Computation of percentile points for Kolmogorov-Smirnov statistic under Type-II censoring

Yash Sethi
18MA40019

Department of Mathematics
IIT Kharagpur

27 May 2020

Outline

- Goodness of fit
- Kolmogorov-Smirnov test
- Kolmogorov Distribution
- Kolmogorov test for censored data
  - Type I censoring
  - Type II censoring
- Different Methods for Type II censoring
- Parameters Unknown
- Analysis and findings

Goodness-of-fit test

- $X_1, X_2, \cdots, X_n$ are i.i.d. random samples from a continuous distribution.
- $H_0: \; X \sim F_0(.)$ Vs $H_1: \; X$ does not follow $F_0(.)$
- Assumption: Hypothesized distribution $F_0(.)$ is completely known.
- Consider the transformation $T = F_0(X)$ and so $T_i = F_0(X_i)$ $\forall i = 1, 2 \cdots, n$
- Equivalent to test $H_0 : T \sim U[0,1]$ Vs $H_1 : T$ does not follow $U[0,1]$

Now onwards we will consider random samples from $U[0,1]$, unless other wise specified.

- The Kolmogorov–Smirnov statistic quantifies a distance between the empirical cumulative distribution function(ecdf) of the sample and the cdf of the reference distribution, or between the ecdfs of two samples.

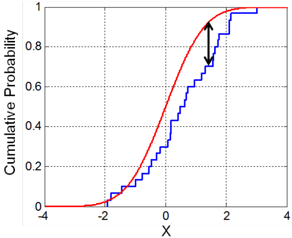

Fig. Illustration of the Kolmogorov–Smirnov statistic.
Red line is CDF, blue line is an ECDF, and **the black arrow** is the KS-statistic.

The KS-statistic for a specified cumulative distribution function F(x) is given by :

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

where

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

is Empirical distribution and $\mathbf{1}_{(-\infty,x]}(X_i)$ is the Dirac delta function,

equal to 1 if $X_i \leq x$ and equal to 0 otherwise.

Kolmogorov Distribution

Definition

A random variable K defined as

$$K = \sup_{t \in [0,1]} |B(t)| \tag{2}$$

is said to have the Kolmogorov distribution, where B(t) is the standard Brownian bridge on [0,1].

Under null hypothesis that the sample comes from the hypothesized distribution $F(x)$,

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

in distribution. If F is continuous then under the null hypothesis $\sqrt{n} D_n$ converges to the Kolmogorov distribution, which does not depend on F.

Data or observation is said to be censored if only a partial information of the data is sample under the scheme.
Types of censoring considered in our work :

1. **Type I censoring** : The duration of a life-testing experiment in a Type-I censoring is predetermined, say X_0, goes in favour of consumer.
   The number of events in that time interval is a non-negative integer-valued random quantity.

2. **Type II censoring** : In a Type-II censoring scheme, the experiment takes a random time to produce the required number of events, say r, which is prespecified, goes in favour of producer.
   The stopping time is a random variable, the rth order statistic, denoted by $X_{(r)}$.

- $T_1, T_2, \cdots T_n$ are i.i.d. $U[0,1]$ under H_0.
- Define $D_n(t) := (F_n(t) - t)$
- Kolmogorov-Smirnov statistic for complete data

$$KS^0 = \sup_{t\in[0,1]} \sqrt{n}|D_n(t)| = \sup_{t\in[0,1]} \sqrt{n}|F_n(t) - t|$$

- Limiting distribution of KS^0 can be obtained as

$$\sup_{t\in[0,1]} \sqrt{n}|D_n(t)| \xrightarrow{L} \sup_{t\in[0,1]} |B(t)| = K \quad \text{when} \quad n \uparrow \infty. \tag{4}$$

- Working formula:

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

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

KS Test for Type I censoring Scheme [1]

- The experiment is terminated at time $T_0 = F_0(X_0) \in [0, 1]$.
- Kolmogorov-Smirnov statistic for Type I censoring will be,

$$KS^I = \sup_{t \in [0,T_0]} \sqrt{n}|D_n(t)| = \sup_{t \in [0,T_0]} \sqrt{n}|F_n(t) - t| \tag{5}$$

