Goodness-of-Fit Test Based on Arbitrarily Selected Order Statistics

Zhenmin Chen

Department of Mathematics and Statistics, Florida International University, Miami, FL33199, USA
*Corresponding Author: chenzh@fiu.edu

Abstract   Checking whether or not the population distribution, from which a random sample is drawn, is a specified distribution is a popular topic in statistical analysis. Such a problem is usually named as goodness-of-fit test. Numerous research papers have been published in this area. The purpose of this short paper is to provide a goodness-of-fit test statistic which works for many kinds of censored data formed by order statistics. This is an extension of the work presented in Chen and Ye (2009). The method can be used for censoredsamplesthat are commonly used in reliability analysis including left censored data, right censored data and doubly censored data.

Keywords   Goodness-of-Fit Test, Censored Data, Arbitrarily Selected Order Statistics, Uniformity

1. Introduction

The goodness-of-fit test has its long history. The primary goal of the goodness-of-fit test is to check how well a specificstatistical model can fit a given data set. The χ^2 test (Pearson (1900)), Kolmogorov-Smirnov test (Kolmogorov (1933) and Smirnov (1939)), Cramer-von Mises test (Cramer (1928)),and Anderson-Darling test (Anderson and Darling (1952)) are the statistical testspresented in early papers and are still widely used by statistics practitioners. All these test statistics are adopted by most statistical software. The Shapiro and Wilk test (Shapiro and Wilk (1965)) is another commonly used test statistic which works specifically for the normal distribution and lognormal distribution. In the recent years, many research papers have been published in the area of goodness-of-fit test. The power of the goodness-of-fit tests has been compared by many authors. See, for examples, Choulakian, Lockhart and Stephens (1994), and Steele and Chaseling (2006). Chen and Ye (2009) proposed a test statistic for checking whether or not the population distribution, from which a random sample is drawn, is a uniform distribution. It has been shown in that paper that the power of the proposed test in that paper is higher than some of the existing test statistics in some cases,

especially for the case that the alternative distributions are V-shaped distributions. In this short paper, the method used in Chen and Ye (2009) will be extended to censored samples. The new test statistic can be used when only some order statistics are available.

2. Uniformity Test Based on Order Statistics

The purpose of the uniformity test is to check whether or not the population distribution, from which a random sample was drawn, is a uniform distribution on interval $[0.1]$. Suppose $X_1, X_2, \cdots, X_n$ form a random sample from a population distribution with support set $[0,1]$. Suppose also that $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ are the corresponding order statistics. For the complete sample case, atest statistic was proposed in Chen and Ye (2009). The test statistic has the form

$$G(X_1, X_2, \cdots, X_n) = \frac{(n+1)\sum_{i=1}^{n+1}\left(X_{(i)} - X_{(i-1)} - \frac{1}{n+1}\right)^2}{n}. \quad (1)$$

Here $X_{(0)}$ is defined as 0, and $X_{(n+1)}$ is defined as 1. Chen and Ye (2009) discussed the properties of this test statistic. The expected value, variance and the shape of this test statistic are described in that paper. When the population distribution is the same as the specified distribution, the value of $G(X_1, X_2, \cdots, X_n)$ should be pretty close to 0. On the other hand, when the population distribution is far away from the specified distribution, the value of $G(X_1, X_2, \cdots, X_n)$ should be pretty close to 1. The quantiles of the test statistics were obtained by Monte Carlo simulation and were tabulated. In order to let statistics users find the quantile values easily, the quantile value for different sample sizes and different significance levels are listed in Table 1. The quantiles can be used to conduct goodness-of-fit test discussed in this paper. In fact, the hypothesis of uniformity should be rejected at significant level α if

