SIMULATING TYPE-II CENSORED DATA FOR DISCRIMINATING BETWEEN THE WEIBULL AND LOG-NORMAL DISTRIBUTIONS

Sai, Harsha, Gupta, Kamra, Bikash*

Indian Institute of Technology Guwahati

Abstract

Log-normal and Weibull distributions are the two most popular distributions for analyzing skewed lifetime data. In this report, we simulate Type-II censored data for verifying results published by Arabin Kr. Dey and D. Kundu.

I. Introduction

It is quite difficult to distinguish between data generated from Weibull or Lognormal distributions because the cumulative distribution functions of both types are very close to each other. In the paper [1], the difference between log-likelihood of Weibull and Lognormal is used to decide on which distribution to model the dataset on. The difference has been proved to be asymptotically normally distributed.

II. The Log-Normal and Weibull Distributions

The density function of the Log-Normal distribution and the Weibull distribution is presented below. It is assumed that the density function of a log-normal random variable with scale parameter $\eta > 0$ and a shape parameter $\sigma > 0$ is

$$f_{LN}(x;\sigma,\eta) = \frac{1}{\sqrt{2\pi}x\sigma} e^{-\frac{1}{2}\{\frac{\ln x - \ln \eta}{\sigma}\}^2} \quad (1)$$

Similarly, the density function of a Weibull distribution, with shape parameter $\beta > 0$ and scale parameter $\theta > 0$ is

$$f_{WE}(x;\beta,\theta) = \beta\theta^{\beta} x^{(\beta-1)} e^{-(\theta x)^{\beta}} \quad (2)$$

III. The Problem

The problem of testing whether some given observations follow one of the two probability distributions is quite old. In this project we consider the following problem: Let $X_1, \ldots, X_n$ be a random sample from a log-normal or Weibull distribution and the experimenter observes only the first r of these, namely $X_{(1)} < \ldots < X_{(r)}$. Based on the sample the experimenter wants to decide which of the two mentioned distributions is preferable to model the data.

In [1] we use the difference of the maximized log-likelihood functions in discriminating between the two distribution functions. Suppose, $(\hat{\beta}, \hat{\theta})$ and $(\hat{\sigma}, \hat{\eta})$ are the MLEs of the Weibull parameters (β, θ) and the log-normal parameters (σ, η) respectively, based on the censored sample $X_{(1)}, \ldots, X_{(r)}$. Then we choose Weibull or log-normal as the pre-

*RollNo. 110123-11,18,28,30,38

ferred model if

$$T_n = L_{WE}(\hat{\beta}, \hat{\theta}) - L_{LN}(\hat{\sigma}, \hat{\eta}) \tag{3}$$

is greater than zero or less than zero respectively. Here $L_{WE}(.,.)$ and $L_{LN}(.,.)$ denote the log-likelihood functions of the Weibull and lognormal distributions. Without the additive constant they can be written as,

$$\begin{aligned} L_{WE}(\beta, \theta) = &\sum_{i=1}^{r} \ln f_{WE}(X_{(i)}; \beta, \theta) \\ &+(n-r)\ln(1 - F_{WE}(X_{(r)}; \beta, \theta)) \end{aligned} \tag{4}$$

and, for Lognormal distributions,

$$\begin{aligned} L_{LN}(\sigma, \eta) = &\sum_{i=1}^{r} \ln f_{LN}(X_{(i)}; \sigma, \eta) \\ &+(n-r)\ln(1 - F_{LN}(X_{(r)}; \sigma, \eta)) \end{aligned} \tag{5}$$

We can obtain the asmptotic distribution of T_n using the approach of [2]. It is observed that the asymptotic distributions are normally distributed and they are independent of the unknown parameters. The asymptotic distributions can be used to compute the probability of correct selection (PCS) in selecting the correct model. Also, PCS can be used to discriminate between the two distributions for modeling the data for a given PCS.

In this project we try to simulate data to verify the above mentioned theory for preset values of n, r, and PCS using Monte-Carlo methods. In the following sections we discuss the backgroung for determination of sample size and the algorithms for generating the data.

IV. Determination of the Sample Size

In this section we discuss a method to determine the minimum sample size needed to discrimante between the two distribution functions for a given user specified probability of correct selection and when the censoring proportion is also known. The asymptotic distributions can also used for testing purposes. Suppose it is assumed that the data are coming from $LN(\sigma, \eta)$ and the censoring proportion is p. Since T_n is asymptotically normally dostributed with mean $E_{LN}(T_n)$ and variance $V_{LN}(T_n)$, therefore, $PCS_{LN} = P(T_n \leq 0)$. Similarly, if it is assumed that the data are coming from $WE(\beta, \theta)$, then for the censoring proportion p, the PCS can be written as $PCS_{WE} = P(T_n > 0)$. Therefore, for a given p, to determine the minimum sample size required to achieved at least $\alpha*$ protection level, we equate

$$\phi\left(-\frac{n \times AM_{LN}(p)}{\sqrt{n \times AV_{LN}(p)}}\right) = \alpha* \tag{6}$$

and,

$$\phi\left(-\frac{n \times AM_{WE}(p)}{\sqrt{n \times AV_{WE}(p)}}\right) = \alpha* \tag{7}$$

and obtain n_1 and n_2 from the above two equations as,

$$n_1 = \frac{z^2_{\alpha*} AV_{LN}(p)}{(AM_{LN}(p))^2} \tag{8}$$

$$n_1 = \frac{z^2_{\alpha*} AV_{WE}(p)}{(AM_{WE}(p))^2} \tag{9}$$

Here, z_α is the α-th percentile point of a standard normal distribution. From this we can take $n = max n_1, n_2$ the minimum sample size required to achieve at least $\alpha*$ protection level for a given p. The asymptotic values for $AM_{LN}(p)$,$AV_{LN}(p)$,$AM_{WE}(p)$,$AV_{WE}(p)$ are provided in [1] along with there derivations. In the next section we mention the algorithm used in generating the data.

