Testing Normality Based on Kullback-Leibler Information With Progressively Type-II Censored Data

Arezou HABIBI RAD *
Department of Statistics, School of Mathematical Sciences,
Ferdowsi University of Mashhad, P.O. Box 1159, Mashhad 91775, Iran.

Abstract

We will use the joint entropy of progressively censored order statistics in terms of an incomplete integral of the hazard function, and provide a simple estimate of the joint entropy of progressively Type-II censored data, has been introduced by Balakrishnan et al. (2007). Then We construct a goodness-of-fit test statistic based on Kullback-Leibler information for normal distribution. Finally, by using Monte Carlo simulations, the power of the test is estimated and compared against several alternatives under different progressive censoring schemes.

***Keywords*: Entropy; Hazard function; Monte Carlo simulation; Order statistics; Progressively Type-II censored data; Test of normality.**

*e-mail address: habibi@math.um.ac.ir>

Notation

n	sample size.
$X_{i:n}$	i-th order statistic from a sample of size n.
$m,\ (m \leq n)$	the number of complete failures in a progressively censored sample.
R_i	the number of surviving units censored at the time of the i-th failure.
$X_{i:m:n}^{(R_1,\cdots,R_m)}$	i-th progressively censored order statistic; for convenience, we will use the simplified notation $X_{i:m:n},\ i = 1, 2, \cdots, m$.
$f_{X_{r:m:n}}$	the probability density function (p.d.f.) of $X_{r:m:n}$.
$f_{X_{1:m:n},\cdots,X_{m:m:n}}$	the joint probability density function of all m progressively Type-II censored order statistics.
$h(x)$	the hazard function (rate), $f(x)/(1 - F(x))$.
$H_{1\cdots m:m:n}$	the joint entropy of $X_{1:m:n}, \cdots, X_{m:m:n}$.
$I_{1\cdots m:m:n}(f : g)$	the Kullback-Leibler information of $X_{1:m:n}, \cdots, X_{m:m:n}$.
$f_{X_{r:m:n}\vert X_{r-1:m:n}}(x_r\vert x_{r-1})$	the conditional p.d.f. of $X_{r:m:n}$, given $X_{r-1:m:n} = x_{r-1:m:n}$.
$H_{r\vert r-1:m:n}$	the expectation of conditional entropy of $X_{r:m:n}$, given $X_{r-1:m:n} = x_{r-1:m:n}$.

1 INTRODUCTION

Suppose a random variable X has a distribution function $F(x)$ and a continuous density function $f(x)$. The differential entropy $H(f)$ of the random variable is defined in Shannon (1948), to be

$$H(f) = -\int_{-\infty}^{\infty} f(x) \log f(x) dx. \tag{1}$$

The first time, the test normality performed based on sample entropy by Vasicek (1976) and the power compared with some leading test statistics for complete samples.
The entropy difference $H(f) - H(g)$ has been considered in Dudewicz et al. (1981) and Gokhale (1983) for establishing goodness-of-fit tests for the class of the maximum entropy distributions.

The Kullback-Leibler (KL) information in favor of $g(x)$ against $f(x)$ is defined in Kullback (1959) to be

$$I(g : f) = \int_{-\infty}^{\infty} g(x) \log \frac{g(x)}{f(x)} dx,$$

which is an extended concept of entropy.
Because $I(g : f)$ has the property that $I(g : f) \geq 0$, and the equality holds if $g = f$, the estimate of the KL information has also been considered as a goodness-of-fit test statistic by some authors including Arizono et al. (1989) and Ebrahimi et al. (1992), for complete samples. Park (2005) and Balakrishnan et al. (2007), respectively, for Type-II censored data and progressively Type-II censored data.
Now, in this paper we will extend the goodness-of-fit test based on KL information with progressively Type-II censored data for normality.

The rest of the paper is organized as follows: In Section 2 as Preliminary, we introduce Type-II progressive censoring data, the joint entropy of progressively censored data in terms of the hazard function and the nonparametric estimate of the joint entropy. In Section 3, we define the KL information for progressively Type-II censored data and propose a goodness-of-fit test for normality based on KL information. In Section 4, we use Monte Carlo simulations to evaluate the power under different Type-II progressive censoring schemes. Finally, in Section 4, we present an illustrative example.

2 PRELIMINARY

2.1 Progressively Type-II Censored Data

Suppose n identical items are placed on a life-testing experiment. Assume that their life-times are independent and identically distributed with probability distribution function (cdf) $F(x;\underline{\theta})$ and probability density function (pdf) $f(x;\underline{\theta})$, where θ is a vector of parameters.

