical methods present descriptive statistics or conduct statistical tests of normality. The descriptive methods are based on the empirical data, whereas the theory-driven methods consider both empirical and theoretical distributions. Although graphical methods are based on subjective visual examination of the data, they are helpful in detecting serious departures from normality and are easy to interpret. Numerical methods provide objective ways of assessing normality [2].

A Stem-and-leaf plot, box plot, dot plot and histogram are descriptive graphical methods, while The Q-Q and P-P plots are theory driven ones. Skewness and kurtosis are descriptive numerical methods, whereas the Shapiro-Wilk, Shapiro- Francia, Kolmogorov- Smirnov (Lilliefors test), Anderson-Darling, Cramer-von Mises, Jarque-Bera, Skewness- Kurtosis tests are some of the theory-driven numerical methods that provide a diagnostic check for possible departure from a normal distribution [2].

Unfortunately, in some papers the one-sample Kolmogorov-Smirnov test (K-S test) has been

used for testing normality while the assumptions of applying this test are not satisfied. To conduct this test, it is assumed that the population distribution is fully specified (i.e. it assumes that you know the mean and Standard deviation (SD) of the overall population perhaps from prior work). When analyzing data, you rarely know the overall population mean and SD. You only know the estimated mean and SD from your sample. In addition, this test tends to be more sensitive near the center of the distribution than at the tails and it appears to waste information by using only the largest discrepancy between cumulative distribution of the sample and a cumulative normal distribution [4].

This test (K-S test) is used to decide if a sample comes from a population with a completely specified continuous distribution [4,5]. The null hypothesis of this test is that the data follow a specified distribution and an alternative hypothesis tells that the data do not follow it. The test statistic is based on the maximum distance between the empirical distribution function (EDF) of the sample and the cumulative distribution function (CDF) of the reference distribution [4]. In the special case of testing for normality, the EDF is compared with the CDF of the normal distribution and the normality hypothesis is rejected if the test statistic exceeds the critical value obtained from tables may be found in conover (1999) or in many of general statistical tables [4,5]. When the mean and SD of population distribution are unknown and are estimated from the sample (practical situation), the power of the test to detect departures from the normal distribution may be seriously reduced. So for this situation, a modification of the Kolomogorov-Smirnov test, Lilliefors test, is used [6].

The null hypothesis for Lilliefors test is that the data is normally distributed with unknown mean and standard deviation and the alternative hypothesis tells that the data is not normally distributed. The critical region of the K-S test is no longer valid when mean and SD of the population is estimated from sample [6].It is suggested that the probability of a type *I* error will be smaller than as given by tables of the K-S statistic [7,8].

Lilliefors used the Monte-Carlo method to compute an approximation of the sampling distribution of the test statistic. For this procedure, a large number of samples are selected from a normal population and the values of the test statistics are calculated for each of these samples. An approximation of the sampling distribution of the test statistic under the normality assumption is obtained by the empirical distribution of the values of the test statistics [4].

In order to conduct the Lilliefors test of normality, first, one can estimate the population mean and variance from the sample data. Then the maximum discrepancy between the EDF and the CDF of the normal distribution can be found with the estimated mean and estimated variance, this will be the test statistic. Finally, finding out whether the test statistic is large enough to be statistically significant is of interest, this is where this test becomes more complicated than the K-S test [4]. Since the hypothesized CDF has been moved closer to the data by estimation based on those data, the maximum discrepancy has been made smaller than it would have been if the null hypothesis had singled out just one normal distribution. Thus the ”null distribution” of the test statistic, i.e. its probability distribution assuming the null hypothesis is true, is stochastically smaller than the KolmogorovSmirnov distribution. This is the Lilliefors distribution [9].

In this paper, we point out the way of checking the normality assumption by the lilliefors test in most widely used statistical software packages such as SPSS, S-PLUS, R and StatXact. We use a dermatology dataset to illustrate how the use of the one-sample K-S test instead of Lilliefors test can be misleading. Also through Monte Carlo simulation, we can find out which tests are more powerful. So a brief Monte Carlo investigation is made to compare the approximate power of the one-sample K-S test (with the estimated population mean and SD) with Lilliefors test in some common specified distributions.