V. Algorithm

I. Generating data from Weibull

Data: p, the censoring proportion and n, size of the dataset
Result: α, the probability of correct selection

1 dataset = rweibull(n, 1, 1);
2 censoredData = dataset(1:n$\times$ p);
3 T = c(T, mleWE(censoredData, n) - mleLN(censoredData, n);
4 $\alpha = \{T_i : T_i > 0\}/n \times p$;

II. Generating data from Log-normal

Data: p, the censoring proportion and n, size of the dataset
Result: α, the probability of correct selection

1 dataset = rlnormal(n, 1, 1);
2 censoredData = dataset(1:n$\times$ p);
3 T = c(T, mleWE(censoredData, n) - mleLN(censoredData, n);
4 $\alpha = \{T_i : T_i < 0\}/n \times p$;

VI. Implementaion using R

I. Generate Data for Weibull

```
source('mlewe.R')
source('mleln.R')

n_sim <- 10000
p <- seq(from = 0.3, to = 0.9, by
    = 0.1)
n <- c(20, 40, 60, 80, 100, 200)
pcs <- matrix(0, nrow=length(p),
    ncol=length(n))
for(iter_n in 1:length(n)) {
  for(iter_p in 1:length(p)) {
    for(i in 1:n_sim) {
      dataset <- rweibull(n[iter_n
          ], shape = 0.8, scale =
          1)
      censored <- dataset[1:n[iter
          _n]*p[iter_p]]
      pcs[iter_p, iter_n] <- pcs[
          iter_p, iter_n] + sign(
          mleweibull(censored, n[
          iter_n]) - mlelnormal(
          censored, n[iter_n]))
    }
    pcs[iter_p, iter_n] <- (n_sim
        + pcs[iter_p, iter_n])/(2*
        n_sim)
  }
}

colnames(o) <- n
rownames(o) <- p
pcs <- as.table(pcs)
```

II. Generate Data for Log-Normal

```
source('mlewe.R')
source('mleln.R')

n_sim <- 10000
p <- seq(from = 0.3, to = 0.9, by
    = 0.1)
n <- c(20, 40, 60, 80, 100, 200)
pcs <- matrix(0, nrow=length(p),
    ncol=length(n))
for(iter_n in 1:length(n)) {
  for(iter_p in 1:length(p)) {
    for(i in 1:n_sim) {
      dataset <- rlnorm(n[iter_n],
          meanlog = 1, sdlog = 1)
      censored <- dataset[1:n[iter
          _n]*p[iter_p]]
      pcs[iter_p, iter_n] <- pcs[
          iter_p, iter_n] + sign(
          mleweibull(censored, n[
          iter_n]) - mlelnormal(
          censored, n[iter_n]))
    }
```

```
    pcs[iter_p, iter_n] <- (n_sim
        + pcs[iter_p, iter_n])/(2*
        n_sim)
  }
}

colnames(o) <- n
rownames(o) <- p
pcs <- as.table(pcs)
```

III. MLE for Weibull

```
library(MASS)

mleweibull <- function(x, L) {
  f <- fitdistr(x, densfun="
      weibull")
  t <- unlist(f, use.names = FALSE
      )
  beta <- t[1]
  theta <- t[2]
  mw <- sum(log(dweibull(x, shape
      = beta, scale = theta))) + (
      L-length(x))*log(1 -
      pweibull(x[length(x)], shape
       = beta, scale = theta))
  return(mw)
}

ldiffbeta <- function(x, b, t)
    length(x)/b + length(x)*log(t)
     + sum(log(x)) - sum((t*x)^b *
     log(t*x))

ldifftheta <- function(x, b, t)
    length(x)/t - t^(b-1) * sum(x^
    b)
```

IV. MLE for Lognormal

```
mlelnormal <- function(x, L) {
  mu <- sum(log(x))/length(x)
  sigma <- sqrt(sum((log(x) - mu)
      ^2)/length(x))
  mw <- sum(log(dlnorm(x, meanlog
      = mu, sdlog = sigma))) + (L-
      length(x))*log(1 - plnorm(x[
      length(x)], meanlog = mu,
      sdlog = sigma))
  return(mw)
}
```

Conclusion

We have faced numerous issues while generating the correct data set. With some careful observation we were able to resolve a few. Noteable among them are

References

[1] [Arabin Kr. Dey and Debasis Kundu, 2009] IEEE Transactions on Reliability , vol. 58, no. 3, 416-424, 2009. " Discriminating among the log-normal, Weibull and generalized exponential distributions"

[2] [Bhattacharyya,G.K., 1985] Journal of American Statistical Association , vol. 80, 398-404, 1985. " The Asymptotic of Maximum Likelihood and Related Estimators Based on Type-II Censored Data"