# Bootstrap and Other Tests For Goodness of Fit.

Ms. Rumee Rani Savapandit and Dr. Bipin Gogoi

Department of Statistics

Dibrugarh University,Dibrugarh,Assam (INDIA)

**Abstract:** Goodness of fit has a long time been a problem of research .It has received a considerable attention in the statistical literature. Goodness of fit techniques can be described as the method of how well a sample of data agrees with a given distribution as its population. Goodness of fit techniques is based on measuring in some way the conformity of the sample data to the hypothesized distribution or equivalently, its discrepancy from it. The techniques usually give formal statistical tests and the data based measures of conformity or discrepancy are referred to as test statistics. In this paper we have studied the performance of the bootstrap based procedure of EDF based tests for testing the goodness of fit for normality of the distribution using simulation technique. Some results are calculated to know the performance of bootstrap based technique and these are displayed in tables. Discussions and conclusions are made on the basis of results obtained.

**Key words**: Bootstrap procedure, Kolmogorov-Smirnov test, Anderson-Darling test, simulation, power.

**1. Introduction:**

Goodness-of-fit has been occupying an important place in statistical inference since long time. In short it is a technique of examining how well a sample of data agrees with a given distribution as its population. Measures of goodness of fit typically summarize the discrepancy between observed values and the expected values under the model in equation. Such measures can be used in statistical hypothesis testing, i.e. to test for normality of residuals, to test whether two samples are drawn from identical distributions, or whether the outcome frequencies follow a specified distribution. Although it is a cornerstone of modern statistical theory, but still no clear notion of optimality for more complicated situations. An important but difficult problem is evaluating the goodness-of -fit of a model and obtains the P-value. The normal theory of likelihood ratio test statistics whose distribution is approximated by a chi-square test is often used in practice. But it is known that the chi-square distribution is not so accurate for small sample sizes even when the latent factors are normally distributed. Similar results are also reported by Anderson-test too.

Fitting of a probability model to observed data is an important statistical problem from both theory and application point of view. There is a multitude of statistical models and procedures that rely on the validity of a given data hypothesis, being the normality of the data assumption one of the most commonly found in statistical studies. As observed in many models and in research on applied statistics and economics, following the normal distribution assumption blindly may affect the accuracy of inference and estimation procedures. The evaluation of this distributional assumption has been addresses, for example, in Min(2007) where the conditional normality assumption in the sample selection model applied to housing demand is examined, or in Lisenfeld and Jung(2000) where the normality assumption has been addressed in the context of stock market data, a type of data that has been found to be typically heavy –tailed in Gel and Gastwirth(2008). The analysis of the normality hypothesis can also be found in the characterization of error terms in the context of regression analysis models applied to economic time-series Giles(2007), Thadewald and Buning(2007), to probit models Wilde(2008) or to other types of time series Onder and Zaman(2005), Quddus(2008). In medical research the assumption of normality is also very common Schoder, Himmelmann and Wilhelm(2006) and Surucu and Koc(2007), but the suitability of this assumption must also be verified with adequate statistical tests. The definition of adequate normality tests can, therefore , be seen to be of much importance since the acceptance or rejection of the normality assumption of a given data set plays a central role in numerous research fields. As such, the problem of testing normality has gained considerable importance in both theoretical and empirical research and has led to the development of a large number of goodness of fit tests to detect departures from normality. Given the importance of this subject and the widespread development of normality tests over the years, comprehensive descriptions and power comparisons of such tests have also been the focus of attention, thus helping the analyst in the choice of suitable tests for his particular needs.