MATERIAL AND METHODS

Applied example

A dermatology data gathered by Shohada-e-Tajrish Skin Research Center is applied to illustrate how wrong is the use of the K-S test (with the estimated population mean and SD) instead of its modification. This study was

performed to compare the serum Antioxidant levels in 30 patients with pemphigus vulgaris, an auto-immune blistering disorder, and 30 healthy individuals referred to two major Hospitals of Shahid Beheshti University of Medical Sciences named Shohada-e-Tajrish and Loghman-e-Hakim Hospitals.

Simulation

For Monte Carlo simulation, ten thousand samples of sizes 30 and 50 are drawn from each of several distributions. These distributions are of different shapes where some look like the normal distribution while others are substantially different. A Uniform (0,1) distribution; a Lognormal (1,1.3) distribution; a Logistic (0,1) distribution; a Normal (0,1) distribution; two t distributions and two chi-square distributions with degrees of freedom 3 and 30 are included in this power investigation and Results for type-one error (Alpha)=0.05 is reported.

The computation of the approximate power is done as follows. A random sample of size n is generated from a given non-normal distribution and it has seen how many times the null hypothesis of normality has been rejected. For applied example and simulation study the SPSS 16.0.0 and R 2.10.1 were used, respectively.

Statistical softwares

The way of computing the Lilliefors test in SPSS, S-PLUS, R and StatXact softwares is as follows:

SPSS: Analyze →Descriptive Statistics → Explore →Plots →normality plots with tests (For testing against a normal distribution with estimated parameters)

S-PLUS: Statistics → Compare Samples → One Sample Kolmogorov-Smirnov GOF (if the mean and SD of the population are not specified by the user)

R software: lillie.test (), one should install the package nortest in advance.

StatXact: Statistics → One Sample Goodness of Fit → Lilliefors

It is worth mentioning that StatXact 8 is professional software (now with 140 exact tests and procedures) for conducting many of nonparametric tests.

RESULTS

We assess the normality assumption of the two variables, Direct Bilirubin and Serum Selenium, in the case group by using the one-sample Kolmogorov-Smirnov (with estimated population mean and variance) and Lilliefors tests. The results are shown in table 1. As it can be seen in table 1, for both variables the one-sample Kolmogorov-Smirnov test dose not rejects the normality assumption while the Lilliefors test dose. Also graphical methods mentioned before, show the non-normality of the two variables.

Table 1: P-values obtained by one-sample K-S (with estimated population mean and Variance) and Lilliefors tests for checking normality of the two variables

variables	K-S test	Lillifors
Direct Bilirubin	0.155	0.002
Serum Selenium	0.336	0.024

Table 2: Probability of rejecting hypothesis of normality using K-S (with estimated population mean and Variance) and lilliefors test when sample sizes are 30 and 50.the numbers are result of Monte Carlo calculations with 10000 samples for each distribution.

	n=30		n=50	
distribution	**K-S test**	**Lilliefors**	**K-S test**	**Lilliefors**
Normal(0,1)	0.0001	0.0505	0.0003	0.0508
Lognormal(1,1.3)	0.7019	0.9848	0.9727	0.9998
Logistic(0,1)	0.0012	0.0894	0.0014	0.1100
t(3)	0.0538	0.3382	0.1085	0.4905
T(30)	0.0003	0.0594	0.0002	0.0554
$\chi^2(3)$	0.0441	0.5749	0.1482	0.8245
$\chi^2(30)$	0.0011	0.1026	0.0012	0.1333
Uniform(0,1)	0.0004	0.1500	0.0007	0.2515