- Limiting distribution of KS^I can be obtained as

$$\sup_{t \in [0,T_0]} \sqrt{n}|D_n(t)| \xrightarrow{L} \sup_{t \in [0,T_0]} |B(t)| = K_1 \text{ when } n \uparrow \infty. \tag{6}$$

- Working formula for Type-I censored data:
$KST^I = \max_{i \le d} \left\{ |T_{(i)} - \frac{i}{n}|, |T_{(i)} - \frac{i-1}{n}|, |T_{(d)} - \frac{d}{n}| \right\}$
where, $\{T_{(1)} < T_{(2)} < \cdots < T_{(d)} < T_0 < T_{(d+1)}\}$

[1] by Dufour and Maag(1975).

- The experiment is stopped when the r^{th} failure, i.e., $T_{(r)} = F_0(X_{(r)})$ takes place.
- Observed data: $T_{(1)}, T_{(2)}, \cdots, T_{(r)}$
- $T_{(r)} \sim \text{Beta}(r, n-r+1)$
- Consider $u \in [0,1)$ and define

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

$$D_n^{II}(u) = \frac{r-1}{\sqrt{n}}u + \sqrt{\frac{r-1}{n}}B(u) - u\sqrt{n}\,\mathrm{B}(r, n-r+1) + \frac{1_{\{u=1\}}}{\sqrt{n}}. \quad (7)$$

Note 1: $F_n(uT_{(r)})$ and $uT_{(r)}$ are independent

Note 2: $nF_n(uT_{(r)}) \sim \;\; bin(r-1, u)$ which is invariant of $T_{(r)}$

[2] by Dufour and Maag(1975) & Banerjee and Pradhan(2018).

Working formula for Type II censoring is given by:

$$\sqrt{n}D_{n:r} = \sqrt{n}\max_{i\le r}\left\{\left|T_{(i)} - \frac{i}{n}\right|, \left|T_{(i)} - \frac{i-1}{n}\right|\right\} = KST^{II} \tag{8}$$

1. **Method 1:** When $n, r \uparrow \infty$ such that $r/n \to \lambda_0$, then approximately

$$\sqrt{n}D_{n:r} \xrightarrow{d} \sup_{t\in[0,\lambda_0]} |B(t)| = K_{2a}. \tag{9}$$

1. **Method 2:** $T_{(r)} \sim Beta(r, n-r+1)$, then approximately

$$\sqrt{n} D_{n:r} \xrightarrow{d} \sup_{t \in [0, T_{(r)}]} |B(t)| = K_{2b}. \tag{10}$$

2. **Method 3:** $F_n(uT_{(r)})$ and $uT_{(r)}$ are independent

   $nF_n(uT_{(r)}) \sim \ bin(r-1, u)$ which is invariant of $T_{(r)}$. Then approximately

$$\sqrt{n} D_{n:r} \xrightarrow{d} \sup_{u \in [0,1]} \sqrt{n} |D_n^{II}(u)| = K_{2c} \tag{11}$$

Simulation

- Studied the closeness between test statistics and limiting distribution

- Sample size, n = **200**
  Censoring time, $T_0 = \mathbf{0.4}$
  Censoring Size, r = **80**

- Test statistic KST^0, KST^I and KST^{II} & Limiting distribution is calculated **10000** times

- Limiting test statistic values from the distribution of K, K_1, K_{2a}, K_{2b} and K_{2c} are generated with different grid sizes for the standard Brownian bridge on $[0, 1]$.

- Compared them with Wilcoxon Rank Sum test

- Whole process was repeated **2500** times and p-values were recorded

Censoring Scheme	Exact Test statistic	Limiting Test statistic	Grid size $n \times 100$	Grid size $n \times 12$
Complete Data	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: Proportion of p-values less than 0.05.