$$G(X_1, X_2, \cdots, X_n) > G_{1-\alpha}.$$

Table 1.   Critical Values of Gstatistic

n	$G_{0.900}$	$G_{0.950}$	$G_{0.990}$	$G_{0.995}$	$G_{0.999}$
2	0.5360	0.6543	0.8349	0.8815	0.9459
3	0.3932	0.4895	0.6751	0.7358	0.8397
4	0.3126	0.3862	0.5498	0.6109	0.7280
5	0.2595	0.3185	0.4568	0.5135	0.6278
6	0.2212	0.2702	0.3874	0.4371	0.5448
7	0.1924	0.2337	0.3337	0.3777	0.4762
8	0.1699	0.2053	0.2922	0.3310	0.4208
9	0.1520	0.1829	0.2591	0.2935	0.3741
10	0.1373	0.1644	0.2316	0.2625	0.3361
11	0.1251	0.1492	0.2090	0.2363	0.3033
12	0.1149	0.1364	0.1903	0.2153	0.2763
13	0.1060	0.1255	0.1739	0.1963	0.2511
14	0.0984	0.1160	0.1599	0.1805	0.2311
15	0.0918	0.1079	0.1482	0.1669	0.2130
16	0.0859	0.1007	0.1375	0.1546	0.1971
17	0.0808	0.0943	0.1282	0.1440	0.1834
18	0.0761	0.0887	0.1200	0.1347	0.1718
19	0.0720	0.0837	0.1128	0.1264	0.1605
20	0.0683	0.0791	0.1062	0.1188	0.1506
21	0.0649	0.0751	0.1003	0.1121	0.1418
22	0.0618	0.0714	0.0950	0.1060	0.1336
23	0.0590	0.0680	0.0901	0.1005	0.1266
24	0.0564	0.0649	0.0856	0.0952	0.1197
25	0.0541	0.0620	0.0816	0.0907	0.1139
26	0.0519	0.0594	0.0778	0.0864	0.1084
27	0.0499	0.0570	0.0745	0.0826	0.1032
28	0.0480	0.0547	0.0713	0.0790	0.0986
29	0.0462	0.0527	0.0684	0.0757	0.0943
30	0.0446	0.0507	0.0656	0.0726	0.0902
31	0.0431	0.0489	0.0631	0.0698	0.0865
32	0.0416	0.0472	0.0608	0.0671	0.0830
33	0.0403	0.0456	0.0586	0.0646	0.0799
34	0.0390	0.0441	0.0565	0.0623	0.0767
35	0.0379	0.0427	0.0545	0.0600	0.0740
36	0.0367	0.0414	0.0528	0.0580	0.0715
37	0.0357	0.0402	0.0510	0.0560	0.0689
38	0.0347	0.0390	0.0494	0.0543	0.0666
39	0.0337	0.0379	0.0479	0.0526	0.0644
40	0.0328	0.0368	0.0465	0.0509	0.0621
41	0.0320	0.0358	0.0450	0.0493	0.0601
42	0.0312	0.0349	0.0438	0.0479	0.0583
43	0.0304	0.0340	0.0426	0.0466	0.0567
44	0.0297	0.0331	0.0414	0.0452	0.0548
45	0.0289	0.0323	0.0403	0.0440	0.0532
46	0.0283	0.0315	0.0393	0.0428	0.0518
47	0.0276	0.0308	0.0383	0.0417	0.0504
48	0.0270	0.0301	0.0373	0.0406	0.0490
49	0.0264	0.0294	0.0364	0.0396	0.0477
50	0.0259	0.0287	0.0355	0.0386	0.0465

This paper focuses on the case that the samples are incomplete. In practice, the available samples may be censored ones. For example, in reliability analysis, the statistical analysis is usually based on left censored samples, right censored samples, or doubly censored samples. To fit the need of such kind of applications and even more generalized situations, it is assumed in this paper that only some of the order statistics are available for testing the hypotheses mentioned above. More specifically, it is assumed that only k order statistics

$$X_{(i_1)} \le X_{(i_2)} \le \cdots \le X_{(i_k)}$$

are available when the test is conducted. Here $1 \le i_1 < i_2 < \cdots < i_k \le n$ are k integers that are arbitrarily picked from set $\{1, 2, \cdots, n\}$. In such a case, the test statistic can be defined as

$$G\left(X_{(i_1)}, X_{(i_2)}, \cdots, X_{(i_k)}\right) = \frac{(n+1)^2}{(n^2-1) + \sum_{j=1}^{k+1}(i_j - i_{j-1})^2} \sum_{j=1}^{k+1}\left(X_{(i_j)} - X_{(i_{j-1})} - \frac{i_j - i_{j-1}}{n+1}\right)^2 \quad (2)$$

for any $1 \le i_1 < i_2 < \cdots < i_k \le n$. Here i_0 is defined as 0, i_{k+1} is defined as $n+1$, The test statistic defined in (2) possesses some properties. It can be seen that $0 \le G\left(X_{(i_1)}, X_{(i_2)}, \cdots, X_{(i_k)}\right) \le 1$. This is because

$$0 \le \sum_{j=1}^{k+1}\left(X_{(i_j)} - X_{(i_{j-1})} - \frac{i_j - i_{j-1}}{n+1}\right)^2 \le \frac{n-1}{n+1} + \frac{1}{(n+1)^2}\sum_{j=1}^{k+1}(i_j - i_{j-1})^2.$$