There have been a quite a few works on goodness-of-fit test based on bootstrap as compared to other test. Blake(2005) discussed the utility of bootstrap method in normally distributed data of violent crime across the states. Matthias von(1997) showed that the parametric bootstrap can be used for analyzing goodness-of-fit, even when the data are very sparse. Alberto and Harry (2008) provided an overview of the new developments in limited information goodness-of-fit assessment of categorical data models. The goodness-of-fit of latent trait models in attitude measurement was discussed by Bartholomew and Tzamourani (1999). There are two versions of the bootstrap, the (naïve) bootstrap and the parametric bootstrap, of which only the parametric bootstrap can be used for goodness-of-fit testing (Bollen & Stine, 1993,Langeheine-1996)

## 2. Goodness of Fit test Based on Empirical Distribution Functions

### 2.1 The Kolmogorov-Smirnov test modified by Lilliefors and Stephens

Kolmogorov and Smirnov ( 1933,1948 ) developed a one sample goodness of fit test based on empirical distribution function(EDF). Kolmogorov-Smirnov(K-S) statistic is popular, although other EDF based statistics such as the Cramer-von-Mises(C-vM) and Anderson –Darling(A-D) statistics have better sensitivity for some data-model differences. However, the goodness of fit probabilities derived from the K-S or other EDF statistics are usually not correct when applied in model fitting situations with estimated parameters.

K-S statistic is no longer distribution –free if some parameters are estimated from the data set under consideration. The K-S probabilities are only valid if the model being tested is derived independently of the data set at hand.

Lilliefors (1967) proposed a modification of Kolmogorov-Smirnov test for normality when the mean and the variance are unknown , and must be estimated from the data. The test statistic K-S is defined as

$$KS = \max_{1 \le i \le n} [\ \Phi(x_i; \bar{x}; s^2) - \frac{(i-1)}{n}; \frac{i}{n} - \Phi(x_i; \bar{x}, s^2)\ ] \qquad (2.1)$$

Where $\Phi(x_i; \bar{x}, s^2)$ is the cumulative distribution function of the normal distribution with parameters estimated from data. The normality hypothesis of the data is then rejected for large values of K-S. Table of percentage points are found in Lilliefors(1967). It can also be obtain give by Stephens(1969). Modification of K-S statistic given by Stephens(1969) from the Lilliefors form is as follows;

$$KS^* = KS(\sqrt{n} - 0.01 + 0.85/\sqrt{n}) \qquad (2.2)$$

Comparing with the upper tail significance points of the distribution on the null hypothesis; may be reject the null hypothesis if value of KS* exceeds the table value at corresponding significance levels. Table of percentage point is available in Stephens(1969).

## 2.2. The Anderson- Darling test

Anderson and Darling(1952,1954 ) introduced a new class of quadratic a test statistics . These are given by

$$Q_n(\psi) = n \int_{-\infty}^{\infty} [F_n(x) - \Phi(x)]^2 \psi(x) dF(x)$$

Where $F_n(x)$ is empirical distribution function(EDF) , $\Phi(x)$ is the cumulative distribution function of the standard normal distribution and $\psi(x)$ is a weight given by

[ $\Phi(x).(1-\Phi(x))]^{-1}$. It can be seen from Anderson-Darling(1954) that AD can be written as

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[In\Phi(z_{(i)}) + In(1-\Phi(z_{(n+1-i)})] \qquad (2.3)$$

Where $z_{(i)} = (x_{(i)} - \bar{x})/s$. In order to increase its power when $\mu$ and $\sigma$ are estimated from the sample, a modification factor has proposed for AD by Stephens(1974) resulting in new statistic AD*:

$$AD^* = AD\left( 1+\frac{0.75}{n} + \frac{2.25}{n^2} \right) \qquad (2.4)$$

The normality hypothesis of the data is then rejected for large values of the test statistic.

Table of percentage points of this statistic is given by D'Agostino(1986).

## 3. Goodness fit test Based on Bootstrap Resampling:

Fortunately, there is an alternative to the erroneous use of K-S procedure, although it require a numerically intensive calculation for each data set and model addressed. It is based on bootstrap resampling, a data-based Monte Carlo method that has been mathematically shown to give valid estimates of goodness of fit probabilities under a very wide range of situations.