There are several scenarios in life-testing and reliability experiments in which units that are subject to test are lost or removed from the experiment before failure. Such units are usually called the censored unites. The two most common censoring schemes are termed as conventional Type-I and Type-II censoring schemes which are extensively studied in statistical and reliability literature, Balakrishnan and Cohen (1991). Briefly, they can be described as follows: Consider n items under observations in a particular experiment. In the conventional Type-I censoring scheme, the experiment continues up a pre-specified time T. The conventional Type-II censoring scheme requires the experiment to continue until a pre-specified number of failures $m(\leq n)$ occur.

One of the drawbacks of the conventional Type-I, Type-II censoring schemes is that they do not allow for removal of units at points other than the terminal point of the experiment. One censoring scheme known as Type-II progressive censoring scheme, which has this advantage, so it becomes very popular for the last few years. It can be described as follows: Consider n units in a study and suppose $m(\leq n)$ is fixed before the experiment. Moreover, m other integers, $R_1, \cdots R_m$ are also fixed before so that $R_1 + \cdots + R_m + m = n$. At the time of the first failure, say $X_{1:m:n}$, R_1 of the remaining units are randomly removed. Similarly, at the time of the second failure, say $X_{2:m:n}$, R_2 of the remaining units are randomly removed and so on. Finally, at the time of the $m-th$ failure, say $X_{m:m:n}$, the rest of the R_m units are removed. Fore further details on Type-II progressive censoring, refer to Balakrishnan and Aggarwala (2000).

The joint probability density function (pdf) of all m progressively Type-II censored order statistics $(X_{1:m:n}, \cdots X_{m:m:n})$ which is define in Balakrishnan (2000) to be

$$f_{X_{1:m:n},X_{2:m:n},\cdots,X_{m:m:n}}(x_1, x_2, \cdots, x_m) = c\prod_{i=1}^{m} f(x_i)\{1 - F(x_i)\}^{R_i}, \quad x_1 < x_2 < \cdots < x_m,$$

where

$$c = n(n - R_1 - 1)\cdots(n - R_1 - R_2 - \cdots - R_{m-1} - m + 1).$$

2.2 Entropy of Progressively Censored Data in Terms of the Hazard Function

The joint entropy of $X_{1:m:n}, \cdots, X_{m:m:n}$ defined in literature (Park, 2005), to be

$$\begin{aligned} H_{1\cdots m:m:n} &= -\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{x_{2:m:n}} f_{X_{1:m:n},X_{2:m:n},\cdots,X_{m:m:n}}(x_1, x_2, \cdots, x_m) \\ &\times \log f_{X_{1:m:n},X_{2:m:n},\cdots,X_{m:m:n}}(x_1, x_2, \cdots, x_m) dx_{1:m:n} \cdots dx_{m:m:n}, \end{aligned}$$

where $f_{X_{1:m:n},X_{2:m:n},\cdots,X_{m:m:n}}(x_1, x_2, \cdots, x_m)$ is the joint pdf of all m progressively Type-II censored order statistics.

$H_{1\cdots m:m:n}$ is an m-dimensional integral, and we need to simplify this multiple integral.

The simple calculation of the entropy of the usual single and consecutive order statistics has been studied in Wong et al. (1990) and Park (1995). The multiple integral of the entopy for Type-II censored data be simplified to a single-integral by Park (2005) and the joint entropy of progressively Type-II censored order statistics in terms of an incomplete integral of the hazard function , $h(x)$, has been simplified by Balakrishnan et al. (2007),

$$H_{1\cdots m:m:n} = -\log c + n\overline{H}_{1\cdots m:m:n},$$

where

$$\overline{H}_{1\cdots m:m:n} = \frac{m}{n} - \frac{1}{n}\int_{-\infty}^{\infty} \sum_{i=1}^{m} f_{X_{i:m:n}}(x) \log h(x) dx.$$

2.3 Nonparametric Entropy Estimate

The nonparametric estimate of the joint entropy ($H_{1\cdots m:m:n}$) was obtained, as

$$H_{1\cdots m:m:n}(w, n, m) = -\log c + nH(w, n, m),$$

where

$$H(w, n, m) = \frac{1}{n}\sum_{i=1}^{m} \log\left(\frac{(x_{i+w:m:n} - x_{i-w:m:n})}{E(U_{i+w:m:n}) - E(U_{i-w:m:n})}\right) - (1 - \frac{m}{n})\log(1 - \frac{m}{n}).$$

3 TESTING NORMALITY BASED ON THE KULLBACK-LEIBLER INFORMATION

3.1 Kullback-Leibler Information and Test Statistic

For a null density function $f^0(x; \theta)$, the KL information from a progressively Type-II censored data is given by

$$\begin{aligned} I_{1\cdots m:m:n}(f : f^0) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{x_{2:m:n}} f_{X_{1:m:n},X_{2:m:n},\cdots,X_{m:m:n}}(x_1, x_2, \cdots, x_m; \theta) \\ &\times \log \frac{f_{X_{1:m:n},X_{2:m:n},\cdots,X_{m:m:n}}(x_1, x_2, \cdots, x_m; \theta)}{f^0_{X_{1:m:n},X_{2:m:n},\cdots,X_{m:m:n}}(x_1, x_2, \cdots, x_m; \theta)} dx_1 \cdots dx_m, \end{aligned}$$

where $f_{X_{1:m:n},X_{2:m:n},\cdots,X_{m:m:n}}(x_1, x_2, \cdots, x_m)$ is the joint pdf of all m progressively Type-II censored order statistics.

