Transactions of the Nigerian Association of Mathematical Physics
Volume 14, (January - March, 2021 Issue), pp111 –116

ORDER STATISTICS APPROACH TO ESTIMATING MISSING DATA IN FLEXIBLE ELLIPTICAL PROCESSES

*[*1]A.T. SÓYÍNKÁ, [2]A.A. OLÓSUNDE, [1]A.O. WÁLÉ-ORÒJO and [1]K.M. YUSUFF*

[1]Department of Statistics Federal University of Agriculture, Alabata, Abeokuta, Ogun State, Nigeria.
[2]Department of Mathematics, Obafemi Awolowo University, Ile-ife, Osun state, Nigeria.

Abstract

Despite the fact that aging effects is a phenomenon that is unavoidable and must be given proper follow up attention by every individual; the inadequacy of good and relatively affordable health services has been the major discouragement for many individuals from following up with their health status. This trend of lack of follow up data has created some missing link and thus pose a huge setback to research and development due to data missing at random (on blood pressure history, liver function history, kidney function history and so on of individuals). Hence in this study we obtain the estimate of the parameters of i th order statistics exponential power distribution which is a member of elliptical contoured family and also develop the model to obtain the missing value(s) within data set that are exponential power distributed using order statistics approach. The maximum likelihood estimation method was used to obtain the i th order statistics scale and the location parameters while kurtosis maximization was used to determine the shape parameter via normalp. The missing value(s) are obtained by maximizing the joint order statistics distribution within a single data set that is truncated at a missing point $x_o < x < x_o < y$. Application to missing data in the weights of depressive patients at the psycho-geriatric clinic of the Federal Neuro-Psychiatric Hospital Aro Abeokuta was used to demonstrate the workability of the model.

***Keywords*:** Order statistics, exponential power distribution, parameter estimation, missing at random.

1. Introduction

Whenever there is a missing data (x_o) within a data set; then ordering the data set in ascending order implies that the missing data occupies an unknown particular position within the data set. Supposing the missing data occupies the position in between i th ordered data and $(n-i-1)$ ordered data. Then starting with the joint i th and j th order statistics distribution, we can obtain the order statistics distribution of the missing data (x_o) defined over the interval $-\infty < x < x_o < y < \infty$.

2. Literature review and Findings

Definition 1: A random variable $(rv)\ X$ is said to have a univariate Exponential Power Distribution (EPD) if

$$f(x) = \frac{1}{\sigma\Gamma(1+\frac{1}{2\beta})2^{1+\frac{1}{2\beta}}} \exp\left(-\frac{1}{2}\left|\frac{x-\mu}{\sigma}\right|^{2\beta}\right); -\infty < |x-\mu| < \infty, \sigma > 0, \beta > 0, \quad 2.1$$

Correspondence Author: Soyinka A.T., Email: soyinkaat@funaab.edu.ng, Tel: +2348038220589

where β is the shape parameter, μ and σ are location and scale parameters respectively. When $\beta = 1/2$ the function becomes a double exponential distribution and when $\beta = 1$ the function becomes a normal density. The function approaches a uniform density as β values increases beyond one towards infinity [1,2,3].

Definition 2: If $f(x)$ and $F(x)$ are the probability density function (pdf) and cumulative distribution function (cdf) of an unordered random samples 'x' with size n then after ordering the samples in order of increasing magnitude such that $X_{1:n} \le X_{2:n} \le X_{3:n} \le ... \le X_{n:n}$ then the pdf of the new RV $X_{i:n}$ is given as

$$f_{i:n}(x) = \frac{n!}{(i-1)!(n-i)!}[F(x)]^{i-1} f(x)[1-F(x)]^{n-i}; -\infty < x < \infty \qquad 2.2$$

Definition 3: if the order of ordering is such that $X_{i:n} < X_{j:n}$ then the joint pdf is given as

$$f_{i,j:n}(x,y) = \frac{n![F(x)]^{i-1} f(x)}{(i-1)!(j-i-1)!(n-j)!}[F(y)-F(x)]^{j-i-1}[1-F(y)]^{n-j} f(y); -\infty < x < y < \infty \qquad 2.3$$

[4,5]. Expression for the order statistics pdf and cdf of logistic distribution, normal distribution, uniform distribution are available in literatures [5,6,7,8]. We thus obtained the pdf and cdf of EPD to accommodates for light and heavy tailed member distributions (normal, double exponential, uniform etc) depending on the shape parameter.