To show $\sum_{j=1}^{k+1}\left(X_{(i_j)} - X_{(i_{j-1})} - \frac{i_j - i_{j-1}}{n+1}\right)^2 \le \frac{n-1}{n+1} + \frac{1}{(n+1)^2}\sum_{j=1}^{k+1}(i_j - i_{j-1})^2$, note that

$$\sum_{j=1}^{k+1}\left(X_{(i_j)} - X_{(i_{j-1})} - \frac{i_j - i_{j-1}}{n+1}\right)^2$$

$$= \sum_{j=1}^{k+1}\left(X_{(i_j)} - X_{(i_{j-1})}\right)^2 + \sum_{j=1}^{k+1}\left(\frac{i_j - i_{j-1}}{n+1}\right)^2 - 2\sum_{j=1}^{k+1}\left(X_{(i_j)} - X_{(i_{j-1})}\right)\left(\frac{i_j - i_{j-1}}{n+1}\right)$$

$$\le \sum_{j=1}^{k+1}\left(X_{(i_j)} - X_{(i_{j-1})}\right)^2 + \frac{1}{(n+1)^2}\sum_{j=1}^{k+1}(i_j - i_{j-1})^2 - \frac{2}{n+1}\sum_{j=1}^{k+1}\left(X_{(i_j)} - X_{(i_{j-1})}\right)^2$$

$$= \frac{n-1}{n+1}\sum_{j=1}^{k+1}\left(X_{(i_j)} - X_{(i_{j-1})}\right)^2 + \frac{1}{(n+1)^2}\sum_{j=1}^{k+1}(i_j - i_{j-1})^2$$

$$\le \frac{n-1}{n+1}\left(\sum_{j=1}^{k+1}\left(X_{(i_j)} - X_{(i_{j-1})}\right)\right)^2 + \frac{1}{(n+1)^2}\sum_{j=1}^{k+1}(i_j - i_{j-1})^2 \le \frac{n-1}{n+1} + \frac{1}{(n+1)^2}\sum_{j=1}^{k+1}(i_j - i_{j-1})^2.$$

It can also be seen that

$$E\left(X_{(i)} - X_{(i-1)}\right) = \frac{1}{n+1}$$

for any $i = 1, 2, \cdots, n+1$. It implies that

$$E\left(X_{(i_j)} - X_{(i_{j-1})}\right) = \frac{i_j - i_{j-1}}{n+1} \quad (j = 1, 2, \cdots, k+1).$$

Therefore, the value of $G\left(X_{(i_1)}, X_{(i_2)}, \cdots, X_{(i_k)}\right)$ should be quite close to zero if the population distribution is the uniform distribution on interval $[0,1]$. On the other hand, if the value of $G\left(X_{(i_1)}, X_{(i_2)}, \cdots, X_{(i_k)}\right)$ is too far away from zero, it could be an indication that the cumulate distribution function of the underlying distribution is not a uniform distribution on $[0,1]$.

Then H_0 should be rejected at significant level α if

$$G\left(X_{(i_1)}, X_{(i_2)}, \cdots, X_{(i_k)}\right) > G_{1-\alpha}\left(i_1, i_2, \cdots, i_k\right).$$

Here $G_{1-\alpha}\left(i_1, i_2, \cdots, i_k\right)$ is a number such that

$$P\left(G\left(X_{(i_1)}, X_{(i_2)}, \cdots, X_{(i_k)}\right) > G_{1-\alpha}\left(i_1, i_2, \cdots, i_k\right)\right) = \alpha.$$

For the complete sample case, the critical values computed in Chen and Ye (2009) can be adopted to conduct the statistical test discussed in this section. For such a general case, a simple computer program is needed to run Monte Carlo simulation to find the critical values of the test statistic.

3. Test for General Distributions

Now suppose $X_1, X_2, \cdots, X_n$ form a random sample from a population distribution with cumulative distribution function $F_0(x)$. Suppose also that $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ are the corresponding order statistics. The purpose is to test

H_0: The cumulative distribution function of the population distribution is $F_0(x)$,

H_a: The cumulative distribution function of the population distribution is not $F_0(x)$.