The KL information can be estimated by

$$I_{1\cdots m:m:n}(f:f^0) = -n\overline{H}_{1\cdots m:m:n} - \sum_{i=1}^{m}\log f^0(x_i;\theta) - \sum_{i=1}^{m} R_i \log\left(1 - F^0(x_i;\theta)\right). \tag{2}$$

Thus, the test statistic based on $\frac{1}{n}I_{1\cdots m:m:n}(f:f^0)$ is given by

$$T(w,n,m) = -H(w,n,m) - \frac{1}{n}\left[\sum_{i=1}^{m}\log f^0(x_i;\widehat{\theta}) + \sum_{i=1}^{m} R_i \log(1 - F^0(x_i;\widehat{\theta}))\right], \tag{3}$$

where $\widehat{\theta}$ is an estimation of θ.

3.2 Test Statistic

Suppose we are interested in goodness-of-fit test for
$H_0 : f^0 = (2\pi\sigma^2)^{\frac{-1}{2}} exp\{-(x-\mu)^2/2\sigma^2\}$ vs $H_A : f^0 \neq (2\pi\sigma^2)^{\frac{-1}{2}} exp\{-(x-\mu)^2/2\sigma^2\}$ where $\underline{\theta} = (\mu, \sigma^2)$ is unknown.
Then, the KL information for a progressively Type-II censored data can be approximated, by (3) and we estimate the unknown parameters (μ, σ^2) by the maximum likelihood estimate.
The Maximum Likelihood Estimation (MLE) for progressively Type-II censored sample from a $Normal(\mu, \sigma^2)$ distribution obtain by solving the below equations, (Balakrishnan, 2000)

$$\frac{\sum_{i=1}^{m} x_i}{m} = \overline{x} = \mu - \frac{\sigma}{m}\sum_{i=1}^{m} R_i Z_i$$

$$\frac{\sum_{i=1}^{m}(x_i - \overline{x})^2}{m} = s^2 = \sigma^2\{1 - \frac{1}{m}\sum_{i=1}^{m} R_i \xi_i Z_i - (\frac{1}{m})^2 \sum_{i=1}^{m}(R_i Zi)^2\},$$

where $Z_i = \frac{\varphi(\xi_i)}{1-\phi(\xi_i)}$ and $\varphi(.)$ is the probability density function of the standard normal distribution.

3.3 Implementation of Test

Because the sampling distribution of $T(w,n,m)$ is intractable, we determine the percentage points using 10,000 Monte Carlo simulations from an exponential distribution. In determining the window size w which depends on n, m and α, we define the optimal window size w to be one which gives minimum critical points. However, we find from the simulated percentage points that the optimal window size w varies much according to m rather than n, and does not vary much according to α, if $\alpha \leq 0.1$. In view of these observations, our recommended values of w for different m are as given in Ebrahimi (1992) and Park (2005).

To obtain the critical values, after deciding about the value of w, simulate the whole procedure by taking the observation from $Normal(0, 1)$ distribution and calculate the value of $T(w,n,m)$, for about 10,000 times. Critical values can then be the percentage points of the thus derived (empirical) distribution of T.

3.4 Power Results

As the proposed test statistic is related to the hazard function of the distribution, we consider the alternatives according to the type of hazard function as follows:
a) Monotone increasing hazard: Weibull (shape parameter 2),
b) Monotone decreasing hazard: Weibull (shape parameter 0.5),
c) Nonmonotone hazard: Center Beta (shape parameter 0.5).
We used 10,000 Monte Carlo simulations for $n = 20$, to estimate the power of our proposed test statistic. The simulation results are summarized in Tables 1.

We can see from Tables 1 that the former censoring scheme shows better power than the other schemes when the alternative is monotone increasing hazard function or monotone decreasing hazard functions. For the alternative with nonmonotone hazard function, sometimes the former censoring scheme gives higher power and sometimes the latter censoring scheme does.