Findings

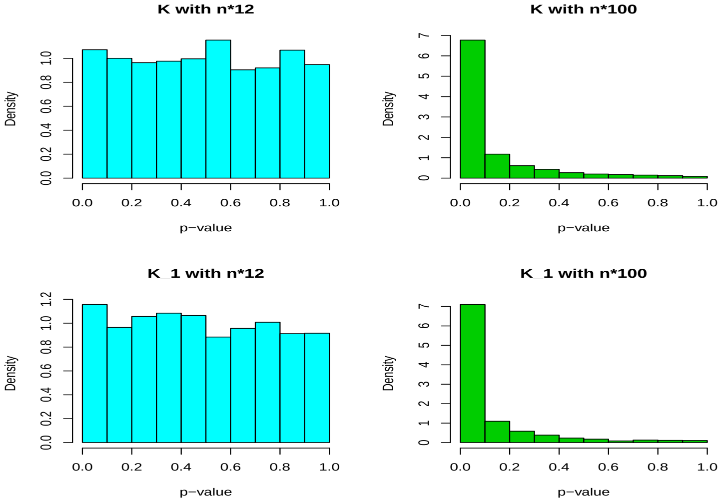

Figure: K with Complete data and K_1 with Type-I censoring

Findings

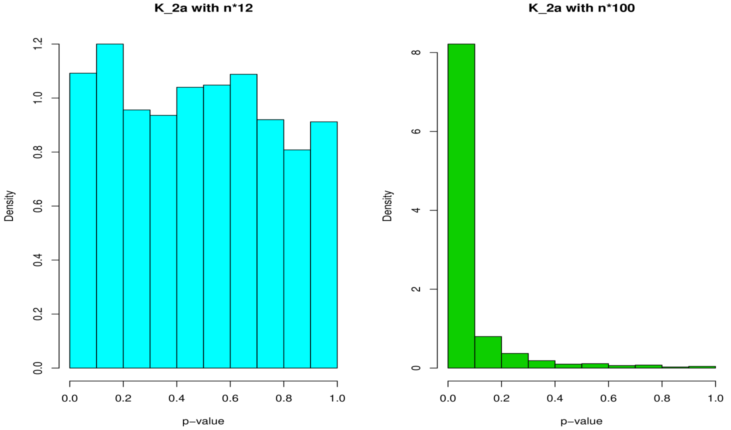

Figure: K_{2a} with Type-II censoring

Findings

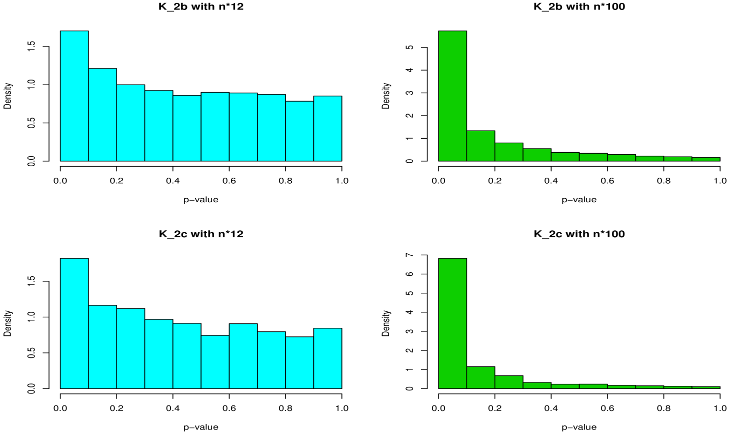

Figure: K_{2b} and K_{2c} with Type-II censoring

A large class of distributions can be covered by location-scale families. All the pdfs in the family are generated by transforming the standard pdf in the prescribed way. Let $f(x)$ be the standard pdf.

- $f(x-\mu)$, indexed by μ, $-\infty < \mu < \infty$, is called the *location family* and μ is called the *location parameter* of the family.
- $\frac{1}{\sigma}f(\frac{x}{\sigma})$, indexed by σ, $\sigma > 0$, is called the *scale family* and σ is called the *scale parameter* of the family.
- $\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)$, indexed by $\sigma(\sigma > 0)$,$\mu(-\infty < \mu < \infty)$, is called the *location-scale family* and μ and σ are called the *location and scale parameter* of the family repectively.

