DOI: http://doi.org/10.4038/sljastats.v21i1.8012

Sri Lankan Journal of Applied Statistics, Vol (21-1)

An Order Statistics Technique for Monitoring and Quantifying Susceptibility to Toxicity

G. A. Dawodu[1*]

[1]Department of Statistics, Federal University of Agriculture (FUNAAB), Nigeria.
[*]Corresponding Author: dawoduga@funaab.edu.ng

Received: 12th January 2020 / Revised: 9th July 2020 / Published: 30th November 2020

ABSTRACT

In launching an environmental technique, statistics plays an important role. The art of using statistical methods, models and principles to work on environmental issues is hereby utilized in dealing with a well-ordered set of quantities on exposures of some Artisans to toxicity. The Poisson was initially assumed to be the intrinsic distribution, but following a strong advocacy in favour of the Lindley over members of the exponential family of distributions, the population, $(X_1, \ldots, X_n)$, of hazardous elements, in the blood/urine of Artisans is assumed to be Lindley. The derivation of other "locational" distributions therefore originates from it. Numerical illustrations of how the proposed technique can be utilized concludes this work.

Key Words: Environmental and Ordered Statistics, Lindley Distribution, Exposure, Quantiles, Toxicity and R.

1. Introduction

Environmental issues such as Pollution manifest in many guises, two of which are; Accidental and Occupational (Dawodu et al., 2011; Eisenbud and Gesell, 1997). With respect to the latter, bad working conditions with occupational hazards and air pollution are of common occurrence in popular industrial areas. A division of occupational health providers need to be sited at major industrial centers in the country. Such a division is to study occupational health hazards and develop effective preventive and control measures. This work purports to; suggest a technique for monitoring and quantifying toxicity attributable to the type of work an Artisan does, identify the distribution of accumulated impurities found in the blood/urine of specific voluntary Artisans, by assuming that the initial intrinsic distribution of the impurities is Lindley and compute quantiles as future apparent susceptibility to toxicity of specified Artisans. The rest of this work is planned as follows; section 2 is captioned "Identification of order statistical measures obtainable on toxicity", section 3 is "On the theory of order statistics", four is on "Derivation of the r^{th} ordered observation's distribution, five is on "Computation of Quantiles", six is on "Numerical Illustrations", seven is on "Discussion and Conclusion" while eight is on the "Preference of this Order Statistics Technique over Existing Standard Methods" which is finally followed by References.

2 Identification of order statistics measures obtainable on toxicity

It is customary that a researcher studies the data he/she collected from a research immensely before he can decide on the most appropriate analysis for it. His study will usually include checking for; size of the data (say n), the presence of extreme values or outliers, minimum and maximum points, an intrinsic distribution whose simulated values could be similar to the data under consideration. Between the maximum, $X_{(n)}$ and minimum, $X_{(0)}$ are ordered values (i.e. Statistics or Quantiles, such as, Percentiles, Deciles, Quartiles. The sorting of data to obtain a result that is described here, initiates the popular order statistics).

2.1 On the theory of order statistics

The ordering, of our continuously distributed population or data, often results in,

$$X_{(0)} \leq X_{(1)} \leq \cdots \leq X_{(r)} \leq \cdots \leq X_{(n-1)} \leq X_{(n)}.$$

The ordered partitions, $[X_{(0)}, X_{(1)}), \dots, [X_{(n-2)}, X_{(n-1)}), [X_{(n-1)}, X_{(n)}]$ will further enable the creation of, $X_{(1)} < X_{(2)} < \cdots < X_{(r)} < \cdots < X_{(n-1)} < X_{(n)}$ by merely collecting the least upper bounds of each partition. The definition of order statistics does not require that the $X_{(i)}$'s be independent or identically distributed. Much of the literatures (Reiss, 1989; Balakrishnan and Rao, 1998; Lieberman and Olkin, 1991; David and Nadaraja, 1981) , on this issue, focus on the case in which they (i.e. the $X_{(i)}$'s) constitute a data from some joint/intrinsic probability function f (with the Independent and Identically Distributed (IID) assumption). The Probability Density Function (pdf) of $X_{(r)}$, the r^{th} order statistics is given by David and Nadaraja (1981) as;

$$f^{(n)}\left(X_{(r)} = x\right) = r\binom{n}{r}\left(F(x)\right)^{(r-1)}\left(1 - F(x)\right)^{(n-r)} f(x), -\infty < x < \infty \qquad (1)$$

