International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:6, No:1, 2012

A Thought on Exotic Statistical Distributions

R K Sinha

***Abstract*—**The statistical distributions are modeled in explaining nature of various types of data sets. Although these distributions are mostly uni-modal, it is quite common to see multiple modes in the observed distribution of the underlying variables, which make the precise modeling unrealistic. The observed data do not exhibit smoothness not necessarily due to randomness, but could also be due to non-randomness resulting in zigzag curves, oscillations, humps etc. The present paper argues that trigonometric functions, which have not been used in probability functions of distributions so far, have the potential to take care of this, if incorporated in the distribution appropriately. A simple distribution (named as, Sinoform Distribution), involving trigonometric functions, is illustrated in the paper with a data set. The importance of trigonometric functions is demonstrated in the paper, which have the characteristics to make statistical distributions exotic. It is possible to have multiple modes, oscillations and zigzag curves in the density, which could be suitable to explain the underlying nature of select data set.

***Keywords*—**Exotic Statistical Distributions, Kurtosis, Mixture Distributions, Multi-modal

I. Introduction

STATISTICS, as a subject, has flourished consistently over the years, and is being used heavily in many disciplines, including multi-disciplinary subjects. Amongst the various uses, the applications of statistical distributions have become an important ingredient in various analyses involving modeling. These distributions help in explaining nature of various types of data sets. The individual statistical distributions could be discrete or continuous. The continuous distributions are more versatile and are used more frequently, because the random variables happen to be continuous quite often. The majority of continuous distributions could be grouped into two categories, viz. Beta family (such as, Pareto, GPD, burr, Dagum, log-logistic, Para log-logistic etc.).

The choice of a statistical distribution purely depends upon the underlying data set, being modeled. For example, one would not choose a discrete distribution if the variate in the data set is continuous. The complexity and versatility of a distribution usually increases with the number of parameters involved in it. This is because the larger number of parameters in a distribution makes the estimation procedure more complicated, but at the same time generally makes density curve more flexible tracking the given data. The number of parameters varies from one to four in the traditional statistical distributions. Generally, a four parameter distribution is more likely to capture first four moments (central tendency, variability, skewness and kurtosis) more precisely than that of distributions with smaller number of parameters. The selection of an appropriate procedure of estimation (such as, method of moments, method of maximum likelihood estimation, method of least squares etc.) is an important aspect of modeling given a data set, as there could be wide variation in the parameter estimates under some of the circumstances. It is common to estimate parameters through more than one procedure to have an insight into their potential variations. Ultimately, the statistical tests (such as, Kolmogorov-Smirnov, Anderson-Darling, Chi-square etc.) need to be carried out to conclude the validity and performance of the model.

It is quite common to see multiple modes in the observed distribution of the underlying variables, consisting of few big ones near central tendency of the data and few-to-many (smaller ones) in the tails. Nevertheless, the observed data does not exhibit smoothness primarily due to the presence of randomness. The non-smoothness of the observed data could still be present, which might be in the forms of specific zig-zag curves, oscillations, humps etc. Having a clear cut trend of these, one may note that these will be very likely due to non-randomness rather than random fluctuations merely. In fact, it is not difficult to find data sets, which have certain well defined humps etc., which is ignored by the model.

The above genuine failure is, in fact, because of the simple reason that the continuous statistical distributions happen to be uni-modal. The density function of such distributions is non-decreasing function of the variate in the range (from lowest value to mode) and non-increasing function thereafter (from mode to highest value). As a special case, the same remains constant throughout, in case of rectangular distribution. The issue remains here is, could we have a distribution, which could capture this behaviour of data.

II. Few illustrations

A. Review of previous literature

There are several studies [5]-[6], which have discussed that even an appropriate statistical distribution often fails to capture the nature of underlying data in the tails, especially in the finance and insurance sector, as these data sets happen to be highly positive skewed and leptokurtic. In case of insurance, under-estimation/over-estimation at the tail end could lead to mis-pricing of products and impact the business and financial position of the insurer and/or re-insurer. [5] applied the extreme value theory and demonstrated to model the tail of such data sets, wherein a Danish fire loss data set was illustrated. The study also gave a link between the conditional and unconditional probabilities, which could be used for the transformation of the variate.