3. Estimating the parameters of EPD order statistics for n ordered samples

Substituting the expression of $F(x)$ and $f(x)$ into $f_{i:n}(x)$ while expanding $[1-F(x)]^{n-i}$ we obtain

$$f_{i:n} = \frac{n!}{(i-1)!(n-i)!}\sum_{j=0}^{n-i}\frac{(n-i)!}{j!(n-i-j)!}(-1)^j \left(\frac{1}{\sigma\Gamma(1+\frac{1}{2\beta})2^{(1+\frac{1}{2\beta})}}\right) \qquad 3.1$$

$$\sum_{k=0}^{n}\left(\frac{1}{4(\Gamma(k+1+\frac{1}{2\beta}))}\right)^{i+j-1} \left|\frac{x-\mu}{\sigma}\right|^{(2\beta k+1)(i+j-1)} e^{-(\frac{i+j}{2})\left|\frac{x-\mu}{\sigma}\right|^{2\beta}}$$

Taking the log likelihood function of equation 3.1 and applying the MLE, then the derivative expression for the estimate of μ, σ, and β for n ordered samples of EPD can be obtained by solving the equation

$$\frac{\partial \ln L_{i:n}(x,\mu,\sigma,\beta)}{\partial\mu} = \sum_{j=1}^{n}(-1)^j\left(-\frac{n(i+j-1)}{|x_i-\mu|}+\frac{\beta(i+j)}{\sigma}\left|\frac{x_i-\mu}{\sigma}\right|^{2\beta-1}\right) = 0 \qquad 3.2$$

$$\frac{\partial \ln L_{i:n}(x,\mu,\sigma,\beta)}{\partial\mu} = \sum_{j=1}^{n}(-1)^j\left(-\frac{n(i+j-1)+1}{\sigma}+\frac{\beta(i+j)}{\sigma}\left|\frac{x_i-\mu}{\sigma}\right|^{2\beta}\right) = 0 \qquad 3.3$$

and

$$\frac{\partial \ln L_{i:n}(x;\mu,\sigma,\beta)}{\partial\beta} = 0 \qquad 3.4$$

in favour of μ, σ, and β respectively.

Hence solving further we obtain

$$\hat{\mu} = E(\mu) = \left(\bar{x} - \frac{s}{n}\left(\sum_{j=1}^{n}(-1)^j\frac{n(j)}{\beta(j+1)}\right)^{\frac{1}{2\beta}}\right)_{\bar{x}\ge median} \qquad 3.5$$

Likewise the expected estimate of σ^2 is

$$\hat{\sigma}^2 = E(\sigma^2) = \frac{ns^2}{\left[\sum_{j=1}^{n}(-1)^j\frac{n(j)+1}{\beta(j+1)}\right]^{\frac{1}{\beta}}} \qquad 3.6$$

Where $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}$. Note β cannot be obtained in closed form and so, we will employ the use of normalp package in the r environment to obtain the shape parameter while ensuring that the input observations are ordered $x_{i:n}$. The r code is 'paramp$(x_{i:n})$'.

4. Distribution of the missing data (x_o)

Proposition 1: The distribution $F_{i,j:n}(x_o)$ of a missing data (x_o) within an exponential power distributed data set is

$$\frac{n!\left(\left[\frac{\frac{1}{2}^{\frac{i}{2\beta}+1}}{\Gamma(\frac{i}{2\beta}+1)}\right]^{j-2}\left[\frac{(\frac{j-1}{2})^{\frac{i-1}{2}}}{(\frac{i-1}{2\beta})!}\right]\frac{\left|\frac{x_o-\mu_x}{\sigma_x}\right|^{i(j-1)}}{i(j-1)}-\frac{(\frac{1}{2})^{j-1}}{j-1}\right)}{(i-1)!(j-i-1)!(n-j)!} \qquad 4.1$$

$$\left[\left(\frac{1}{n-j+1}[1-\frac{1}{2}^{n-j+1}]-\left[\frac{\frac{1}{2}^{\frac{n-i-1}{2\beta}+1}}{\Gamma(\frac{n-i-1}{2\beta}+1)}\right]\right)^{n-j+1}\left[\frac{(\frac{n-j+2}{2})^{\frac{n-i-2}{2\beta}}}{(\frac{n-i-2}{2\beta})!}\right]\frac{\left|\frac{x_o-\mu_y}{\sigma_y}\right|^{(n-i-1)(n-j-2)}}{(n-i-1)(n-j-2)}\right]$$