Where; $f(x)$ is the joint/intrinsic probability function and $F(x)$ is the corresponding cumulative density function. The respective pdfs of the first and last order statistics are;

$$f^{(n)}\left(X_{(1)} = x\right) = n\left(1 - F(x)\right)^{(n-1)} f(x), -\infty < x < \infty \tag{2}$$

and

$$f^{(n)}\left(X_{(n)} = x\right) = n(F(x))^{(n-1)} f(x), -\infty < x < \infty \tag{3}$$

The cdfs of the first and last order statistics are easily derived integrating Equations (2) and (3) respectively to obtain Equations (4) and (5):

$$F^{(n)}\left(X_{(1)} = x\right) = 1 - \left(1 - F(x)\right)^{(n)}, -\infty < x < \infty \tag{4}$$

and

$$F^{(n)}\left(X_{(n)} = x\right) = (F(x))^{(n)}, -\infty < x < \infty \tag{5}$$

In general, the cdf of the r^{th} order statistics is given as;

$$F^{(n)}\left(X_{(r)} = x\right) = P\left(X_{(r)} \leq x\right) \tag{6}$$

$$= P\left(\text{at least r of } X_{(1)}, X_{(2)}, \ldots, X_{(r)}, \ldots, X_{(n-1)}, X_{(n)} \leq x\right) \tag{7}$$

$$= \sum_{i=r}^{n} P\left(\text{exackly i of } X_{(1)}, X_{(2)}, \ldots, X_{(r)}, \ldots, X_{(n-1)}, X_{(n)} \leq x\right) \tag{8}$$

$$= \sum_{i=r}^{n} \binom{n}{r} (F(x))^{i} (1 - F(x))^{(n-i)}, -\infty < x < \infty. \tag{9}$$

2.2 Derivation of the r^{th} ordered observation's distribution

Karlin and Taylor (1975) suggested the Poisson on the issue of rate of radiation emission and thus exposure rate, but Sankaran (1970) and Ghitany et al., (2008), in turn, advocated strongly in favour of the Lindley distribution amongst the exponential family of distributions. Their submissions bother on; suitability for discretization (i.e. when Lindley is convoluted with the Poisson to form Poisson-Lindley, a continuous distribution becomes discretized) and flexibility of properties (including those of shape). Armed with the order statistics models and the assumed intrinsic Lindley, the derivations of first, r^{th} and last order statistics goes thus; From Ghitany et al., (2008), the pdf, f(x) and cdf, F(x) of Lindley are;

$$f(x) = \frac{\lambda^2}{\lambda+1}(1 + x)e^{-\lambda x}, x > 0, \lambda > 0 \tag{10}$$

and

$$F(x) = 1 - \frac{\lambda x+\lambda+1}{\lambda+1} e^{-\lambda x}, x > 0, \lambda > 0 \tag{11}$$

With mean, μ and its cdf, $F(\mu)$ given, respectively, as;

$$\mu = E(X_i) = \frac{\lambda+2}{\lambda(\lambda+1)}, i = 1,2, \dots, n, \lambda > 0 \tag{12}$$

and

$$F(\mu) = 1 - \frac{\lambda^2+3\lambda+3}{(\lambda+1)^2} e^{-\frac{(\lambda+2)}{(\lambda+1)}}, \lambda > 0 \tag{13}$$

Now, using Equations (1) through (5) in conjunction with (9), (10) and (11) to derive the; first, r^{th} and last order statistics pdfs and cdfs respectively, we have;

$$f^{(n)}(X_{(1)} = x) = \frac{n\lambda^2}{(\lambda+1)^n} (\lambda x + \lambda + 1)^{(n-1)} (1+x) e^{-n\lambda x}, x > 0, \lambda > 0 \tag{14}$$

$$f^{(n)}\left(X_{(r)} = x\right) = \binom{n}{r} \frac{(1+x) r \lambda^2 e^{-\lambda(n-r+1)x}}{\lambda+1} \left(1 - \frac{\lambda x + \lambda + 1}{\lambda+1} e^{-\lambda r}\right)^{(r-1)} \left(\frac{\lambda x + \lambda + 1}{\lambda+1}\right)^{(n-r)}, x > 0, \lambda > 0 \tag{15}$$

