COMPUTATION OF PERCENTILES POINTS FOR KOLMOGROV-SMIRNOV STATISTIC UNDER TYPE-II CENSORING

YASH SETHI

Abstract. This work considers goodness-of-fit test for the life test data with a few censoring schemes. Kolmogorov–Smirnov(KS) test is one of the most popular non-parametric test for goodness-of-fit with complete data. KS test is easily extendable for the Type-I censoring. But its generalization for the Type-II censoring is much more challenging because of the involvement of censoring sample size and the complete sample size. KS statistic is a continuous functional of the standard Brownian Bridge which is an infinite dimensional object. As a consequence sampling from the standard Brownian Bridge demands finer partitions on $[0, 1]$ and it becomes computationally intensive. But the finer partition neither guaranty to meet the level of the test nor to maintain the power of the test under Type-II censoring. **We want to propose a computationally efficient method to find the percentile points of the distribution of KS statistic under Type-II censoring scheme overcoming the above problems. We are intending to generalize the idea for the case when the parameter is unknown.**

Contents

[0]The project is done under the supervision of Dr. Buddhananda Banerjee.

1. Introduction

Testing for goodness-of-fit is one of the fundamental problem in statistics. A density based approach provides χ^2-test and a distribution based approach provides Kolmogorov–Smirnov (KS) test for the same. Among many other, the most popular and worth performing goodness-of-fit tests with complete data are Cramér-von Mises test and Anderson-Darling test etc.. When the data are complete and the distribution is pre-specified, the KS-test statistic asymptotically follows the Kolmogorov distribution under the null hypothesis. The Kolmogorov distribution can be viewed as a distribution of the supremum norm of a standard Brownian bridge on $[0, 1]$.

When the data are censored with different schemes, the generalization of the KS-test becomes a more interesting challenge with life testing data. The two most important censoring schemes are Type-I and Type-II censoring. The duration of a life-testing experiment, denoted by random variable X, in a Type-I censoring is predetermined, say X_0, goes in favour of consumer. But the number of failures in that time interval $[0, X_0]$ is a non-negative integer-valued random variable. On the contrary, in a Type-II censoring scheme, the experiment takes a random time to produce a pre-specified number of failures, say r, goes in favour of the producer. Hence, the stopping time of the experiment is a random variable, the r^{th} order statistic, usually denoted by $X_{(r)}$.

Kolmogorov (1933) and Smirnov (1948) provided a complete methodology to compare an empirical cumulative distribution function (ECDF) with a pre-specified cumulative distribution function (CDF). The two-sided one-sample KS test is modified by Barr and Davidson (1973) to use it for the censored and truncated samples. According to their observation, the goodness-of-fit tests based on the modified test statistic are inappropriate when parameters of the hypothesized distribution are estimated from the data and used for the test. It reduces the power of the test. Some correction factors are also suggested by Dufour and Maag (1978) to the KS statistic obtained from Type-I and Type-II censoring schemes to make the statistic compatible with the tabulated critical values provided by Koziol and Byar (1975). A generalization of KS test was done by Fleming *et al.* (1980) for the one-sample and the two-sample

problems of an arbitrarily right-censored data. A new modified goodness-of-fit testing based on Type-II right censored data was proposed by Lin *et al.* (2008). The goodness-of-fit test for censored data from a location-scale distribution, especially for exponential distribution, has been discussed by Castro-Kuriss (2011).

Zhang (2002) proposed even more powerful test than the existing ones. For Type-I censored data, the KS test statistic has the asymptotic distribution similar to that of complete data on a suitable interval contained in $[0, 1]$. For Type-II censored data, when the sample size (n) and the number of failures (r) be quite large such that the ratio r/n approaches to a constant, the distribution of KS test statistic has a similar behavior to that of the Type-I censored data. If the hypothesized distribution is completely known to be $F_0(\cdot)$ then we also know that the stopping time follows the $Beta(r, n-r+1)distribution$ after a transformation $T = F_0(X)$ following $U[0, 1]$.