Table 1: Power for different hazard alternatives at 10% significance level for several progressively censored samples when the sample size is $n = 20$.

		monotone increasing hazard alternatives	monotone decreasing hazard alternatives	nonmonotone hazard alternatives
m	schemes $(R_1, \cdots, R_m)$	Weibull shape 2	Weibull shape 0.5	Center Beta shape 0.5
5	15,0,0,0,0	.323	.336	.341
5	0,15,0,0,0	.340	.330	.125
5	3,3,3,3,3	.420	.345	.235
5	0,0,0,15,0	.245	.101	.098
5	0,0,0,0,15	.103	.104	.205
10	10,0,0,...,0,0,0	.452	.446	.344
10	0,10,0,...,0,0,0	.460	.456	.241
10	1,1,1,...,1,1,1	.542	.480	.338
10	0,0,0,...,0,10,0	.321	.303	.215
10	0,0,0,...,0,0,10	.250	.289	.296
15	5,0,0,...,0,0,0	.712	.708	.401
15	0,5,0,...,0,0,0	.785	.766	.394
15	1,1,...,1,...,1,1	.835	.821	.459
15	0,0,0,...,0,5,0	.450	.310	.257
15	0,0,0,...,0,0,5	.445	.308	.409
18	2,0,0,...,0,0,0	.862	.845	.485
18	0,2,0,...,0,0,0	.880	.860	.452
18	1,0,0,...,0,0,1	.916	.926	.495
18	0,0,0,...,0,2,0	.654	.402	.501
18	0,0,0,...,0,0,2	.540	.423	.435

REFERENCES

[1] I. Arizono and H. Ohta, "A test for normality based on Kullback-Leibler information," *The American Statistician,* vol. 43, pp. 20–23, 1989.

[2] N. Balakrishnan and R. Aggarwala, *"Progressive Censoring: Theory, Methods, and Applications,"* Birkhauser, Boston, 2000.

[3] N. Balakrishnan, H. K. T. Ng and N. Kannan, "Goodness-of-fit tests based on spacings for progressively type-II censored data from a general location-scale distribution," *IEEE Trans. on Rel.,* vol. 53, pp. 349–356, 2004.

[4] N. Balakrishnan, A. Habibi Rad and N. R. Arghami, "Testing Normality Based on Kullback-Leibler Information With Progressively Type-II Censored Data," *IEEE Trans. on Rel.,* vol. 53, pp. 349–356, 2007.

[5] U. Balasooriya and N. Balakrishnan, "Reliability sampling plans for lognormal distribution, based on progressively censored samples," *IEEE Trans. on Rel.,* vol. 49, pp. 199–203, 2000.

[6] A. W. Bowman, "Density based tests for goodness-of-fit," *J. Statistical Computation and Simulation,* vol. 40, pp. 1–13, 1992.

[7] E. J. Dudewicz and E. C. van der Meulen, "Entropy-based tests of uniformity," *J. Amer. Statist. Assoc.,* vol. 76, pp. 967–974, 1981.

[8] N. Ebrahimi and M. Habibullah, "Testing exponentiality based on Kullback-Leibler information," *J. Royal Statist. Soc. B,* vol. 54, pp. 739–748, 1992.

[9] A. Foldes, L. Rejto and B. B. Windter, "Strong consistency of nonparametric estimators for randomly censored data," In *Handbook of Statistics*: North Holland, 1980.

[10] D. V. Gokhale, "On entropy-based goodness-of-fit tests," *Computational Statistics and Data Analysis,* vol. 1, pp. 157–165, 1983.

[11] S. Kullback, "*Information Theory and Statistics,"* New York: Wiley, 1959.

[12] H. K. T. Ng, P. S. Chan and N. Balakrishnan, "Estimation of parameters from progressively censored data using EM algorithm," *Computational Statistics and Data Analysis,* vol. 39, pp. 371–386, 2002.

[13] S. Park, "The entropy of consecutive order statistics," *IEEE Trans. Inform. Theory,* vol. 41, pp. 2003–2007, 1995.

[14] S. Park, "Testing exponentiality based on the Kullback-Leibler information with the type II censored data," *IEEE Trans. on Rel.,* vol. 54, pp. 22–26, 2005.

[15] C. E. Shannon, "A mathematical theory of communications," *Bell System Tech. J.,* vol. 27, pp. 379–423, 1948.

[16] S. Teitler, A. K. Rajagopal and K. L. Ngai, "Maximum entropy and reliability distributions," *IEEE Trans. on Rel.,* vol. 35, pp. 391–395, 1986.

[17] O. Vasicek, "A test for normality based on sample entropy," *J. Royal Statist. Soc. B,* vol. 38, pp. 730–737, 1976.

[18] K. M. Wong and S. Chan, "The entropy of ordered sequences and order statistics," *IEEE Trans. Inform. Theory,* vol. 36, pp. 276–284, 1990.