$$f^{(n)}\left(X_{(n)} = x\right) = \frac{n\lambda^2 (1+x) e^{-\lambda x}}{\lambda+1} \left(1 - \frac{\lambda x+\lambda+1}{\lambda+1} e^{-\lambda x}\right)^{(n-1)}, x > 0, \lambda > 0 \tag{16}$$

$$F^{(n)}\left(X_{(1)} = x\right) = 1 - \left(\frac{\lambda x+\lambda+1}{\lambda+1} e^{-\lambda x}\right)^n, x > 0, \lambda > 0 \tag{17}$$

$$F^{(n)}\left(X_{(r)} = x\right) = \sum_{i=r}^{n} \binom{n}{r} \left(1 - \frac{\lambda x+\lambda+1}{\lambda+1} e^{-\lambda x}\right)^i \left(\frac{\lambda x+\lambda+1}{\lambda+1} e^{-\lambda x}\right)^{(n-i)}, x > 0, \lambda > 0 \tag{18}$$

$$F^{(n)}\left(X_{(n)} = x\right) = \left(1 - \frac{\lambda x+\lambda+1}{\lambda+1} e^{-\lambda x}\right)^n, x > 0, \lambda > 0 \tag{19}$$

2.3 Computation of Quantiles

To initiate the computation of a quantile, x^*, Equation (12) is first solved to obtain the value of "universal" $\lambda > 0$;

$$\lambda = \frac{\sqrt{\mu^2 + 6\mu + 1} - (\mu - 1)}{2\mu}$$

After which, the pdf or cdf of order statistics that is of interest will be respectively; integrated or evaluated in the interval, $[0, x^*]$ (located at the corresponding values of the order statistics, $X_i, i = 1,2,\dots,n$ and universal $\lambda > 0$), equated to a corresponding fraction, $0 < \rho < 1$ and solved for x^*.

2.4 Numerical Illustrations

The following data (i.e. lblood and lurine) is an extract of the data that was obtained on Artisans working within Abeokuta metropolis, about a decade ago, the work was done by the former Department of Biochemistry, College of Natural Sciences, University of Agriculture, Abeokuta. Data was collected on the quantities of lead (Pb), Calcium (Ca), Sodium (Na), Potassium (K) etc. in the blood and urine samples from volunteers (the unit was in part-per-million (ppm)). The Artisans were in seven categories with respect to their trades, there were two Petrol Attendants categories as well. The population size is 118 (i.e. $n = 118$) Dawodu et al., (2011).

Example I

The lead-in-blood (i.e. lblood) is used here, the summary and order statistics are as stated in tables 1 and 2 below; With respect to table 1 (it is noteworthy, the difference in measures of central tendencies, with or without the presence of outliers (i.e. median and mean respectively), because if the data size is as large as 118 and yet the two measures are not approximately equal then, the "idea" of using the normal distribution in place of Lindley should be "laid" to rest, at least, until when the data size is in "tune" of thousands), set $\mu = 32.14$ to obtain λ as contained in Equation (20);

$$\lambda = \frac{\sqrt{(32.14)^2 + 6(32.14) + 1} - (32.14 - 1)}{2(32.14)} = 0.06045411 \approx 0.06 \qquad (20)$$

Now, assuming an "Apprentice" or a new entrant join the Artisans later, after working for some time his blood sample was taken and the quantity of lead (Pb) in his blood (say, 10.50 ppm) is brought forward, then the identification of his lead-in-blood distribution and "status" (i.e. quantile state) through the existing population data can be done as follows;

In Equations (14) and (17), set $n = 1$ to obtain Equation (10), which is his distribution. As for the quantile, Equation (11) will give that by evaluating it in the interval $0 < x < 10.5$ after setting $\lambda = 0.06$. That is;

$$[F(x)]_0^{10.5} = \left[1 - \frac{0.06x+1.06}{1.06} e^{0.06x}\right]_0^{10.5} = 1 - 0.8491324 \approx 0.15 \tag{21}$$

Consequently, his mean is also 32.14 ppm whilst his quantile state is 15th percentile.

