The Ubiquity of Statistical Fractal Observations

[1]Roberto N. Padua and [2]Randy K. Salazar

Abstract

The natural world is recognized to be fractal: from the growth of a leaf to how the trees propagate to form a forest, the neural networks, the DNA and the galaxies in space. The patterns and geometry that nature creates seem familiar and predetermined; actually it is random and unpredictable. It is this characteristic that made fractals a convenient method for such studies. Insights into the unpredictability could be a key in understanding the natural random events, like earthquakes, typhoons and other more subtle, natural occurrence like growths and cell developments. Data in different studies have always been thought of as normally distributed; however, recent investigations found that most of these statistical data are non-normal, and in a lot of cases are found to be fractal. According to Padua et al 2013, statistical fractal observations are random observations that possess stochastically self-similar patterns at various scale. In this light, this paper aims to illustrate the pervasiveness of statistical fractal observations in real-life by examining old data sets that used to be modeled in the framework of normal distribution theory. Samples of such observations will be presented in this paper.

Keywords: fractal, self – similar, statistical data

1.0 Introduction

Statistical fractal observations are random observations that possess stochastically self-similar patterns at various scale (Padua et al, 2013). Selvam (2008) illustrated some observations that display those characteristics in real-life such as data sets of Dow Jones Index, human DNA, where in particular, the Takifugu rubripes (puffer fish) DNA are analyzed to show that the space-time data sets are close to model predicted power law distribution. This paper aims to illustrate the pervasiveness of statistical fractal observations in real-life by examining old data sets that used to be modeled in the framework of normal distribution theory.

In the field of Educational Research, test scores are often analyzed through a normal distribution framework: a few large scores, a few small scores and more scores are clustered around a mean value. In reality test scores may not be distributed this way. Depending on the difficulty of the subject, there will be more small scores than larger scores, and a normal assumption about the distribution of these scores can lead to erroneous conclusions. In Social Sciences, income distribution exhibit fractal characteristics where the preponderance of smaller incomes than large incomes appears to be the rule rather than the exception. In a study, in psychology, a psychologist linked the schizophrenic tendency of people to their intake of vitamin C and found the range of intake of the schizophrenic to be from 0.01 μg to 2.18 μg, which compared with the range of intake non-schizophrenic from 0.98 μg to 2.98 μg, with the former having a much smaller dose of vitamin C than the latter. The data obtained from this experiment are decidedly fractals.

Having recognized that, in fact, there are more non-normal data than normal ones in real situations, statisticians tried to transform non-normal data to

[1]Corresponding Fractal Statistics Expert and Consultant
[2]Faculty, College of Engineering, University of San Jose - Recoletos

normal data through the well-known formula of Box-Cox transformation (Johnson and Wichern, 2000):

$$1.)\ Y_i = cx_i^{\alpha}; \qquad -2 \le \alpha \le 2$$

Where by convention, α = 0 is the logarithmic transformation; however, there is no guarantee that an appropriate transformation can be found that will change non-normality to normality. The other route, which had been tried in the past, is to deepen the effort of extreme observation on a statistical measure (e.g. the mean) through the use of robust statistical method Huber (1981), Stigler (1977). Such an approach, however, choose to ignore reality by de-weighing observations that, in fact, occur often. For instance, Stigler (1977) proposed the use of an L – estimator:

$$2.)\ \ldots\ L = \sum_{i=1}^{n} w_i x_{(i)}$$

where w_i's are properly chosen weights of the order statistics $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$.

2.0 Fractal Statistics

According to Padua(2013), for the random variable *X* whose support is non-negative and is scale-invariant probability distribution *f(x)*, that takes the form of

$$3.)\ldots f(x) = \frac{\lambda - 1}{\theta}\left(\frac{x}{\theta}\right)^{-\lambda} \qquad ,\theta < x < \infty,$$

and a power law distribution of

$$4.)\ldots f(x) = \frac{\lambda - 1}{\theta}\left(\frac{x}{\theta}\right)^{-\lambda} \qquad ,\theta < x < \infty,$$

where the exponent of this power distribution corresponds to the fractal distribution of *X*. The corresponding cumulative distribution *X* is shown to be

$$5.)\ F(x) = 1 - \left(\frac{x}{\theta}\right)^{1-\lambda} \qquad ,\theta < x < \infty.$$