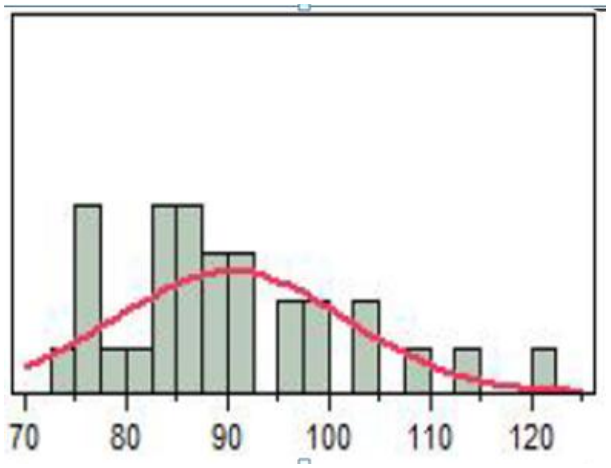

Figure 1: Serum selenium histogram chart Shapiro-Wilk: p-value=0.0494

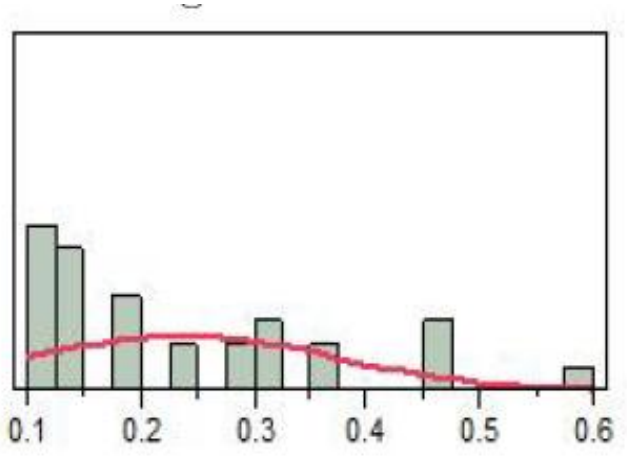

Figure 2: Direct bilirubin histogram Shapiro-Wilk: p-value=0.0003

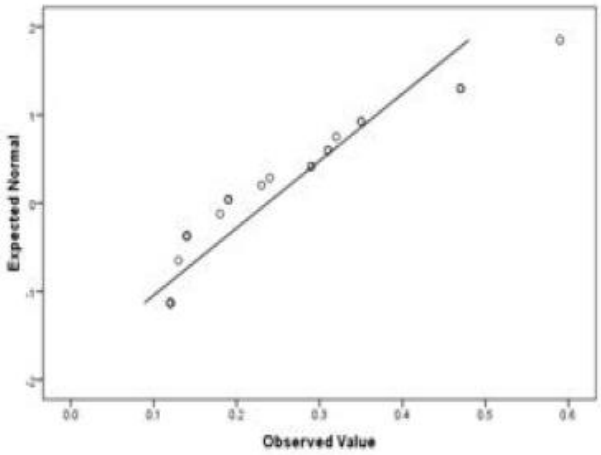

Figure 3: Normal quantile plot of Direct bilirubin

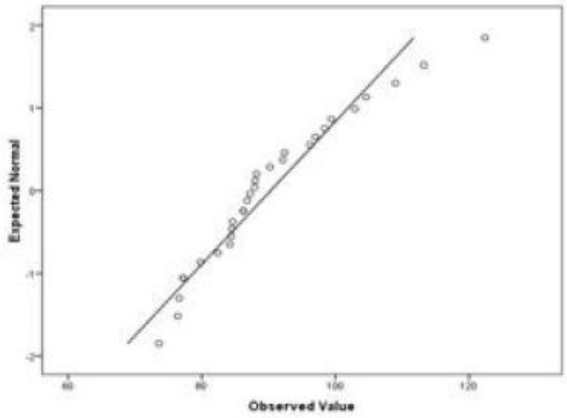

Figure 4: Normal quantile plot of serum selenium

The p-value obtained by the Shapiro-wilk checking normality assumption of Serum Seleniume and Direct Bilirubin are 0.0494 and 0.0003, respectively. as we can see, these values are in the direction of Lilliefors test results. Also,Q-Q plots and histograms of the two variables suggest the non normality of data (figure 1-4). The results of the simulation study are shown in table 2. From this table we can see, for two tests the power was quite large for lognormal distribution. The power of Lilliefors test is better than Kolmogorov-Smirnov test for all distributions that assumed. For the chi-square distribution with 3 degree of freedom, the Kolmogorov-Smirnov test has a much lower power than the Lilliefors test. When sample size of simulation was 30, it was 0.04 and 0.57 for K-S and Lilliefors test, respectively and those were 0.15 and 0.82 when sample size was 50.