We now outline the mathematics underlying bootstrap calculations. Let {F( . ; $\theta$) : $\theta \in \Theta$} be a family of continuous distributions parameterized by $\theta$. We want to test whether the univariate data set $X_1$,$X_2$, …., $X_n$ comes from F = F( . ;$\theta$) for some $\theta$ = $\theta_0$. The K-S C-vM and A-D statistics ( and a few other goodness of fit tests) are continuous functional of the process. Yn(x; $\hat{\theta}_n$ ) = $\sqrt{n}(F_n(x) - F(x;\hat{\theta}_n)$. Here $F_n$ denotes the EDF of $X_1$,$X_2$, …, $X_n$ , $\hat{\theta}_n = (X_1, X_2,..., X_n)$ is an estimator of $\theta$ derived from the dataset, and $F(x;\hat{\theta}_n)$ is the model being tested. For a simple example , if {F( . ; $\theta$) : $\theta \in \Theta$} denotes the Gaussian family with $\theta = (\mu, \sigma^2)$, then $\hat{\theta}_n$ can be taken as $(\bar{X}_n, s_n^2)$ where $\bar{X}_n$ is the sample mean and $s_n^2$ is the sample variance based on the data $X_1$, $X_2$ , … , $X_n$ .

In the case of evaluating goodness of fit for a model where the parameters have been estimated from the data, the bootstrap can be computed in two different ways: the parametric bootstrap and the nonparametric bootstrap. The parametric bootstrap may be familiar to a well established technique of creating fake datasets realizing the parametric model by Monte Carlo realizations of the observed EDF using a "random selection with replacement" procedure.

We now outline the mathematics underlying these techniques. Let $\hat{F}_n$ be an estimator of F , based on $X_1$, $X_2$, … , $X_n$ . In order to bootstrap, we generate data $X_1^*, X_2^*,...., X_n^*$ from the estimated population $\hat{F}_n$ and then construct $\hat{\theta}_n^* = \theta_n(X_1^*, X_2^*,...., X_n^*)$

Using the same functional form. For example, if $F(.;\theta)$ is Gaussian with $\theta = (\mu, \sigma^2)$ and if $\hat{\theta} = (\bar{X}, s_n^2)$, then $\hat{\theta}_n^*(\bar{X}_n^*, s_n^{*2})$.

### 3.1 Parametric Bootstrap

The bootstrapping procedure is called parametric if $\hat{F} = F(.;\hat{\theta}_n)$; that is, we generate data $X_1^*, X_2^*, ...., X_n^*$ from the model assuming the estimated parameter values $\hat{\theta}_n$. The process $Y_n^P(x) = \sqrt{n}(F_n^*(x) - F(x;\hat{\theta}_n^*))$ and the sample process $Y_n(x) - \sqrt{n}(F_n(x) - F(x;\hat{\theta}_n))$ converges to the same Gaussian process Y. Consequently , Ln = $\sqrt{n}\sup_x \left| F_n(x) - F(x;\hat{\theta}) \right|$ and

$L_n^* = \sqrt{n}\sup_x \left| F_n^*(x) - F(x;\hat{\theta}_n^*) \right|$ have the same limiting distribution. For the K-S statistic, the critical values of $L_n$ can be derived as follows: construct B resample based on the parametric model ( B $\approx$ 1000 ) , arrange the the resulting $L_n^*$ values in increasing order to obtain 90 or 99 percentile points for getting 90% or 99% critical values. This procedure replaces the incorrect use of the standard probability distribution.

### 3.2 Nonparametric Bootstrap

The nonparametric bootstrap involving resample from the EDF;

$$Y_n^N(x) = \sqrt{n}(F_n^*(x) - F(x;\hat{\theta}_n^*)) - B_n(x)$$

$$= \sqrt{n}(F_n^*(x) - F_n(x) - +F(x;\hat{\theta}_n) - F(x;\hat{\theta}_n^*))$$

Is operationally easy to perform but requires an additional step of bias correction

$$B_n(x) = \sqrt{n}(F_n(x) - F(x;\hat{\theta}_n))$$

The sample process $Y_n$ and the bias corrected nonparametric process $Y_n^N$ converges to the same Gaussian process Y. That is $L_n$ = $\sqrt{n}\sup_x \left| F_n(x) - F(x;\hat{\theta}) \right|$ and

$J_n^* = \sup_x \left| \sqrt{n} F_n^*(x) - F(x; \hat{\theta}_n^*)) - B_n(x) \right|$ have the same limiting distribution. The critical value of the distribution of $L_n$ can then be derived as in the case of parametric bootstrap.