Example II

If there were two readings of lead-in-blood for identification of distributions and quantile states, for the sake of precision, three or more readings should be avoided because it will require numerical integrations and multiple approximations. Equations (15) and (16) will be utilized for identification of the respective distributions whilst Equations (18) and (19) will be used for the respective quantile states. The smaller reading will be used with respect to Equations (15) and (18) whilst the second will be used with respect to Equations (16) and (19). Now, let the readings be 18.2 and 55.1, then, with respect to 18.2 ppm, its corresponding pdf is obtained by setting $n = 2, r = 1$ in Equation (15) to obtain Equation (22);

$$f^{(2)}\left(X_{(1)} = x = 18.2\right) = \frac{2(1+x)\lambda^2 e^{-2\lambda x}}{(\lambda+1)^2}(\lambda x + \lambda + 1), \lambda > 0 \tag{22}$$

An attempt to estimate his mean in the long-run will involve putting $\lambda = 0.06$ into Equation (22) to obtain approximately

$$f^{(2)}\left(X_{(1)} = x = 18.2\right) \approx (0.0004x^2 + 0.0072x + 0.007)e^{(-0.12x)} \tag{23}$$

And integrating Equation (23) in $x > 0$ will give his mean

$$E(X_1 = 18.2) = \int_0^\infty x\, f^{(2)}\left(X_{(1)} = x = 18.2\right)dx \approx \int_0^\infty (0.0004x^3 + 0.0072x^2 + 0.007x)\, e^{(-0.12x)}dx = 20.39 \tag{24}$$

The estimation of second reading in the long-run is done using Equation (16) to obtain;

$$E(X_2 = 55.1) = \int_0^\infty 0.0068\, x(1+x)e^{(-0.06x)}\left(1 - (0.0566x + 1)e^{(-0.06x)}\right)dx \approx 44.93 \tag{25}$$

Actually, the readings do not play any role in the estimation of "long-run" means. That is, any single reading, $x < 32.14$, is supposed to have 32.14 ppm as its long-run mean and any pair of readings $x_1 < 32.14$, $x_2 > 32.14$ are supposed to have long-run means, 20.39 and 44.93 respectively. However, readings do perform roles in the location of their quantile states. As for the first reading in the pair, its quantile state is obtained, using Equation (15);

$$F^{(2)}(X_1 = 18.2) = \left[1 - (\frac{(0.06x+1.06)e^{(-0.06x)}}{1.06})^2\right]_0^{18.2} \approx 0.536 \qquad (26)$$

It is 53.6 percentile in its distribution. With respect to the second, Equation (19) is utilized to obtain;

$$F^{(2)}(X_2 = 55.1) = \left[(1 - \frac{(0.06x+1.06)e^{(-0.06x)}}{1.06})^2\right]_0^{55.1} \approx 0.72 \qquad (27)$$

This is the 72th percentile in its distribution.

3 Discussion and Conclusion

Pollution matters should not be handled with levity, most especially avoidable ones. In the cases of Energy stations accidents at Chernobyl and Fukushima (Eisenbud and Gesell, 1997), little or nothing could have been done to avoid them, except that their aftermaths ought to be better managed. It may not be possible to totally eradicate environmental pollution because, at some locations, where one least expects them to be, presence of radioactive and carcinogenic elements like uranium and radon, for instance, could be accidentally felt (Rosenbaum, 1995; Dawodu and Mustapha, 2017). However, for avoidable ones, like exposure to radiation or hazardous elements, while at work, they could be well-managed, such that, their effects would have been usurped or reduced to the minimum, through the use of; awareness campaigns, safety gadgets and effective continuous monitoring. This work is part of the attempt to device an effective continuous monitoring technique to assist Artisans that are exposed to hazardous elements at their workshops. In section 5, "ties" may be involved with the ordered statistics. However, their presence would not distort the results because whenever they occur, the ties will possess the same long-run means and quantile states. In conclusion, the technique herein proposed, is effective (for instance, it did detected that the reading, 55.1 of the second example is past its central tendencies (i.e. mean and median) and goes to fix its quantile state, accurately as the 72th percentile and paves the way for works involving Bayesian approaches whose intrinsic simulation method; Gibbs sampling, Metropolis hasting, Markov Chain Monte Carlo etc.), can be used to evaluate the numerous integrations and approximations whenever three or more sample readings are taken, further Equation (13) even supplies the cdf of μ freely to whoever intends to go Bayesian.