It is well known that

$$F_0(X_1), F_0(X_2), \ldots, F_0(X_n)$$

form a random sample from the $U[0,1]$ distribution. Then

$$F_0(X_{(1)}), F_0(X_{(2)}), \ldots, F_0(X_{(n)})$$

are the ordered observations of a random sample from the $U[0,1]$ distribution because of the monotonic property of the cumulative distribution function. Therefore, testing whether or not the cumulative distribution of the population, from which $X_1, X_2, \cdots, X_n$ are sampled, is $F_0(x)$ is the same as testing whether or not the population distribution, from which $F_0(X_1), F_0(X_2), \ldots, F_0(X_n)$ are sampled, is the uniform distribution on $[0,1]$. For the complete sample case, the test statistic can be defined as

$$G(X_1, X_2, \cdots, X_n) = \frac{(n+1)\sum_{i=1}^{n+1}\left(F_0(X_{(i)}) - F_0(X_{(i-1)}) - \frac{1}{n+1}\right)^2}{n} \tag{3}$$

Here $X_{(0)}$ is defined as $-\infty 0$, and $X_{(n+1)}$ is defined as $+\infty$. The hypothesis In fact, H_0 should be rejected at significant level α if $G(X_1, X_2, \cdots, X_n) > G_{1-\alpha}$.

Now it is assumed that only k order statistics

$$X_{(i_1)} \le X_{(i_2)} \le \cdots \le X_{(i_k)}$$

are available when the test is conducted. Here $1 \le i_1 < i_2 < \cdots < i_k \le n$ are k integers that are arbitrarily picked from set $\{1, 2, \cdots, n\}$. Then the test statistic can be defined as

$$G\left(X_{(i_1)}, X_{(i_2)}, \cdots, X_{(i_k)}\right) = \frac{(n+1)^2}{(n^2-1)+\sum_{j=1}^{k+1}(i_j - i_{j-1})^2}\sum_{j=1}^{k+1}\left(F_0\left(X_{(i_j)}\right) - F_0\left(X_{(i_{j-1})}\right) - \frac{i_j - i_{j-1}}{n+1}\right)^2 \tag{4}$$

for any $1 \le i_1 < i_2 < \cdots < i_k \le n$. Here i_0 is defined as 0, and i_{k+1} is defined as $n+1$. Consequently, $X_{(i_0)}$ is defined

as $-\infty$ and $X_{(i_{k+1})}$ is defined as $+\infty$. The null hypothesis H_0 should be rejected at significance level α if

$$G\left(X_{(i_1)}, X_{(i_2)}, \cdots, X_{(i_k)}\right) > G_{1-\alpha}\left(i_1, i_2, \cdots, i_k\right).$$

The value of $G_{1-\alpha}\left(i_1, i_2, \cdots, i_k\right)$ can be obtained using Monte Carlo simulation for any combination of $i_1, i_2, \cdots, i_k$ and α.

For example, suppose the test mentioned above is used to check if the population distribution of a data set is a three-parameter Weibull distribution with location parameter μ_0, shape parameter β_0 and scale parameter η_0. The cumulative distribution function of the three-parameter Weibull distribution is

$$F_0(x) = 1 - \exp\left\{-\left(\frac{x - \mu_0}{\eta_0}\right)^{\beta_0}\right\} \quad (x > \mu_0).$$

Then the formula to calculate the value of the test statistic defined in (3) is

$$G(X_1, X_2, \cdots, X_n) = \frac{n+1}{n}\left[\left(\frac{n}{n+1} - \exp\left\{-\left(\frac{X_{(1)} - \mu_0}{\eta_0}\right)^{\beta_0}\right\}\right)^2 + \sum_{i=2}^{n}\left(F_0(X_{(i)}) - F_0(X_{(i-1)}) - \frac{1}{n+1}\right)^2\right.$$

$$\left. + \left(\exp\left\{-\left(\frac{X_{(n)} - \mu_0}{\eta_0}\right)^{\beta_0}\right\} - \frac{1}{n+1}\right)^2\right]$$

for complete samples. The calculated value of $G(X_1, X_2, \cdots, X_n)$ can then compared with the critical value listed in Table 1 to determine whether or not to reject the null hypothesis that the population distribution is the three-parameter Weibull distribution.