**P Values based on Resampling Methods:** P-values for goodness-of-fit statistics can be obtained by generating the empirical sampling distribution of goodness-of-fit statistics using a resampling method such as the parametric bootstrap method. However, there is strong evidence that parametric bootstrap procedures do not yield accurate P-values. Furthermore, resampling methods may be very time consuming if the researchers is interested in computing the fit of several models.

## 4. Simulation Study

A simulation study is presented in the following to estimate the level and power of the selected normality tests. The effects on the power of the tests due to the sample size, the selected significance level and the type of the alternative distribution are shown with the help of simulation method. The study carried out for seven (n = 10,15,20,25 ,50 and 100 )sample sizes and considering significance levels 0.10,0.05 and 0.01 (for 1 percent level not shown in table due to space ) considering the alternative of nonnormal symmetric and asymmetric . Results obtained are shown in different tables given below. Here, normal observations are generate using Box-Muller(1958) formula and for the other distribution method of inverse integral transformation are used. For each result 10,000 repetitions are made. The ratio of number of test statistic value greater than critical value divided by the total number of repetition gives the empirical level of test statistic under null case and power of the test statistic under the alternative hypothesis .

**Table 1. Empirical levels of test under normal(0,1) Distribution at 0.05 and 0.10 levels**

| Sample Size n | **Test Statistics** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | K-S | | AD | | KS(Bootstrap) | | AD(Bootstrap) | |
| | $\alpha$ =.10 | .05 | .10 | .05 | .10 | .05 | .10 | .05 |
| 10 | .0997 | .0467 | .1025 | .0521 | .0945 | .0475 | .0925 | .0500 |
| 15 | .1036 | .0430 | .1030 | .0527 | .0935 | .0435 | .0900 | .0535 |
| 20 | .1002 | .0548 | .1043 | .0510 | .0950 | .0480 | .1050 | .0485 |
| 25 | .0950 | .0494 | .0976 | .0518 | .1010 | .0530 | .1055 | .0520 |
| 30 | .1086 | .0598 | .0984 | .0526 | .1040 | .0565 | .0980 | .0535 |
| 50 | .0920 | .0480 | .0960 | .0450 | .0870 | .0380 | .0845 | .0410 |
| 100 | .0942 | .0492 | .0970 | .0525 | .1100 | .0545 | .0995 | .0495 |

**Table 2 Empirical power of test Normal Vs Cauchy (0,1) Distribution at 0.05 and 0.10 levels**

| Sample Size n | **Test Statistics** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | K-S | | AD | | KS(Bootstrap) | | AD(Bootstrap) | |
| | $\alpha$ =.10 | .05 | .10 | .05 | .10 | .05 | .10 | .05 |
| 10 | 5274 | .4548 | .7355 | .6580 | .5172 | .4468 | .7243 | .4278 |
| 15 | .5755 | .5037 | .8495 | .7640 | .5634 | .4836 | .8312 | .7526 |
| 20 | .6206 | .5478 | .9050 | .8395 | .6124 | .5372 | .8942 | .8245 |
| 25 | .6512 | .5821 | .9545 | .9070 | .6478 | .5747 | .9456 | .8942 |
| 30 | .6862 | .6152 | .9740 | .9440 | .6804 | .6124 | .9678 | .9376 |
| 50 | .7538 | .6836 | .9995 | .9965 | .7512 | .6818 | .9934 | .9834 |
| 100 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