DISCUSSION

In this paper, different distributions were used to compare the powers of Lilliefors and onesample K-S test (with the estimated population mean and SD). For all the distributions and two sample sizes mentioned in the table 2, the power of the Lilliefors test is consistently better than the one-sample K-S test.

For these two tests, detecting non-normality is difficult when the observed distribution looks like to normal distribution and this difficulty increases with larger degrees of freedom. We can see from table 2 that with increasing the degrees of freedom of the t and chi-square distributions, the

power of two tests decreases. In theory, with increasing degrees of freedom, the t-distribution behaves like the normal distribution. So the K-S test has much lower power than Lilliefors test when the distribution gets closer to normal.

Although the Lilliefors test is more powerful than one-sample K-S (with the estimated population mean and SD) but, there are more powerful tests for checking the normality assumption such as Shapiro-Wilk test and Anderson-Darling test[10].

No matter which normality test is used, it may fail to detect the actual non-normality of the population distribution if the sample size is small and with large sample sizes, a small deviation from normality will lead to rejection of the normality hypothesis. As a guideline, for sample sizes smaller than 30, one can always assume non-normality of the distribution. For large samples (n>100) If formal test is not significant, one can accept normality otherwise double-check the assumption using graphical methods. For moderate sample sizes (30-100), if the test is significant, one can accept non-normality otherwise double-check using graphs [11].

CONCLUSION

In practical situation where the mean and SD of population distribution is not specified in advance, one should use a modification of the K-S test (Lilliefors test) for checking the normality assumption and specially in SPSS package, one should be aware of not using the nonparametric one-sample K-S option.

ACKNOWLEDGEMENTS

The authors wish to thank the chief of the Skin Research Center, Dr. Parviz Toosi, Dr.Hoda Rahimi, mrs Maryam Poorsani and miss Mahshid Namdari for their contribution to this research.

REFERENCES

1.Mcclave J T, Benson P G, Sincich T. Statistics for business and economics. 8th ed. New jersey:prentice hall international; 2001, p: 234.

2.Park, H M. Univariate Analysis and Normality Test Using SAS, Stata, and SPSS.[working paper]. Indiana: University Information Technology Services, Indiana University; 2008. Available at: URL: http://www.indiana.edu/statmath . Accessed september, 1 2012.

3.Johnson R A, Bhattacharyya G. statistics principles and methods. 2nd ed. John Wiley and Sons; 1992.

4.Lilliefors H. On the KolmogorovSmirnov test for normality with mean and variance unknown.JASA. June 1967; 62: 399402.

5.Sprent P, Smeeton N C. Applied Nonparametric Statistical Methods. 4th ed. Florida: Chapman and Hall/CRC; 2001.

6.Abdi H, Molin P. Lilliefors/Van Soest's test of normality. Available at: URL: http://www.utd.edu herve . Accessed september 2, 2012 .

7.Massey F J. The KolmogorovSmirnov test for goodness of fit. ASA 1951; 46: 68-78

8.Birnbaum Z W. Numerical Tabulation of the Distribution of Kolmogorovs Statistic for Finite Sample Size. Journal of the American Statistical Association 1952; 47: 425-41

9.nist/sematech e-handbook of statistical methods. 2009, Available at: URL: http://www.itl.nist.gov/div898/handbook/,Accessed september 2, 2012.

10.Islam, Tanweer ul.International Islamic University, Islamabad, Pakistan. 2008, Available at:URL:http://mpra.ub.uni-muenchen .de/.Accessed september 2, 2012.

11.Chan Y H. Basic statistics for doctors .2003 vol 44(6) 280-85 ;Available at: URL: http://www.sma.org.sg/smj/. Accessed september 2, 2012.

12.Thode H C. Testing for normality. New York: Marcel Dekker;2002.