We have studied the asymptotic behavior of the KS test statistic when the data are coming from the Type-II censoring schemes. We have compute the percentiles of the distribution of the KS-statistic under Type-II censoring. We have also computed the different functional of the standard Brownian Bridge on $[0, 1]$ to resemble with the exact simulation of the KS-test statistic. We have come up with an efficient computational scheme and we are willing to extend this project for parameter unknown case.

2. Definitions and Assumptions

Suppose $X_1, X_2, \ldots, X_n$ be independently and identically distributed (i.i.d.) random variables form a continuous distribution $F(\cdot)$. Then the KS-statistic quantifies the supremum norm between the ECDF $F_n(x)$ from the sample and the true CDF $F(x)$ on $\mathbb{R}$ defined as

$$D_n = \sup_x |F_n(x) - F(x)| \tag{1}$$

where,

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(-\infty, x]}(X_i) \tag{2}$$

is the ECDF and $\mathbf{1}_{(-\infty,x]}(X_i)$ is the Dirac delta function that equals to 1 if $X_i \leq x$ and equals to 0 otherwise. Then it can be shown that (see Durrett, 2019, Ch 8)

$$\lim_{n\to\infty} \sqrt{n} D_n \xrightarrow{d} \sup_t |B_0(F(t))|. \tag{3}$$

If $F(\cdot)$ is $U[0,1]$ then under the null hypothesis $\sqrt{n}D_n$ converges in distribution to the Kolmogorov distribution. A random variable K defined as

$$K = \sup_{t\in[0,1]} |B_0(t)| \tag{4}$$

is said to have the Kolmogorov distribution, where $B_0(t)$ is the standard Brownian bridge on $[0,1]$. If $Z(t)$ is a standard brownian motion,then

$$B(t) = Z(t) - \frac{t}{T} Z(T) \tag{5}$$

is a Brownian bridge for $t \in [0,T]$. Note that $B(t)$ is independent of $Z(T)$. Brownian motion is a stochastic process $\mathcal{Z} = \{Z(t) : t \geq 0\}$ is satisfying the following conditions:

- For $t_1 < t_2 < \ldots < t_n$, the random variables $Z(t_n) - Z(t_{n-1}), ..., Z(t_1) - Z(0)$ are independent
- $Z(t+s) - Z(s) \overset{d}{=} Z(t) - Z(0)$ for $s, t \geq 0$
- $Z(t) \overset{d}{=} N(\mu t, \sigma^2 t)$ for $t \geq 0$, drift $\mu \in \mathbb{R}$ and scale $\sigma > 0$
- $Z(t)$ has a continuous path.

A Brownian motion is known as a standard brownian motion if it has mean 0 and variance t i.e. $Z(t) \sim N(0,t)$ for all $t \geq 0$. Data or observation are said to be censored if only a partial information of the data is sampled under the scheme. Types of censoring considered in this work are

- **Type-I censoring** : The duration of a life-testing experiment in a Type-I censoring is predetermined, say $X_0 \in \mathbb{R}^+$. The number of events in that time interval is a non negative integer-valued random variable.
- **Type-II censoring** : In a Type-II censoring scheme, the experiment takes a random time to produce the required number of events, say $r \in \mathbb{N}$, which is prespecified. The stopping time is a random variable, the r^{th} order statistic, denoted by $X_{(r)}$.

Suppose X be a continuous random variable stands for the life distribution supported on positive part of real line i.e. $\mathbb{R}^+ = \{x|x > 0\}$. It is assumed that under the null hypothesis $H_0 : X \sim F_0(\cdot)$, where $F_0(\cdot)$ is completely specified and the alternative hypothesis H_1 claims that H_0 is not true. We can easily do the transformation $T = F_0(X)$ which always follows $U[0,1]$ under the null hypothesis. Now onward we will discuss about the random variable with $U[0,1]$ distribution only. As a consequence $T_0 = F_0(X_0) \in (0,1)$ is the stopping time for Type-I censoring and $T_{(r)} = F_0(X_{(r)}) \sim Beta(r, n-r+1)$ for Type-II censoring.