For the samples containing k order statistics only, the formula to calculate the value of the test statistic defined in (4) becomes

$$G\left(X_{(i_1)}, X_{(i_2)}, \cdots, X_{(i_k)}\right) = \frac{(n+1)^2}{(n^2-1) + \sum_{j=1}^{k+1}(i_j - i_{j-1})^2}\left[\left(\frac{n+1-i_1}{n+1} - \exp\left\{-\left(\frac{X_{(i_1)} - \mu_0}{\eta_0}\right)^{\beta_0}\right\}\right)^2\right.$$

$$\left. + \sum_{j=1}^{k}\left(F_0\left(X_{(i_j)}\right) - F_0\left(X_{(i_{j-1})}\right) - \frac{i_j - i_{j-1}}{n+1}\right)^2 + \left(\exp\left\{-\left(\frac{X_{(i_k)} - \mu_0}{\eta_0}\right)^{\beta_0}\right\} - \frac{n+1-i_k}{n+1}\right)^2\right].$$

A simple computer program is needed to find the critical value $G_{1-\alpha}\left(i_1, i_2, \cdots, i_k\right)$ using Monte Carlo simulation so that the decision of the statistical test can be made. The null hypothesis that the population distribution is the three-parameter Weibull distribution should be rejected at significance level if the calculated value of $G\left(X_{(i_1)}, X_{(i_2)}, \cdots, X_{(i_k)}\right)$ exceeds the critical value $G_{1-\alpha}\left(i_1, i_2, \cdots, i_k\right)$.

4. Conclusion and Discussion

In this short paper, a goodness-of-fit test statistic is proposed for checking whether or not the cumulative distribution function of a population distribution has a specified form for the case that sampled data are arbitrarily censored. More specifically, the test is performed based on order statistics $X_{(i_1)} \leq X_{(i_2)} \leq \cdots \leq X_{(i_k)}$ for arbitrarily selected $1 \leq i_1 < i_2 < \cdots < i_k \leq n$. Such a setting covers various censoring cases used in reliability analysis and other fields. Because of its flexibility in regard of selection of sample items, the method can be used for many complicated cases. For example, in life test, the data collectors may collect failure times during one time period and leave the test items running without observing the failure times in the second time period. Then the data collectors may come back and record failure times in the third time period. The procedure can continue with the same pattern. The method presented in this paper can be used for this kind of data without any technical difficulties. It has been shown that the value of the proposed test statistic is always between 0 and 1. The critical values of the proposed test statistic can be obtained by Monte Carlo simulation. The test statistic discussed in Chen and Ye (2009) is for univariate uniformity. By applying the probability integral transformation, the test can be used to check if the cumulative distribution of a population distribution is of any specified distribution. It is worth being noted here is that the test statistic discussed in this paper is for testing whether or not the population distribution, from which a random sample is drawn, is a specified distribution. It means that everything about the distribution is known. In practice, however, sometimes it is desired to test if the population distribution is of certain type when the parameters of the distribution are unknown. In that case, it is believed that the method suggested by Lilliefors (Lilliefors (1967) and Lilliefors (1969)) can be used.

REFERENCES

[1] Anderson, T. W. and Darling, D. A. (1952) Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Process, *Annals of Mathematical Statistics*, 23, 193-212.

[2] Chen, Zhenmin and Ye, Chunmiao (2009) An alternative test for uniformity, *International Journal of Reliability, Quality and Safety Engineering*, 16, 343-356.

[3] Choulakian, V., Lockhart, R. A. and Stephens, M. A. (1994) Cramer-von Mises Statistics for Discrete Distributions, *The Canadian Journal of Statistics*, 22, 125-137.

[4] Cramer, H. (1928) On the Composition of Elementary Errors, *SkandinaviskAktuarietidskrift*, 11,13-74, 141-180.

[5] Kolmogorov, A. N. (1933) Sulla Determinazione Empirica di Una Legge di Distribuzione, *Giornale dell' Istituto degli Attuari*, 4, 83-91.

[6] Lilliefors, W. H. (1967) On the Kolmogorov-Smirnov test for normality with mean and variance unknown, *Journal of the American Statistics Association*, 62, 399-402.

[7] Lilliefors, W. H. (1969) On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown, *Journal of the American Statistics Association*, 64, 387-389.

[8] Pearson, K. (1900) On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be Reasonably Supposed to Have Arisen from Random Sampling, *Philosophical Magazine*, 5, 157–175.

[9] Shapiro, S. S. and Wilk, M. B. (1965) An analysis of variance test for normality (complete samples), *Biometrika*, 52, 591-611.

[10] Smirnov, N. V. (1939) Estimate of Deviation Between Empirical Distributions (Russian), *Bulletin Moscow University,* 2, 3-16.

[11] Steele, M. and Chaseling, J. (2006) Powers of Discrete Goodness-of-Fit Test Statistics for a Uniform Null Against a Selection of Alternative Distributions, *Communications in Statistics – Simulation and Computation*, 35, 1067-1075.