Simulation

- Let $X \sim F(\mu, \sigma)$ and $Z \sim F(\mu = 0, \sigma = 1)$, where F be any distribution with μ and σ as location and scale parameters(both unknown).
- Data is generated with sample size,n= **200**, for both **X** and **Z**.
- Parameters μ and σ are estimated from data as $\hat{\mu}(\mathbf{X})$ and $\hat{\sigma}(\mathbf{X})$ respectively using Method of Moments(MoM) or Maximum likelihood estimator(MLE) whichever is best suited.
- Transformed into X^* and Z^* as $X^* = \frac{T(\mathbf{X}) - \hat{\mu}(\mathbf{X})}{\hat{\sigma}(\mathbf{X})}$ and $Z^* = \frac{T(\mathbf{Z}) - \hat{\mu}(\mathbf{Z})}{\hat{\sigma}(\mathbf{Z})}$, where $T(\cdot)$ is any function of random variable.
- KS-statistic values are formulated for X^* and Z^* **20000** times and are compared with respect to WRS test.
- Whole process was repeated **2500** times and p-values are speculated.

- Suppose $X_1, X_2, \cdots, X_n$ are n *i.i.d.* samples from $U(a, b)$ distribution. Let $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ be the ordered statistic. We have estimated parameters a and b as :

$$\hat{a}(\mathbf{X}) = \frac{n * X_{(1)} - X_{(n)}}{n-1} \tag{12}$$

$$\hat{b}(\mathbf{X}) = \frac{n * X_{(n)} - X_{(1)}}{n-1} \tag{13}$$

- Both $\hat{a}$ and $\hat{b}$ are uniformly minimum-variance unbiased estimators(UMVUE) of a and b respectively.

- If $X \sim U(a, b)$ and $Z \sim U(0, 1)$ then define $X^* = \frac{X - \hat{a}(\mathbf{X})}{\hat{b}(\mathbf{X}) - \hat{a}(\mathbf{X})}$ and $Z^* = \frac{Z - \hat{a}(\mathbf{Z})}{\hat{b}(\mathbf{Z}) - \hat{a}(\mathbf{Z})}$.

- We can see that p-values are following $U[0,1]$ distribution and quantile difference between KS-statistic follows Normal distribution with mean approximating to $0(-1.332225 * 10^{-5})$ and small variance $(6.564052 * 10^{-5})$.
- Hence, $X^* \stackrel{d}{=} Z^*$.

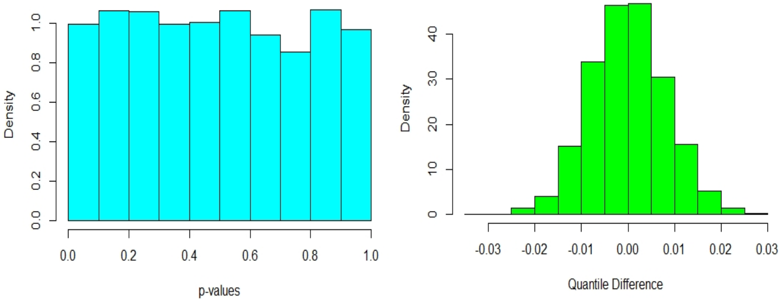

Figure: Uniform Distribution

- Suppose $X_1, X_2, \cdots, X_n$ are n *i.i.d.* samples from $Normal(\mu, \sigma)$ distribution.We have estimated parameters μ and σ using MLE as:

$$\hat{\mu}(\mathbf{X}) = \frac{\sum_{i=1}^{n} X_i}{n} \tag{14}$$

$$\hat{\sigma}(\mathbf{X}) = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \hat{\mu})^2}{n-1}} \tag{15}$$

- Both $\hat{\mu}$ and $\hat{\sigma}$ are unbiased estimators of μ and σ respectively.
- If $X \sim N(\mu, \sigma)$ and $Z \sim N(0, 1)$ then define $X^* = \frac{X - \hat{\mu}(\mathbf{X})}{\hat{\sigma}(\mathbf{X})}$ and $Z^* = \frac{Z - \hat{\mu}(\mathbf{Z})}{\hat{\sigma}(\mathbf{Z})}$.

Normal Distribution, $N(\mu, \sigma)$

- We can see that p-values are following $U[0,1]$ distribution and quantile difference between KS-statistic follows Normal distribution with mean approximating to $0(0.0017058182)$ and small variance$(2.165865 * 10^{-5})$.
- Hence, $X^* \stackrel{d}{=} Z^*$.