Proof: Doubly Integrate equation 3.1 with respect to x within limits $\int_{-\infty}^{x_0}$ and with respect to y within limits $\int_{x_0}^{\infty}$ we obtain equation 4.1. Taking the logarithm of equation 4.1 and its derivative with respect to x_o; we obtain the median of the likely values of x_o within the region where the difference between the former and latter ordered data set is minimal as

$$\frac{\sigma_y}{\sigma_x}\left(\frac{\left|\frac{x_o-\mu_x}{\sigma_x}\right|^{i(j-1)-1}}{\left|\frac{x_o-\mu_y}{\sigma_y}\right|^{(n-i-1)(n-j+2)-1}}\right)=\frac{k_2}{k_1}\left[\frac{\left(\frac{1}{2}\right)^{j-1}}{1-\left(\frac{1}{2}\right)^{n-j+1}}\right]\left(\frac{n-j+1}{j-1}\right) \qquad 4.2$$

5. Order Statistics approach to estimating missing data in flexible elliptical processes

Whenever there is a missing data (x_o) within a data set; then ordering the data set in ascending order implies that the missing data occupies an unknown particular position within the data set. Supposing the missing data occupies the position in between i th ordered data and $(n-i-1)$ ordered data. Then starting with the joint i th and j th order statistics distribution, we can obtain the order statistics distribution of the missing data (x_o) defined over the interval $-\infty < x < x_o < y < \infty$.

Distribution of the missing data (x_o)

Proposition 2: The distribution of a missing data x_o within an exponential power distributed data set is

$$F_{i,j:n}(x_o) = \frac{n!\left(\left[\frac{\frac{1}{2}^{\frac{i}{2\beta}+1}}{\Gamma(\frac{i}{2\beta}+1)}\right]^{j-2}\left[\frac{(\frac{j-1}{2})^{\frac{i-1}{2}}}{(\frac{i-1}{2\beta})!}\right]\left|\frac{x_o-\mu_x}{\sigma_x}\right|^{i(j-1)}}{i(j-1)} - \frac{(\frac{1}{2})^{j-1}}{j-1}\right)}{(i-1)!(j-i-1)!(n-j)!}$$

$$\left(\frac{1}{n-j+1}\left[1-\frac{1}{2}^{n-j+1}\right] - \left[\frac{\frac{1}{2}^{\frac{n-i-1}{2\beta}+1}}{\Gamma(\frac{n-i-1}{2\beta}+1)}\right]\left[\frac{(\frac{n-j+2}{2})^{\frac{n-i-2}{2\beta}}}{(\frac{n-i-2}{2\beta})!}\right]\frac{\left|\frac{x_o-\mu_y}{\sigma_y}\right|^{(n-i-1)(n-j-2)}}{(n-i-1)(n-j-2)}\right) \quad 5.1$$

Proof: From equation 2.3 expand $[F(y)-F(x)]^{(j-i-1)}$ and then doubly integrate it with respect to x within limits $\int_{-\infty}^{x_0}$ and with respect to y within limits $\int_{x_0}^{\infty}$. The integrals are $\int_{-\infty}^{x_o}[F(x)]^{i+r-1}f(x)dx = \int_{-\infty}^{0}[F(x)]^{i+r-1}f(x)dx + \int_{0}^{x_o}[F(x)]^{i+r-1}f(x)dx$ and

$\int_{x_o}^{\infty}[F(y)]^{j-i-1-r}[1-F(y)]^{n-j}f(y)dy = \int_{0}^{\infty}[F(y)]^{j-i-1-r}[1-F(y)]^{n-j}f(y)dy - \int_{0}^{x_o}[F(y)]^{j-i-1-r}[1-F(y)]^{n-j}f(y)dy$.