3. Methodology

In case of complete data, $T_1, T_2, \cdots T_n$ are i.i.d. $U[0,1]$ random variables under H_0. The CDF is given by

$$F(t) = \begin{cases} 0, & \text{if } t < 0 \\ t, & \text{if } t \in [0,1] \\ 1, & \text{if } t > 1 \end{cases} \tag{6}$$

and the ECDF is given by

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} I_{\{T_i \leq t\}} \tag{7}$$

Hence the KS-statistic for complete data is defined as

$$D_n(t) := F_n(t) - t \tag{8}$$

So, limiting distribution of KS- statistic with complete data denoted by KS^0 as $n \to \infty$ converges to K in distribution i.e.

$$KS^0 = \sup_{t\in[0,1]} \sqrt{n}|D_n(t)| = \sup_{t\in[0,1]} \sqrt{n}|F_n(t) - t| \xrightarrow{d} \sup_{t\in[0,1]} |B(t)| = K \tag{9}$$

The working formula to compute from complete data

$$KST^0 = \max_i \max \left\{ |T_{(i)} - \frac{i}{n}|, |T_{(i)} - \frac{i-1}{n}| \right\} \tag{10}$$

where, $\{T_{(1)}, T_{(2)}, \cdots T_{(n)}\}$ are the order statistic of $T_1, T_2, \cdots T_n$.

In Type-I censoring scheme, the experiment is terminated at time $T_0 \in [0,1]$. So, the modified KS-statistic, KS^I say,for Type-I censoring will be :

$$KS^I = \sup_{t\in[0,T_0]} \sqrt{n}|D_n(t)| = \sup_{t\in[0,T_0]} \sqrt{n}|F_n(t) - t| \tag{11}$$

Here the distribution of the test statistic is immediate because the stopping time T_0 is independent of the process $B_0(t)$. So the limiting distribution of KS^I when $n \to \infty$ can be obtained as

$$KS^I = \sup_{t\in[0,T_0]} \sqrt{n}|D_n(t)| \xrightarrow{d} \sup_{t\in[0,T_0]} |B(t)| = K_1 \tag{12}$$

Dufour and Maag (1978) suggested the working formula for Type-I censored data

$$KST^I = \max_{i\leq d} \left\{ |T_{(i)} - \frac{i}{n}|, |T_{(i)} - \frac{i-1}{n}|, |T_{(d)} - \frac{d}{n}| \right\} \tag{13}$$

where, $\{T_{(1)} < T_{(2)} < \cdots < T_{(d)} < T_0 < T_{(d+1)}\}$

In Type-II censoring scheme the experiment is stopped when the r^{th} failure, i.e., $T_{(r)} = F_0(X_{(r)})$ takes place. We observe the realizations till the rth failure as $T_{(1)}, T_{(2)}, ..., T_{(r)}$. This test can be formulated using three different methods. First of all the working formula suggested by Dufour and Maag (1978) for Type-II censored data is

$$\sqrt{n}D_{n:r} = \sqrt{n} \max_{i\leq r} \left\{ \left|T_{(i)} - \frac{i}{n}\right|, \left|T_{(i)} - \frac{i-1}{n}\right| \right\} = KST^{II} \tag{14}$$

In the first method we follow the idea by Koziol and Byar (1975). They calculated the percentage points for different truncation points of Type-I censoring and claimed that it will work equivalently well for Type-II censoring when $r/n \longrightarrow \lambda_0$, some fixed truncation point as used in Type-I censoring. Let us assume that both r and n move to ∞ the mode of convergence can be stated as

$$\sqrt{n}D_{n:r} \xrightarrow{d} \sup_{t\in[0,\lambda_0]} |B(t)| = K_{2a}. \tag{15}$$

Note that the correction factor $(-0.24/\sqrt{n})$ suggested by Koziol and Byar (1975) to accommodate the sample size is only a function of n but not r which is also crucial for the Type-II censoring.