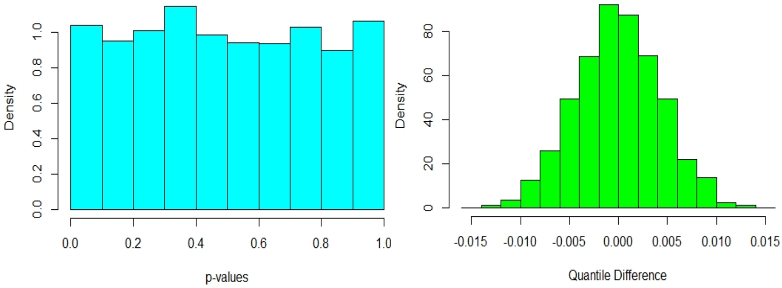

Figure: Normal Distribution

- Suppose $X_1, X_2, \cdots, X_n$ are n *i.i.d.* samples from $Lognormal(\mu, \sigma)$ distribution. As we know that samples $ln(X_1), ln(X_2), \cdots, ln(X_n)$ will follow normal distribution, then we have estimated parameters(unbiased) μ and σ using MLE as:

$$\hat{\mu}(\mathbf{X}) = \frac{\sum_{i=1}^{n} \ln X_i}{n} \tag{16}$$

$$\hat{\sigma}(\mathbf{X}) = \sqrt{\frac{\sum_{i=1}^{n} (\ln X_i - \hat{\mu})^2}{n-1}} \tag{17}$$

- If $X \sim Lognormal(\mu, \sigma)$ and $Z \sim Lognormal(0, 1)$ then define $X^* = \frac{ln(X) - \hat{\mu}(\mathbf{X})}{\hat{\sigma}(\mathbf{X})}$ and $Z^* = \frac{ln(Z) - \hat{\mu}(\mathbf{Z})}{\hat{\sigma}(\mathbf{Z})}$

- We can see that p-values are following $U[0,1]$ distribution and quantile difference between KS-statistic follows Normal distribution with mean approximating to $0(-8.292911 * 10^{-5})$ and small variance$(2.013959 * 10^{-5})$.
- Hence, $X^* \overset{d}{=} Z^*$.

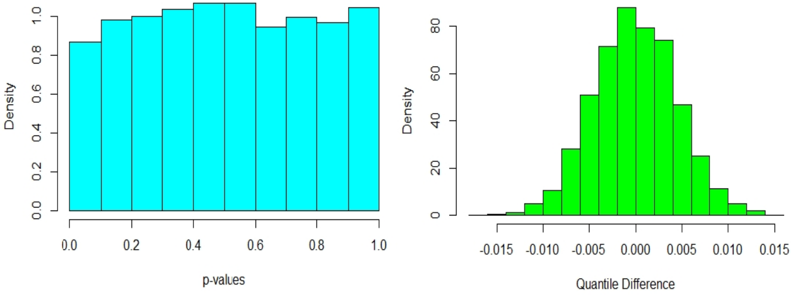

Figure: Lognormal Distribution

- Suppose $X_1, X_2, \cdots, X_n$ are n *i.i.d.* samples from $Logistic(\mu, s)$ distribution.We have estimated parameters μ and σ(unbiased) using MLE. We know that $\sigma^2 = \frac{\pi^2 s^2}{3}$ for logistic distribution. Using this relation, we have estimated the parameter s. The parameters μ and s are estimated as:

$$\hat{\mu}(\mathbf{X}) = \frac{\sum_{i=1}^{n} X_i}{n} \tag{18}$$

$$\hat{\sigma}(\mathbf{X}) = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \hat{\mu})^2}{n-1}} \tag{19}$$

$$\hat{s}(\mathbf{X}) = \sqrt{\frac{3 * \hat{\sigma}^2}{\pi^2}} \tag{20}$$

- If $X \sim Logistic(\mu, s)$ and $Z \sim Logistic(0, 1)$ then define $X^* = \frac{X - \hat{\mu}(\mathbf{X})}{\hat{s}(\mathbf{X})}$ and $Z^* = \frac{Z - \hat{\mu}(\mathbf{Z})}{\hat{s}(\mathbf{Z})}$.

- We can see that p-values are following $U[0,1]$ distribution and quantile difference between KS-statistic follows Normal distribution with mean approximating to $0(-0.000181438)$ and small variance$(2.45671 * 10^{-5})$.
- Hence, $X^* \stackrel{d}{=} Z^*$.