Several researchers [1]-[3], [9] have developed mixture / composite models by mixing two distributions appropriately to cope with this problem, which requires estimation of tail

Dr. R. K. Sinha is with Insurance Regulatory and Development Authority (IRDA), Hyderabad, India (phone: +9140-2349-1813; fax: +9140-6678-9768; e-mail: rksinha@irda.gov.in).

separately and consolidating complete data set subsequently. These models have also been tested on the same data set viz. Danish fire insurance data set, as indicated above. Incidentally, [7] demonstrated that a 4-parameter Burr distribution is a good choice for the said data set, which is found to be quite competitive with respect to these mixture models. Further, the study identifies statistically significant variations in the fire losses/claims pertaining to the summer and winter season. At the end, it highlights some of the statistical limitations of mixture distributions, and prescribes that these applications should be need based.

The existence of multi-modes can be observed in many data sets, which have logical interpretation too. As an illustration, [8] attempted to model the length of stay (LOS) of patients at hospitals, which witnessed two clear cut modes, one at around 2-3 days and the other one near 30 days. The study modeled LOS using a log-normal distribution, although the limitation of modeling two modes by a single statistical distribution remained an issue.

The above mentioned bi-modality is presented in Fig. 1A and Fig. 1B.

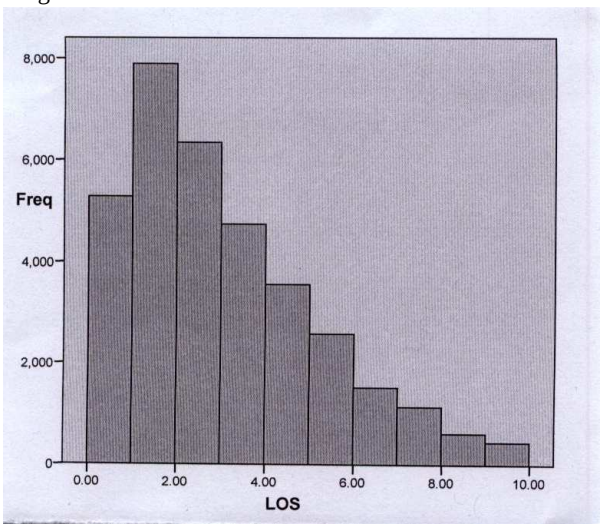

Fig. 1A Histogram - LOS of upto 10 days

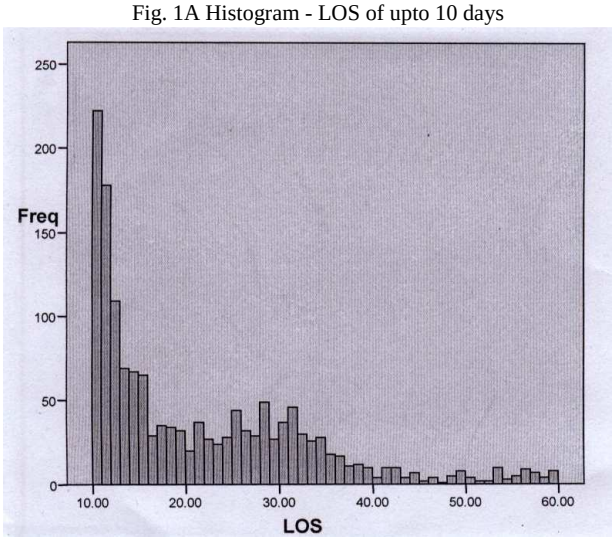

Fig. 1B Histogram – LOS of more than 10 days

The study of [4] introduced spectral analysis of distributions (SAD), a method for detecting and evaluating possible periodicity in experimental data distributions of arbitrary shape. The SAD was used in determining whether a given empirical distribution contains a periodic component. The paper proposed a system of probabilistic mixture distributions to model such data sets having periodicity. The analysis was demonstrated to the eukaryotic enzyme length data. The presence of well defined multi-modes can be seen in their study, (Fig. 2).