Which implies

$$\int_{-\infty}^{x_o}[F(x)]^{i+r-1}f(x)dx = \left(\frac{(\frac{1}{2})^{(\frac{n}{2\beta}+1)}}{\Gamma(\frac{n}{2\beta}+1)}\right)\left(\frac{(\frac{i+r}{2})^{(\frac{n-1}{2\beta})}}{(\frac{n-1}{2\beta})!}\right)\left(\frac{(\frac{x_o-\mu_x}{\sigma_x})^{(n(i+r))}}{n(i+r)} - \frac{(\frac{\mu_x}{\sigma_x})^{(n(i+r))}}{n(i+r)}\right)$$

$$\sigma_x\Gamma(1+\frac{1}{2\beta})2^{(1+\frac{1}{2\beta})}$$

$$-\frac{(\frac{1}{2})^{(i+r)}}{i+r}\int_{x_o}^{\infty}[F(y)]^{j-i-1-r}[1-F(y)]^{n-j}f(y)dy = \frac{1}{n-j+1}\left(1-(\frac{1}{2})^{(n-j+1)}\right)$$

$$-\left(\frac{(\frac{1}{2})^{(\frac{n}{2\beta}+1)}}{\Gamma(\frac{n}{2\beta}+1)}\right)\left(\frac{1}{\sigma_y\Gamma(1+\frac{1}{2\beta})2^{(1+\frac{1}{2\beta})}}\right)\left(\frac{(\frac{n-j+2}{2})^{(\frac{n-1}{2\beta})}}{(\frac{n-1}{2\beta})!}\right)\left(\frac{(\frac{x_o-\mu_y}{\sigma_y})^{(n(n-j+2))}}{(n(n-j+2))} - \frac{(\frac{\mu_y}{\sigma_y})^{(n(n-j+2))}}{(n(n-j+2))}\right)$$

Simplifying further we have

$$F_{i,j:n}(x_o) = \frac{n!\,j-i-1C_r(-1)^r}{(i-1)!(j-i-1)!(n-j)!}\left(\frac{(\frac{1}{2})^{(\frac{n}{2\beta}+1)}}{\Gamma(\frac{n}{2\beta}+1)}\right)\left(\frac{(\frac{i+r}{2})^{(\frac{n-1}{2\beta})}}{(\frac{n-1}{2\beta})!}\right)$$

$$\sigma_x\Gamma(1+\frac{1}{2\beta})2^{(1+\frac{1}{2\beta})}$$

$$\left(\frac{(\frac{x_o-\mu_x}{\sigma_x})^{(n(i+r))}}{n(i+r)} - \frac{(\frac{\mu_x}{\sigma_x})^{(n(i+r))}}{n(i+r)}\right) - \frac{(\frac{1}{2})^{(i+r)}}{i+r}\times\int_{x_o}^{\infty}[F(y)]^{j-i-1-r}[1-F(y)]^{n-j}f(y)dy = \frac{1}{n-j+1}\left(1-(\frac{1}{2})^{(n-j+1)}\right)$$

$$-\left(\frac{(\frac{1}{2})^{(\frac{n}{2\beta}+1)}}{\Gamma(\frac{n}{2\beta}+1)}\right)\left(\frac{1}{\sigma_y\Gamma(1+\frac{1}{2\beta})2^{(1+\frac{1}{2\beta})}}\right)\left(\frac{(\frac{n-j+2}{2})^{(\frac{n-1}{2\beta})}}{(\frac{n-1}{2\beta})!}\right)\left(\frac{(\frac{x_o-\mu_y}{\sigma_y})^{(n(n-j+2))}}{(n(n-j+2))} - \frac{(\frac{\mu_y}{\sigma_y})^{(n(n-j+2))}}{(n(n-j+2))}\right)$$

Taking the logarithm of equation 5.1 and its derivative with respect to x_o; we obtain the median of the likely values of x_o within the region where the difference between the former and latter ordered data set is minimal.