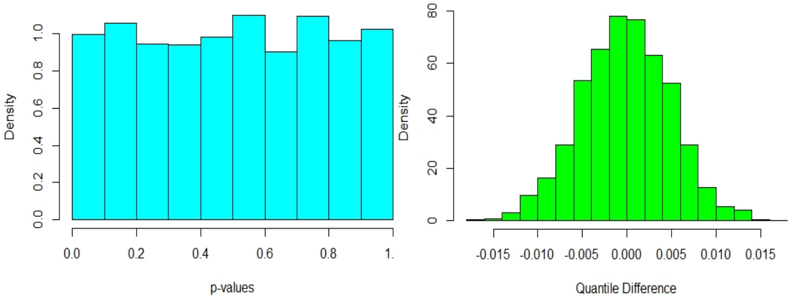

Figure: Logistic Distribution

- Suppose $X_1, X_2, \cdots, X_n$ are n *i.i.d.* samples from $Laplace(\mu, b)$ distribution.We have estimated parameters(unbiased) μ and σ using MLE. We know that $\sigma^2 = 2b^2$ for laplace distribution. Using this relation, we have estimated the parameter b.

$$\hat{\mu}(\mathbf{X}) = \frac{\sum_{i=1}^{n} X_i}{n} \tag{21}$$

$$\hat{\sigma}(\mathbf{X}) = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \hat{\mu})^2}{n-1}} \tag{22}$$

$$\hat{b}(\mathbf{X}) = \sqrt{\frac{\hat{\sigma}^2}{2}} \tag{23}$$

- If $X \sim Laplace(\mu, b)$ and $Z \sim Laplace(0, 1)$ then define $X^* = \frac{X - \hat{\mu}(\mathbf{X})}{\hat{b}(\mathbf{X})}$ and $Z^* = \frac{Z - \hat{\mu}(\mathbf{Z})}{\hat{b}(\mathbf{Z})}$.

- We can see that p-values are following $U[0,1]$ distribution and quantile difference between KS-statistic follows Normal distribution with mean approximating to $0(-5.745044 * 10^{-5})$ and small variance$(5.219668 * 10^{-5})$.
- Hence, $X^* \stackrel{d}{=} Z^*$.

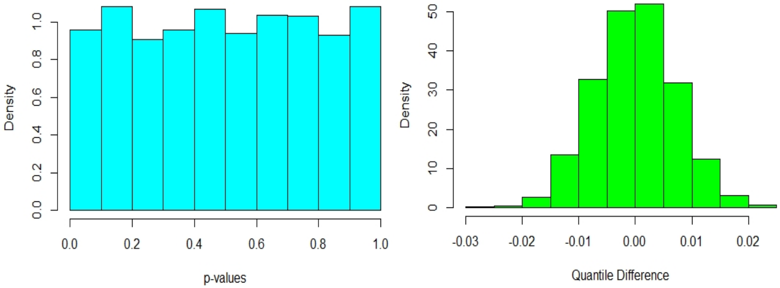

Figure: Laplace Distribution

Conclusion

- For complete data, Type I and Type II(Method 1) censoring grid size $n \times 12$ is working very well

- Method 2 and 3 are not as good as Method 1 for Type II censoring with grid size $n \times 12$

- Generalized the idea for the case when the parameters are unknown for complete data and can also be done for Type-I and Type-II censored data.

References

1. 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,
2. Koziol, J. A., and D. P. Byar.(1975). Percentage points of the asymptotic distributions of one and two sample K-S statistics for truncated or censored data. *Technometrics* **17** (4):507–510.
3. Durrett, R. 2019. *Probability: Theory and Examples*, Chapter 8. New York, USA: Cambridge University Press.
4. Fleming, T. R., J. R. O'Fallon, and P. C. O'Brien. 1980. Modified Kolmogorov–Smirnov test procedures with application to arbitrarily right-censored data. *Biometrics* **36** (4):607–625.
5. Dufour, R., and U. R.Maag. 1978. Distribution results formodified Kolmogorov–Smirnov statistics for truncated or censored samples. *Technometrics* **20** (1):29–32.
6. Zhang, J. 2002. Powerful goodness-of-fit tests based on the likelihood ratio. *J. Royal Stat. Soc. Ser B.* **64** (2):281 294.


Thank you!