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

Daniel M. Sultana, Charlyn J. Suarez, Bruce E. Trumbo, Eric A. Suess

April 18, 2006

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.

Specifically, we explore differences among the tests in several ways, focusing on reasons for their disagreement, on their relative power for several kinds of nonnormal data, and effects of using the tests in combination (for example, in terms of maximum and minimum p-values of several tests). Methods and R code are at an appropriate level for classroom use.

1 Introduction

Imagine that a colleague wanted to know if the data in Table 1 were normal.

8.3	8.6	8.8	10.5	10.7	10.8	11.0
11.0	11.1	11.2	11.3	11.4	11.4	11.7
12.0	12.9	12.9	13.3	13.7	13.8	14.0
14.2	14.5	16.0	16.3	17.3	17.5	17.9
18.0	18.0	20.6				

Table 1: R trees data variable girth

Both the Anderson-Darling and the Cramér-von Mises tests seem to indicate not. In contrast, results of both the Shapiro-Wilk and the Lilliefors tests are consistent with normality.

Test	Statistic	p-value
Anderson-Darling	0.7455	0.04668
Cramér-von Mises	0.1283	0.04353
Shapiro-Wilk	0.9412	0.08893
KS/Lilliefors	0.1414	0.1179

Table 2: Is the data sample from a normal distribution?

A qqplot of the data indicates some departure from normality.

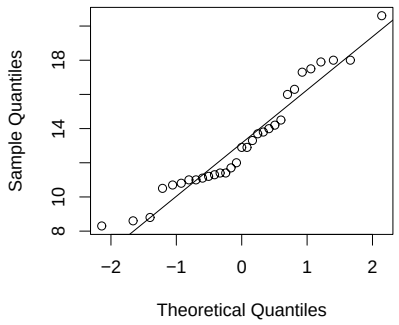

Figure 1: qqplot of trees girth

2 Description of the tests

There are a number of statistical tests available for checking if data are normal. We provide informal de-

scriptions of these tests. Formal definitions are available in [1] [2]. In R, some of the tests are in the supplementary package "'nortest"'.

- The Kolmogorov-Smirnov statistic measures the maximum vertical distance between the empirical cdf and a normal cdf with mean μ and variance σ^2. The KS test uses extreme values of the KS statistic to detect nonnormality. When the mean and variance are estimated from the data, the test changes slightly and is known as Lilliefors test. In R, the commands are ks.test and lillie.test.

- The Cramér-von Mises statistic is an average of the the squared distances between the empirical cdf and the normal cdf. The Cramér-von Mises test uses large values of the statistic to detect non-normality. In R, the command is cvm.test.

- The Anderson-Darling statistic is a weighted average of the squared distances between the empirical cdf and the normal cdf. It gives more weight to the differences between the tails of the edf and normal cdf. The Anderson-Darling test uses large values of the statistic to detect non-normality. In R, the command is ad.test.

- The Shapiro-Wilk statistic is the ratio two estimators of the variance. In R, the command is shapiro.test.

3 When do the tests agree and disagree?

The example in the introduction illustrates that the normality tests can give conflicting results about whether or not a data set is normal. We wanted to determine which tests tend to give similar results and which tend to disagree. 10,000 samples of size 30 were pulled from a Normal(0,1) distribution. Table 3 displays the number of times the various tests rejected normality at the 5% level.

The Cramér-von Mises and the Anderson-Darling tend to agree most of the time on which data sets are non-normal. These tests are based on similar

Test	Count
Anderson-Darling	500
Anderson-Darling and Cramér-von Mises	420
Anderson-Darling and Shapiro-Wilk	344
Anderson-Darling and Lilliefors	262
Cramér-von Mises	489
Cramér-von Mises and Shapiro-Wilk	278
Cramér-von Mises and Lilliefors	290
Shapiro-Wilk	499
Shapiro-Wilk and Lillifors	197
Lilliefors	472

Table 3: Reject normality 10,000 samples of size 30

similar statistics [3]. The Shapiro-Wilk test agrees with the Anderson-Dariling more than with Cramér-von Mises. Lilliefors test agrees with the other tests about half the time.

Since the tests disagree for many of the data sets, it is reasonable to try and identify characteristics of the datasets that are making the different tests reject.

- Skewness: Recall that the $Gamma(n, \frac{1}{n})$ distribution is almost normal $N(1, \frac{1}{n})$ for large n but is skewed right for small n. To determine which test best detects skewed data sets, we pulled samples from $Gamma(n, \frac{1}{n})$ and ran the different normality tests on the samples. The results are plotted in Figure 2. Note that the Shapiro-Wilk test rejects skewed data most often.

- Kurtosis: Recall that the t distribution with df=n is almost normal for large n but has thick tails for small n. To determine which test best detects fat tailed data sets, we pulled samples from $t(df = n)$ and ran the different normality tests on the samples. The Shapiro-Wilk test rejected most often. The Anderson-Darling and Cramér-von Mises were nearly as effective. The Lilliefors tests was least effective.

- Bimodal: Samples were drawn at random from one of two normal distributions $N(0, 1)$ and $N(\mu, 1)$. The resulting samples have a bimodal distribution. The parameter μ was varied from near zero to 10. The Anderson-Darling and

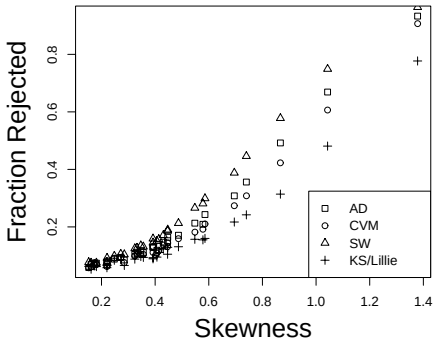

Figure 2: Fraction rejected vs. skewness, gamma distribution

Cramér-von Mises were equally effective at detecting this type of non-normality followed by the Shapiro-Wilks and Lilliefors tests.

References

[1] W. J. Conover. *Practical Nonparametric Statistics*. John Wiley and Sons Inc., 1971.

[2] Ralph B. D'Agostino and Michael A. Stephens. *Goodness-of-Fit Techniques*. Marcel Decker Inc., 1986.

[3] www.sas.com.