In the second method we consider the known fact that the stopping time $T_{(r)}$ follows a $Beta(r, n-r+1)$ distribution. So, in this method the mode of convergence can be states as

$$\sqrt{n}D_{n:r} \xrightarrow{d} \sup_{t\in[0,T_{(r)}]} |B(t)| = K_{2b}. \tag{16}$$

We consider the third method following the work by Banerjee and Pradhan (2018). Assuming $u \in [0, 1)$ define

$$D_n^{II}(u) = D_n(uT_{(r)}) = F_n(uT_{(r)}) - uT_{(r)}. \tag{17}$$

$F_n(uT_{(r)})$ and $uT_{(r)}$ are independent and $nF_n(uT_{(r)}) \sim bin(r-1, u)$ which is invariant of $T_{(r)}$. For the Type-II censoring the KS-statistic define as

$$KS^{II} = \sup_{u\in[0,1]} \sqrt{n}|D_n^{II}(u)| = K_{2c} \tag{18}$$

which can be computed as

$$\max \sqrt{n} \left\{ \max_{u\in[0,1)} \left| \frac{r-1}{\sqrt{n}}u + \sqrt{\frac{r-1}{n}}B_0(u) - u\sqrt{n}T_{(r)} \right|, \left| \frac{r}{n} - T_{(r)} \right| \right\} \tag{19}$$

4. Simulations and Findings

We have conducted an extensive simulation to study the performances of these test statistics and the closeness of their behaviour with the limiting distribution. For testing purpose mostly the 99th, 95th and 90th percentiles are used. The methodology we have discussed can be used for any computation of percentile. But for the ease of demonstration we consider the 95th percentile point. For sample size $200(= n)$ data are generated from $U[0, 1]$ distributions and the test statistic values are computed for KST^0, KST^I and KST^{II} following the equations (10), (13) and (14) respectively. On the other hand, for the limiting test statistic values K, K_1, K_{2a}, K_{2b} and K_{2c} are generated with different grid sizes for the standard Brownian bridge on $[0, 1]$. We have chosen grid sizes $5n$, $12n$, $24n$, $25n$, $50n$ and $100n$ in separate cases. In each of the cases the data are generated from the exact test statistic and the limiting test statistic for **10000** times. It is providing the an opportunity to compare the closeness of the distribution of KST^0 and K for complete data; the distribution of KST^I and K_1 for the Type-I censoring; the distribution of KST^{II} and K_{2a}, K_{2b}, K_{2c} for the Type-II censoring. We have compared them with respect to the Wilcoxon Rank Sum (WRS) Test. This entire

Censoring Scheme	**Exact Test statistic**	**Limiting Test statistic**	**Grid size** $n \times 100$	**Grid size** $n \times 12$
	KST^0	$\mathbf{K}$	0.5576	0.0492
Type-I	KST^I	$\mathbf{K_1}$	0.5968	0.0596
Type-II	KST^{II}	$\mathbf{K_{2a}}$	0.7248	0.0616
Type-II	KST^{II}	$\mathbf{K_{2b}}$	0.4484	0.0988
Type-II	KST^{II}	$\mathbf{K_{2c}}$	0.5612	0.1156

TABLE 1. Proportion of p-values less than 0.05.

process is repeated **2500** time and the p-values are speculated. We know that under the null hypothesis the distribution of p-value will follow $U[0,1]$. As a consequence the computed p-value less than the level 0.05 will be approximately close to 0.05. Only that feature has been observed when the grid size is chosen to be $12n$ and comparisons are done between the distributions of KST^0 and K, KST^1 and K_1, and KST^{II} and K_{2a}. The closeness of K_{2b}, K_{2c} with KST^{II} are not that significant. On the contrary the other specifications of grid sizes shows even more mismatch with the exact distribution of the test statistic. For example we have reported the result for $100n$ grid size in Table 1.

From the histograms we also can see that with grid size $12n$ simulated data have more resemblance with the exact simulation. As a consequence we observe uniform distribution of the p-values in complete data and Type-I censoring. Figure 1 is showing the same.