$$\frac{\sigma_y}{\sigma_x}\left(\frac{\left|\frac{x_o-\mu_x}{\sigma_x}\right|^{i(j-1)-1}}{\left|\frac{x_o-\mu_y}{\sigma_y}\right|^{(n-i-1)(n-j+2)-1}}\right)=\frac{k_2}{k_1}\left[\frac{\left(\frac{1}{2}\right)^{j-1}}{1-\left(\frac{1}{2}\right)^{n-j+1}}\right]\left(\frac{n-j+1}{j-1}\right) \qquad 5.2$$

6. Application

The following are the weights of two different patient at every clinic day in a psychogeriatric clinic

Patient 1 :58,57,57,62,59,60,60,54,62, -,53,54,46,46,48,47,47,46

Patient 2 :50,54,51,60,67,63,63,57, -,55,52,52,45,49,47,45,45,45

Estimate the missing data for patient 1 and patient 2.

Steps to Solution: First obtain the shape parameter β assuming no part of the data set is missing. Using normalp package with code $paramp(x)$ in r software; where x is the data sets. Then obtain the estimate of the location and scale parameter for the former data sets before the missing value and the latter data sets after the missing value independently. Using the notation b as shape parameter, a and s has location and scale parameters estimate with subscript 1 for former and subscript 2 for latter data sets then we evaluate the code below and obtain the median of the data set at the region where the plot of v against x is constant. That is where the difference between the former and the latter data set is zero. The estimate of the missing data in the record of patient 1 is 53 while that of patient 2 is 57. The r code to reproduce the results and also demonstrate its workability on any set of data is in the appendix.

7. Conclusion

The estimation of single data missing at random from a data set has been demonstrated in this study via the order statistics tool. The extension of the tool to the estimation of double and multiple missing data will be considered in future study alongside with improving the paper quality via expectation maximization techniques within the region that contains the likely values of the missing data as against the use of median engaged in the paper.

8. Appendix

```
y<-c(46,47,48,53,54,57,58,59,60,62)
paramp(y)
    Mean       Mp       Sd       Sp        p
54.400000 54.020646  5.462600  6.300695  2.915005
no.conv = TRUE
b=2.915005
j<- seq(1, length(y), by=1)
a<-mean(y)-((sqrt(var(y)))/(length(y)))*(length(y)/b)^(1/(2*b))*((abs(sum((-1)^(j+1)*(j/(j+1)))))^(1/(2*b)))
d<-sqrt((length(y)*var(y))/(abs(sum((-1)^(j)*(((((length(y))*j)+1)/(b*(j+1)))))^(1/b)))
f<-(((exp(-0.5*((abs((y-a)/d))^(2*b))))*((abs((y-a)/d))^(1-2*b)))/(gamma(1/(2*b))*2^(1/(2*b))))
> plot(y,f)
> lines(y,f)
```

References

[1] Gomez, E., Gomez-Villegas, M.A., and Martin, J.M. (1998): A multivariate generalization of the exponential power family of distributions. *Communications in Statistics*, A27, 589-600.

[2] Lindsey J.K. (1999). Multivariate Elliptically Contoured Distributions for Repeated Measurements. Biometrics 55, 1277-1280.

[3] Nadarajah S. (2005). A Generalized Normal Distribution. Journal of Applied Statistics Vol 32, No 7, 685-694.

[4] David H.A. and Nagaraja H. N. (2003). Order Statistics Third Edition. John Wiley & Sons, Inc., Hoboken, New Jersey. Published simultaneously in Canada.

[5] Arnold Barry C. , Balakrishnan N and Nagaraja H.N (2008). A First Course in Order Statistics. Society for Industrial and Applied Mathematics Philadelphia.

[6] Joshi P.C. and Balakrishnan N. (1981). An identity for the moments of normal order statistics with applications. *Scandinavian Actuarial Journal*. Netherland 203-213.

[7] Balakrishnan N. and Clifford Cohen A. (1991). Order Statistics and Inference. Academic press, inc. Publisher.

[8] Lopez Blazquez, F., Balakrishnan, N., Cramer, E., and Kamps, U (2008). A course on Order Statistics and Related Models. Universidad de Sevilla. statmath.wu.ac.at.