3.1 Preference of this Order Statistics Technique over Existing Standard Methods

The "susceptibility to toxicity" of the Artisans' blood can only be obtained after they have been exposed to the elements of their trades for a period (preferably in years). This statistic is not usually exhibited but is of paramount importance in occupational epidemiology. Prior to the publishing of Dawodu et al., (2011), what the epidemiologist does while reporting occupational surveys (Eisenbud and Gesell, 1997; Steenland, 1993) amounts to:

1. Quantify the exposure externally or internally through the blood and/or urine of Volunteers (probably in cohorts).

2. Monitor the existence of Volunteers through their survival analyses. For example, identifying their number when the study began, keeping the records of those who died or suffer ailments traceable to the exposure and using that to calculate the statistic.

3. Giving ratios and percentages of survivors as the study continues as measure of risk and/or odds ratio after using the relevant logistic formula.

In Dawodu et al., (2011), the blood and/or urine of individual Artisan was examined with respect to the presence of some element (e.g. lead(Pb)). The table of correlation coefficients for quantified exposures with respect to the elements under study was given. The descriptive statistics of the exposure elements were given and Quality Control Methodology (QCM) was used to derive the model for predicting the statistic. QCM assumes the accumulation of individual element in the blood/urine of an Artisan is a process that will go out of control when the Artisan is susceptible. In the present technique, individual Artisans can now be statistically examined through their unique distributions. This enables the researcher to know their individual Cumulative Density Functions (cdf) and quantiles. This will enable the researcher to know (provided the Artisan lives long enough), when (in years, after joining a trade) he will be mildly ($\leq 25\%$), averagely ($\leq 50\%$) and highly ($> 50\%$) susceptible. Finally, the determination of individual quantiles connoting the time in years Artisans ought to retire from the work (Artisan's) was not available through known techniques before now.

References

1. Balakrishnan, N. and Rao, N. C., eds. (1998). *Handbook of Statistics,* Elsevier Science, Vol. 17.

2. Dawodu, G.A., Asiribo, O.E., Adelakun, A.A., Ozoge, M.O., Ademuyiwa O. and Akinwale, T.A. (2011). On the Vulnerability of the blood of some Artisans to Toxicity, *Journal of Environmental Statistics.*, 2:4, 1-17.

3. Dawodu, G.A. and Mustaph, A.O. (2017). Modelling Indoor Radon Through Dimensional Analysis, *Transactions of the Nigerian Association of Mathematical Physics*, Vol. 5, 231-240.

4. David, H.A. and Nagaraja, H. N. (1981). *Order Statistics (Third Edition),* Wiley Series in Probability and Statistics, (1981), North Carolina.

5. Eisenbud, M. and Gesell, T. (1997). *Environmental Radioactivity (from Natural, Industrial and Military sources) Fourth Edition*, Academic Press (An imprint of Elsevier). San Diego.

6. Ghitany, M. E., Atieh, B. and Nadarajah, S. (2008). Lindley distribution and its application, *Mathematics and Computers in Simulation*, Vol. 78, 493-506.

7. Karlin, S. and Taylor, H. M. (1975). *A First Course in Stochastic Processes (Second Edition),* Academic Press, New York.

8. Lieberman, G. L. and Olkin, I., eds. (1991). *Statistical Modelling and Decision Science*, Academic Press Inc., Vol. 1. Stanford University, Stanford California.

9. R Core Team. (2016). *R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing,* Vienna, Austria.

10. Reiss, R. D. (1989). *Approximate Distributions of Order Statistics: With Applications to Nonparametric Statistics*, Springer-Verlag, New York Inc., ISBN-13:978-1-4613-9622-2, e-ISBN-13:978-1-4613-9620-8, DOl: 10.1007/978-1-4613-9620-8.

11. Rosenbaum P. R. (1995). Quantiles in Nonrandom Samples and Observational Studies, *Journal of the American Statistical Association*, Vol. 90:432, 1424-1431.

12. Steenland, K. (editor) (1993). *Case Studies in Occupational Epidemiology*, Oxford University Press, Oxford.

13. Sankaran, M. (1970). The discrete Poisson-Lindley distribution. *Biometrics*, Vol. 26, 145-149.