For Type-II censoring when the exact simulation is compared with K_{2a} is also providing closed to $U[0,1]$ distribution for the p-value. Figure 2 is showing the same. Other two statistics K_{2b} and K_{2c} are lagging a bit. But for grid size $100n$ we observe a drastic mismatch among the exact and the limiting distribution. We have observed the same for the other grid sizes also. It is reflected in Figure 3.

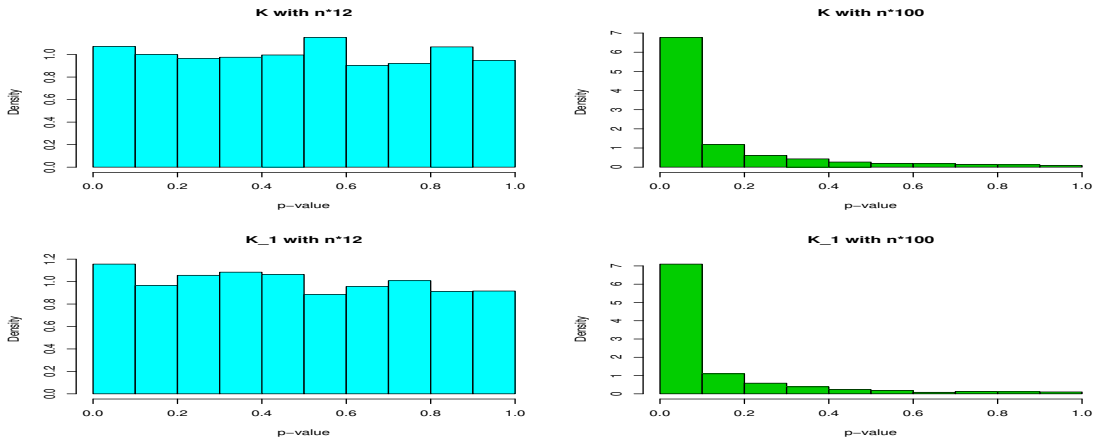

FIGURE 1. K in Complete data and K_1 Type-I censoring

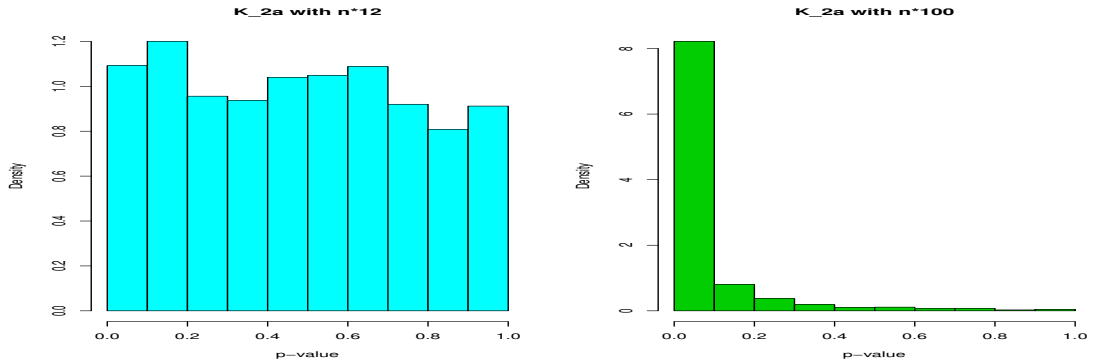

FIGURE 2. K_{2a} in Type-II censoring

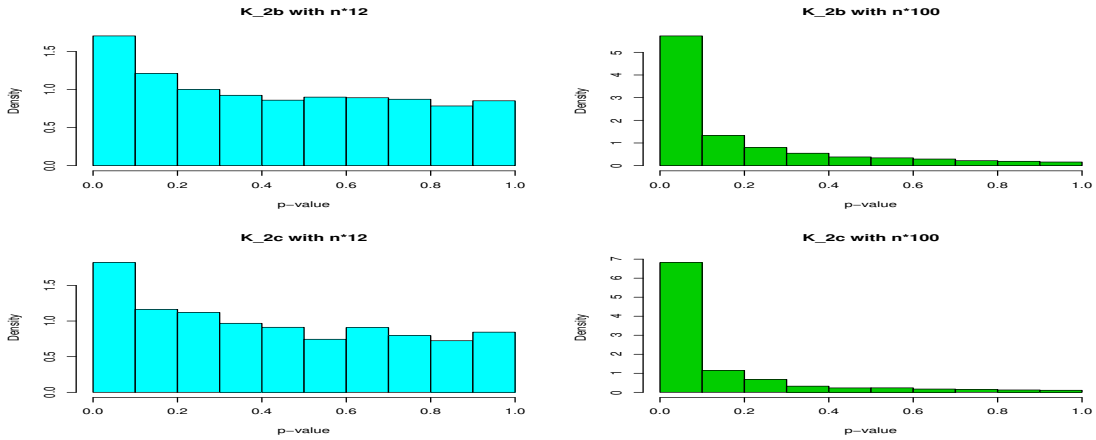

FIGURE 3. K_{2b} and K_{2c} in Type-II censoring

5. Conclusion and Future work

In this analysis we have compared the closeness of the exact distribution of KS-statistic with the limiting distribution of that for complete data, Type-I censoring scheme and Type-II censoring scheme. We have considered different grid sizes on $[0, 1]$ to generate data from Brownian bridges. We observed that not only for the Type-II censoring scheme but also for complete data and Type-I censoring scheme grid size as $\mathbf{n \times 12}$ is working exceptionally well. This is consuming less computational cost and time too. We would like to theorize the finding has a connection with $U[0, 1]$ distribution because it has the variance $\frac{\mathbf{1}}{\mathbf{12}}$. We will provide proof of the same in future. Also, we are aiming to generalize this idea for the case when the parameter is unknown.

6. Acknowledgement

I am heartily thankful to my supervisor, Dr.Buddhananda Banerjee, whose encouragement, guidance and support from the initial to final level enabled me to develop an understanding of this topic. Also, I would like to thank the Department of Mathematics, IIT Kharagpur, which provided me all facilities required to finish my project smoothly.