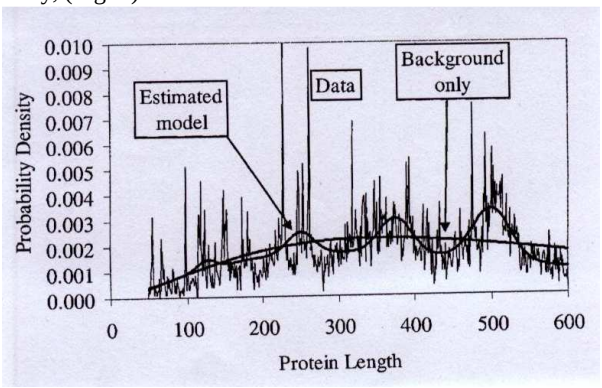

Fig. 2 Estimated probability density of eukaryotic enzyme length, entire data set [4]

The above illustrations reveal that while the observed data can have the characteristics of multi-modes, humps and oscillations, a single distribution will never have such characteristics. Accordingly, these characteristics can never be captured through a single distribution. As can be observed from the above studies, the alternatives are largely categorized in two ways. One way is to split the data sets and model each parts separately and consolidate the estimates at the end. The other way is to use mixture distributions. Clearly, both ways have their own limitations and side-effects. The author raises concerns for these genuine limitations of traditional statistical distributions and attempts to find a way to overcome the same.

The mathematical functions can be broadly classified into 5 categories viz., ILATE, where I is Inverse, L is Logarithmic, A is Algebraic, T is Trigonometric and E is Exponential. It is interesting to note that although Inverse, Logarithmic, Algebraic and Exponential functions have been used extensively, as a sum or product etc., the Trigonometric functions have rarely been considered in the expressions of statistical distributions. Author believes that one of the problems could be its mathematical tractability, which might have led to the non-existence of distributions involving trigonometric functions. A few distributions contain SINE/COSINE hyperbolic functions, but these ultimately lead to exponential functions. Similarly, Cauchy distribution contains ARCTAN function, which is not trigonometric, as such. However, the trigonometric functions have been used for some of the other statistical distributions more appropriately, such as Hann function or raised cosine function etc.

The Hann function deals with discrete variates (integer-valued) and has its probability mass function as below:

$f(n) = 0.5 [1 – \cos\{2\pi n/(N-1)\}]$

Similarly, the raised cosine distribution has the probability density function defined as:

$f(x) = (1/2s) [1 + \cos\{(x-\mu)\pi/s\}], \quad \mu - s \leq x \leq \mu + s$

Fig.3 represents the PDF of raised cosine distribution. From the diagram, it can be seen that it is a symmetric distribution with mean = 0. Its variance is given by $s^2 (1/3 - 2/\pi^2)$.

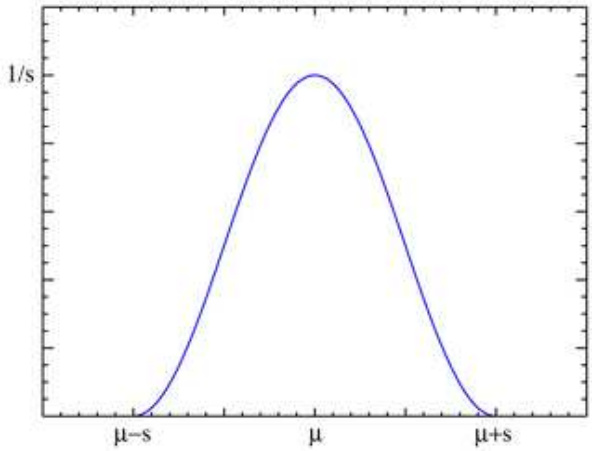

Fig. 3 PDF of Raised Cosine Distribution

Although these distributions contain trigonometric functions, these are uni-modal only. The author feels that the trigonometric functions should be attempted further to invent new statistical distributions, which could be multi-modal and non-smooth, as illustrated earlier. As discussed, the observed data is very unlikely to be smooth, and could consist of multiple modes, humps, oscillations, zig-zag curves. These attributes could be split into two parts, one because of randomness (which will always be there) and the other one because of the genuine nature of underlying data (which may or may not be present). In case the latter is present, it becomes important to identify it and capture it appropriately through a model. One could see plenty of such data sets in specific disciplines, such as, demography, insurance, medical sciences, physics etc.

III. Model

Author attempts to give an insight as to how a trigonometric function can change the density curve. A simple uniform / rectangular distribution is used to include trigonometric functions.

The probability density function of a uniform distribution is defined as:

$f(t) = 1/(b-a), \quad a \leq t \leq b$

Where a and b are parameters for boundary conditions.

