Is It Normal? A Simulation Study of Properties of Some Normality Tests

Daniel M. Sultana, Charlyn J. Suarez, Bruce E. Trumbo, Eric A. Suess
Department of Statistics
California State University, East Bay

Abstract

Statistical packages can perform several different goodness-of-fit tests of normality. We consider the normality tests of Anderson-Darling, Shapiro-Wilk, Cramér-von Mises, and Kolmogorov-Smirnov. For a given dataset these tests sometimes lead to different conclusions, possibly leaving students and practitioners confused about which test to believe. We use the statistical package R to simulate normal and nonnormal data and to compare behaviors of these four tests. We also consider the effects of using the tests in combination (for example, in terms of maximum and minimum p-values of several tests).

Introduction

Imagine that a colleague asked you if the data set shown in the stem and leaf plot were normal. A qq-plot of the data is in figure 1. The four tests check normality in different ways. Here the tests disagree about whether the data set is normal .

```
20 | 8
22 | 1160222355678
24 | 3469146677
26 | 11259001179
28 | 1357045
30 | 122571
32 |
34 | 1
36 | 8
```

Test	p-value
Anderson-Darling	0.08
Cramér-von Mises	0.19
Lilliefors (Kolmogorov-Smirnov)	0.40
Shapiro-Wilk	0.01

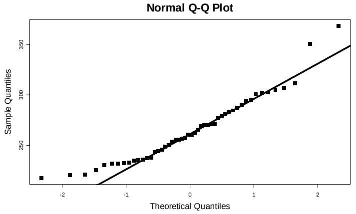

Figure 1: qq-plot of data in stem and leaf plot

Methods

Simulations were done with the computer package R ver 2.3.1. In R, the Shapiro-Wilks test is available in the base package. The other normality tests are in the "nortest" package.

The purpose of the first simulation was to determine if the normality tests tended to identify the same data sets as non-normal. 10,000 normal samples of size 20, 50, and 100 were created and the four normality tests were run against the data. The data sets where normality was rejected were recorded. The lists of rejected data sets from the different normality tests were compared.

The purpose of the second simulation was to compare the power of the four tests. 10,000 samples of size 20, 50, and 100 were drawn from various non-normal distributions. The four normality tests were run at the 5% level and the fraction of data sets where normality was rejected was recorded.

The purpose of the third simulation was to investigate how using the information from all four tests at once influenced the size of the test. 50,000 samples of size 20, 50, and 100 were drawn from the normal distribution. For each sample, the minimum and maximum p-value from the four tests was stored. The fraction of the data sets where the minimum p-value was less than alpha was stored. Similarly, the fraction of the data sets where the maximum p-value was less than alpha was stored.

The purpose of the fourth simulation was to compare the power of the two joint tests. The sizes of the individual tests were set so that the size of the joint test was 5%.

Results

The results from the first simulation are shown in table 1. The normality tests do not always agree about which data sets should be identified as non-normal.

The results from the second simulation are shown in figure 2. The figure shows that the for the joint test to reject at the 5% level, the size of the individual component tests must be adjusted.

The results from the third and fourth simulations are shown in table 2. The Shapiro-Wilks test had more power than the other individual tests in most cases. The joint test formed by taking the minimum p-value from the four tests was more powerful than the joint test formed by taking the maximum p-value.

	Number Rejected out of 10,000 normal samples		
Individual Tests	n=20	n=50	n=100
Cramer-von Mises (CVM)	495	493	563
Anderson-Darling (AD)	489	510	559
Shapiro-Wilk (SW)	489	495	531
Lilliefors (KS)	458	507	534
Combined Test			
AD and CVM	421	424	471
AD and SW	364	320	328
CVM and KS	305	309	354
CVM and SW	302	267	267
AD and KS	269	284	321
SW and KS	203	196	214

Table 1: Shows that the normality tests identify different data sets as non-normal.

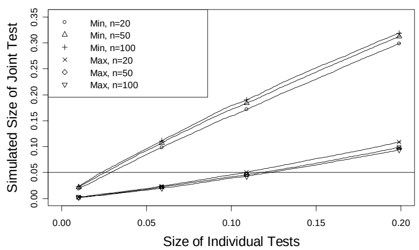

Figure 2 : Two joint tests were studied. In the first joint test, the p-value of the test was set equal to the minimum p-value from the four individual tests. This was equivalent to rejecting normality when at least one test rejected normality. In the second joint test, the p-value of the test was set equal to the maximum p-value from the four individual tests. This was equivalent to rejecting normality when all of the tests rejected normality.

	n=20						n=50						n=100					
	AD	CVM	SW	KS	max	min	AD	CVM	SW	KS	max	min	AD	CVM	SW	KS	max	min
Symmetric alternatives with shorter tails than normal																		
Beta(1/2,1/2)	62	51	73	32	47	61	99	96	100	79	92	100	100	100	100	100	100	100
Uniform	17	14	20	10	16	14	57	44	75	25	42	60	95	85	100	59	100	59
Triangular	4	4	3	4	4	4	5	4	5	4	5	4	8	6	11	5	11	5
Symmetric alternatives with longer tails than normal																		
t(5)	17	16	19	13	16	18	30	27	35	21	26	33	48	44	56	34	43	52
Logistic	10	10	12	8	10	11	15	13	19	11	13	18	24	21	30	16	22	27
Skewed alternatives																		
Weibull(10)	14	13	16	11	16	11	29	26	35	20	27	31	52	47	63	36	48	56
Weibull(3)	5	5	5	5	5	5	6	6	6	5	6	5	7	7	8	7	8	6
Gamma(1,1)	77	73	84	58	84	58	100	99	100	96	99	100	100	100	100	100	100	100
Gamma(10,0.1)	12	11	15	10	15	10	26	23	32	18	24	28	48	42	61	34	45	52

Table 2: Power of 6 normality tests against various alternative distributions. The four individual normality tests are Anderson-Darling (AD) , Shapiro-Wilk (SW) , Cramér-von Mises (CVM), and Lilliefors (KS). The two joint tests are formed by using the minimum and maximum p-value from each of the four tests. All tests were run at the 5% level.

Conclusions

The Shapiro-Wilk test is available in most commercial statistical software packages. Within the scope of our simulations, it should be used in lieu of the other normality tests because of its power.

Casually combining or choosing normality tests results in tests whose size and power differ from the original tests.

References

Ralph B. D'Agostino and Michael A. Stephens. *Goodness-of-Fit Techniques.* Marcel Decker Inc., 1986
Ralph B. D'Agostino, Albert Belanger, Ralph B. D'Agostino, Jr.: `A Suggestion for Using Powerful and Informative Tests of Normality.' *American Statistician,* Vol. 44, No. 4 (Nov., 1990) , pp. 316-321
W. J. Conover. *Practical Nonparametric Statistics.* John Wiley and Sons Inc., 1971.
Stephen W. Looney, Thomas R. Gulledge, Jr.: `Use of the Correlation Coefficient with Normal Probability Plots.' *American Statistician,* Vol. 39, No. 1 (Feb., 1985) , pp. 75-79.
www.sas.com