**Table 3 Empirical power of test Normal Vs logistic (0,1) Distribution at 0.05 and 0.10 levels**

| Sample Size n | Test Statistics K-S | | AD | | KS(Bootstrap) | | AD(Bootstrap) | |
|---|---|---|---|---|---|---|---|---|
| | $\alpha$ =.10 | .05 | .10 | .05 | .10 | .05 | .10 | .05 |
| 10 | .1415 | .0755 | .5955 | .4710 | .1344 | .0662 | .5726 | .4674 |
| 15 | .1435 | .0800 | .7060 | .5860 | .1378 | .0752 | .6784 | .5752 |
| 20 | .1835 | .1080 | .8025 | .6900 | .1778 | .0948 | .7898 | .6856 |
| 25 | .1885 | .1145 | .8680 | .7685 | .1790 | .1064 | .8594 | .7544 |
| 30 | .2055 | .1125 | .9135 | .8465 | .2042 | .8356 | .9034 | .8268 |
| 50 | .2260 | .1370 | .9820 | .9530 | .2212 | .1272 | .9812 | .9510 |
| 100 | .2590 | .1755 | 1.000 | 1.000 | .2578 | .1752 | .9910 | .9890 |

**Table 4 Empirical power of test Normal Vs Exponential ($\lambda = 1$) Distribution at 0.05 and 0.10 levels**

| Sample Size n | Test Statistics K-S | | AD | | KS(Bootstrap) | | AD(Bootstrap) | |
|---|---|---|---|---|---|---|---|---|
| | $\alpha$ =.10 | .05 | .10 | .05 | .10 | .05 | .10 | .05 |
| 10 | .5410 | .3920 | 1.000 | .9880 | .5328 | .3841 | 1.000 | .9765 |
| 15 | .7040 | .5310 | 1.000 | 1.000 | .6878 | .5244 | 1.000 | 1.000 |
| 20 | .7920 | .7195 | 1.000 | 1.000 | .7856 | .7054 | 1.000 | 1.000 |
| 25 | .8660 | .7805 | 1.000 | 1.000 | .8566 | .7784 | 1.000 | 1.000 |
| 30 | .9275 | .8745 | 1.000 | 1.000 | .9128 | .8646 | 1.000 | 1.000 |
| 50 | .9900 | .9705 | 1.000 | 1.000 | .9884 | .9674 | 1.000 | 1.000 |
| 100 | 1.0 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

**Table 5. Empirical power of test Normal Vs lognormal (0,1) Distribution at 0.05 and 0.10 levels**

| Sample Size n | K-S $\alpha$ =.10 | K-S .05 | AD .10 | AD .05 | KS(Bootstrap) .10 | KS(Bootstrap) .05 | AD(Bootstrap) .10 | AD(Bootstrap) .05 |
|---|---|---|---|---|---|---|---|---|
| 10 | .6880 | .5610 | 1.000 | 1.000 | .6742 | .5578 | 1.000 | .9855 |
| 15 | .8495 | .7505 | 1.000 | 1.000 | .8365 | .7455 | 1.000 | 1.000 |
| 20 | .9080 | .8650 | 1.000 | 1.000 | .8884 | .8568 | 1.000 | 1.000 |
| 25 | .9520 | .9200 | 1.000 | 1.000 | .9478 | .9146 | 1.000 | 1.000 |
| 30 | .9815 | .9625 | 1.000 | 1.000 | .9776 | .9568 | 1.000 | 1.000 |
| 50 | 1.000 | .9985 | 1.000 | 1.000 | 1.000 | .9924 | 1.000 | 1.000 |
| 100 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

## 5. Conclusion

We investigate the KS statistics and Anderson-Darling statistics when location and scale parameters are unknown and studied the power of both the exact and bootstrap test against different alternatives, viz. Cauchy ,exponential, lognormal and logistic distribution. The tables presented here lead to the following conclusions.

1. The estimates of powers of exact tests are lightly larger than those of the bootstrap tests in case of K-S test, but it is hard to say which test is more powerful because the estimating are subject to variability. In case of Anderson –Darling test both the exact and bootstrap based test power are almost similar.

2. The difference between the powers of two tests gets smaller as the sample size n increases. When n>10 both tests have very similar powers. This conclusion is also verified by the strong correlation between the exact and bootstrap p-values.

3. Both kinds of tests appear to be unbiased when the sample sizes is large enough.

From the study we may conclude that test based on bootstrap technique be used for goodness of fit of normality without any hesitation. More work can be done using bootstrap technique to know the performance of the tests based on bootstrap with various situations.In future we hope extend more research on this line.