The cumulative distribution function is given as:

$F(t) = (t-a)/(b-a)$

Now a SINE term is incorporated in this uniform distribution, which has a density as:

$f(t) = [1/(b-a)] [1+ \delta \sin\{(2\pi(t-a)/n\}],$

$a \leq t \leq b, 0 \leq \delta \leq 1/(b-a), n = 1, 2, \ldots..$

Where, a, b, δ and n are 4-parameters of the distribution. We may call it "Sinoform Distribution".The cumulative distribution function is given as:

$F(t) = [(t-a)/(b-a)] + [\delta n/2\pi] [1 – \{1/(b-a)\}* \cos\{2\pi(t-a)/n\}],$

It can easily be seen that it is a statistical distribution, satisfying necessary conditions, for example,

$\int f(t)\, dt = 1, F(a) = 0, F(b) = 1$ etc.

Further, the Sinoform becomes uniform as a special case when $\delta = 0$.

This could perhaps be the simplest example of distribution involving a trigonometric function with multi-modes. This could capture oscillations in the data sets.

Fig. 4 exhibits the probability density function of Sinoform distribution with parameter values, a = 0, b = 5, n = 8 and δ = 0.02.

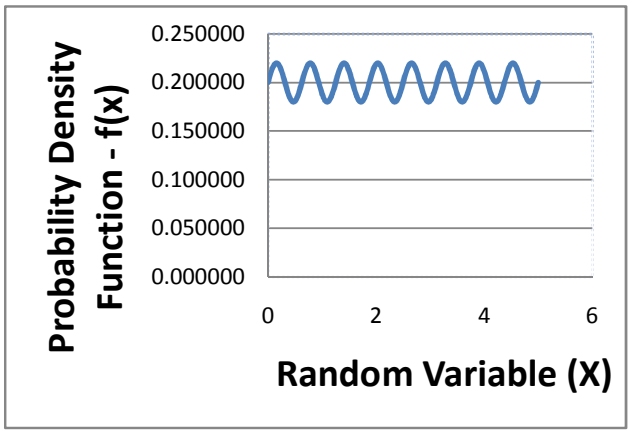

Fig. 4 PDF of Sinoform (a = 0, b = 5, n = 8 and δ = 0.02)

A more generalized uniform can be considered by allowing δ to vary with the variate.

Similarly, it is believed that if trigonometric functions are used in a more complex manner, for example, Exp[Sin(.)], $\text{Sin}^3(.)$ etc. [where, (.) is a function of variate], appropriately, it can make the density zig-zag, but with a non-random behaviour, which is what we are looking for.

IV. Conclusion

It is argued that trigonometric functions have the characteristics, which could make statistical distributions more versatile if incorporated appropriately and make the density exotic. It is possible to have multiple modes, oscillations and zig-zag curves in the density, which could be suitable to explain the underlying nature of select data sets.

It may be interesting to develop and derive mathematics for the estimation procedures for such new distributions. It is expected that these exotic distributions shall be better alternatives to many competitive mixture distributions. The author believes that there is wide scope for further work/study in this unexplored area of statistics.

References

[1] R. Ciumara, "An actuarial model based on composite Weibull-Pareto distribution," Mathematical Reports, vol. 8 (58), no. 4, 2006.

[2] K. Cooray and M. M. A. Ananda "Modeling actuarial data with a composite lognormal-Pareto model," Scandinavian Actuarial Journal, vol. 5, pp. 321-334, 2005.
[3] A. Frigessi, O. Haug and A. Rue "Dynamic mixture model for unsupervised tail estimation without threshold selection," Extremes, vol. 5, pp. 219-235, 2002.
[4] E. Kolker, B. C. Tjaden, R. Hubley, E. N. Trifonov and A. F. Siegel "Spectral analysis of distributions: finding periodic components in eukaryotic enzyme length data," OMICS, Journal of Integrated Biology, vol. 6, no.1, pp. 123-130, 2002.
[5] A. McNeil "estimating the tails of loss severity distributions using extreme value theory," ASTIN Bulletin, vol. 27, no. 1, pp. 117-137, 1997.
[6] S. Resnick "Discussion of the Danish data on large fire insurance losses," ASTIN Bulletin, vol.27, no.1, 139-151, 1997.
[7] D. V. S. Sastry and R. K. Sinha "A revisit to Danish fire loss data," Conference Proceedings, 12th Global Conference of Actuaries (GCA), Mumbai, India, 2010.
[8] D. V. S. Sastry and R. K. Sinha "Length of stay – a data analytic approach," Journal of Quantitative Economics, The Indian Econometric Society, vol. 8, no. 2, pp. 42-60, 2010.
[9] D. P. M. Scollnik "On composite lognormal-Pareto model," Scandinavian Actuarial Journal, vol. 7, no. 1, pp. 20-33, 2007.