References

Banerjee, B. and Pradhan, B. (2018). Kolmogorov–smirnov test for life test data with hybrid censoring. *Communications in Statistics-Theory and Methods*, **47**(11), 2590–2604. 7

Barr, D. R. and Davidson, T. (1973). A kolmogorov-smirnov test for censored samples. *Technometrics*, **15**(4), 739–757. 2

Castro-Kuriss, C. (2011). On a goodness-of-fit test for censored data from a location-scale distribution with applications. *Chilean Journal of Statistics*, **2**(1), 115–136. 3

Dufour, R. and Maag, U. (1978). Distribution results for modified kolmogorov-smirnov statistics for truncated or censored. *Technometrics*, **20**(1), 29–32. 2, 6

Durrett, R. (2019). *Probability: theory and examples*, volume 49. Cambridge university press. 4

Fleming, T. R., O'Fallon, J. R., O'Brien, P. C., and Harrington, D. P. (1980). Modified kolmogorov-smirnov test procedures with application to arbitrarily right-censored data. *Biometrics*, pages 607–625. 2

Kolmogorov, A. (1933). Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, **4**, 83–91. 2

Koziol, J. A. and Byar, D. P. (1975). Percentage points of the asymptotic distributions of one and two sample ks statistics for truncated or censored data. *Technometrics*, **17**(4), 507–510. 2, 6

Lin, C.-T., Huang, Y.-L., and Balakrishnan, N. (2008). A new method for goodness-of-fit testing based on type-ii right censored samples. *IEEE Transactions on Reliability*, **57**(4), 633–642. 3

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, **19**(2), 279–281. 2

Zhang, J. (2002). Powerful goodness-of-fit tests based on the likelihood ratio. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(2), 281–294. 3

Department of Mathematics Indian Institute of Technology Kharagpur, India-721302

Email address: sethiyash87@gmail.com