**References:**


Anderson,T.W. Darling, D.A. (1952): Asymptotic theory of certain “goodness of fit” criteria based on stochastic processes, Ann. Math. Stat. 23(2),193-212.

Anderson,T.W. Darling, D.A. (1954): A test of goodness of fit , Jour. Amer. Stat. Assoc. 49(268),765-769.

Babu,G.J. and Rao,C.R.(2004): Goodness of fit tests when parameters are estimated. Sankhya,66,1,63-74.

Bartholomew, D.J. & Tzamourani, P.(1999) : The goodness-of-fit of latent trait models in attitude measurement. Sociological Methods and Research, 27, 525-546.

Bollen, K. A. & Stine, R. A. (1993): Bootstrapping Goodness-of-fit Measures in Structural Equation Models in Bollen, K. Long, J. Testing Structural Equestion Models, 111-135

Devier Matthias von (1997): Bootstrapping Goodness-of-fit Statistics for Sparse Categorical Data- Results of a Monte Carlo Study. Methods of Psychological Research Online, 29-48.

Efron, B. (1979): Bootstrap Methods:Another look at the Jackknife. Ann. Statistics, 7, 1-26.

Efron, B. & Tibshirani,R.J.(1993): An Introduction in the Bootstrap.Chapman & Hall,New York.

Fleming, S.E. (1979): The Use of Chi-squared Statistics for Categorical Data problems. Journal of the Royal Statistical Society, Series B 41(1), 54-64.

Giles,D. (2007): Spurious regressions with time series data: Further asumptotic results, Comm. Stat. Theory and Methods 36(5),967-979.

Kolmogorov,A.N.(1933): Sulla determinazione empirica di una legge di distribuzione,Giornale dell’ Istituto degli Attuari 4, 83-91.

Langeheine, R.,Pannekoek, J.& von de Pol, F.(1996): Bootstrapping Goodness-of-fit Measures in Categorical Data Analysis. Sciological Methods and Research, 24(4), 492-516.

Lilliefors,H.W.(1967): On Kolmomogorov-Smirnov Test for Normality with Mean and Variance Unknown. Jour. Amer. Statist. Assoc. Vol. 62,No. 318,339-402.

Lisenfeld,R and Jung ,R.c. (2000): Stichastic volatility models: conditional normality versus heavy tailed distributions,Jour. App.Econom.,15(2),137-160.

Min, I. (2007): A nonparametric test of the conditional normality of housing demand, Appl.Econ. Lett 14(2),105-109.

Olivares Alberto Maydeu & Joe Harry (2008): An Overview of Limited Information Goodness-of-fit Testing in Multidimensional Contingency Tables. New trends in psychometrics, 253-262.

Onder,A. and Zaman, A(2005): Robust tests of normality of errors in regression models, Econom. Lett.86(1)63-68.

Quddus,M.A.(2008):Time series count data models:An empirical application to traffic accidents,Accid. Anal. And Prev. 40(5),1732-1741.

Schoder,V.,Himmelmann,A. and Wilhelm,K.P.(2006): Preliminary testing for normality: Some statistical aspects of a common concept.

Smirnov,N.V.(1939):Estimate of deviation between empirical distributions,(russians). Bulletin Moscow University 2,3-36.

Stephens,M.A.(1974): EDF Statistics for Goodness of Fit and Some Comparisons, Jour.Amer. Statist. Assoc.,vol. 69,730-737.

Surucu,B. abd Koc, E. (2007): Assessing the validity of a statistical distribution:Some illustrative example from dermatological research, clin.Exp.Dermatol 33(3),239-242.

Blake, T. (2005): The Bootstrap Theorem: Creating Empirical Distributions. E- Journal.

Thadewald,T and Buning, H.(2007): Jarque-Bera test and its competitors for test of normality-A power comparison, Jour. Appl. Stat. 34(1),87-105.

Wilde,J.(2008): A simple repression of Bera-Jarque –Lee test for probit models, Econom. Lett.,101(2),110-121.