In a set of data, where smaller values are more substantial than the normal occurrence, often the Power-law distribution is used. For $x_1, x_2, \ldots x_n$ an iid of *F(x)*, the maximum likelihood estimator of λ is given by

$$6.)\ \ldots\ \hat{\lambda} = 1 + n\left[\sum_{i=1}^{n} \ln\left(\frac{x_i}{\theta}\right)\right]^{-1}$$

$$7.)\ \ \hat{\theta} = \min\{x_1, x_2, \ldots, x_n\},$$

Let $x_1, x_2, \ldots, x_n$ be iid *G(.)* where *G(.)* is an unknown absolutely continuous function. Without loss of generality, assuming that $x_i > 0$ and $x_i > \theta$ for *i* = *1, 2, …., n*. Fitting a fractal distribution $f(x)$ to the quantile of the distribution *G(.)*, where x_α be the α^{th} quantile of *G(.);*

$$8.)\ldots G(x_\alpha) = \int_0^{x_\alpha} g(x)dx = \alpha \qquad (\alpha th\ quantile)$$

$$9.)\ldots\ G(x_\alpha) = F(x_\alpha) = \alpha$$

the following can be obtained;

$$10.)\ \ldots \ddot{e}_\alpha = 1 - \frac{ln(1-\ )}{ln\left(\frac{x_{\acute{a}}}{\grave{e}}\right)} \qquad \text{for all } \alpha \in (0,1)$$

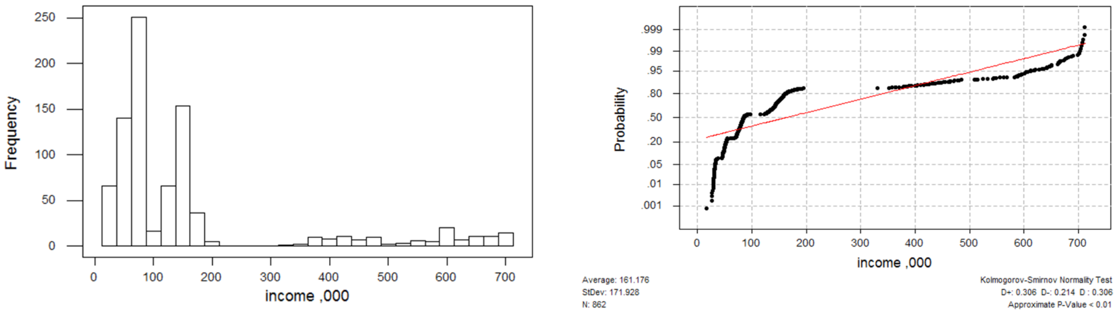
or alternately the following equation can be used;

$$11.)\ \ldots\ \ddot{e}_\alpha = \frac{ln\left(\frac{x_{\acute{a}}}{\grave{e}(1-\acute{a})}\right)}{ln\left(\frac{x_{\acute{a}}}{\grave{e}}\right)} \qquad \text{for all } \alpha \in (0,1)$$

3.0 Sample Data Results

1) Family Income per Year

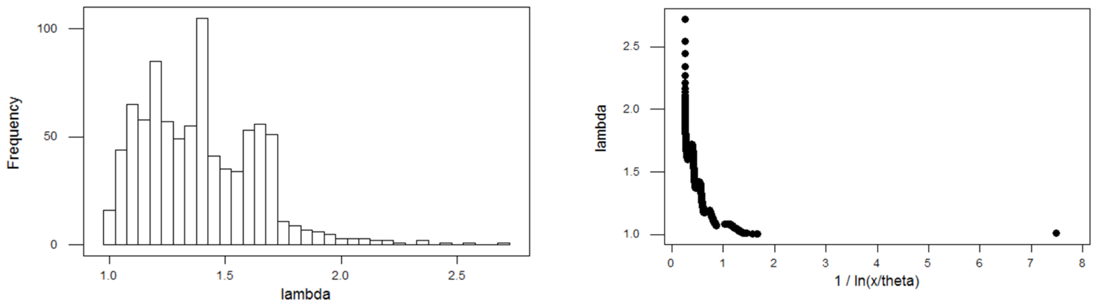
The following is the income distribution of Region VIII Eastern Visayas as surveyed by the National Statistics Office in 2009. The histogram, Figure 1a, showed a highly disproportionate distribution of the yearly income with more low income groups than high income groups. Figure 1b is the Normality test plot, with P-value < 0.01. Such distribution does not reflect any customary characteristics of a normal or other type of distribution.



a.) Histogram **b.) Normality test Plot**

Figure 1. Family Income /Yr, sorted data

Figure 2a, is the histogram of the computed λα, and the scatter-plot showing the multifractal spectrum of the data.



a.) Histogram **b.) Multifractal Spectrum**

Figure 2. Computed λα for Family Income /Yr,

Figure 3a shows clusters of data indicating different fractal dimensions. There are several ways a λ_α-class can be estimated, from (11) the following can be generated in the form of the slope-intercept equation (12), (where; *b is approximated as 0.00*).

$$12.)\ldots \quad \ln\left(\frac{x_\alpha}{\theta\,(1-\alpha)}\right) = \lambda_\alpha \ln\left(\frac{x_\alpha}{\theta}\right) + 0$$

where: λ_α forms the slope of the linear equation

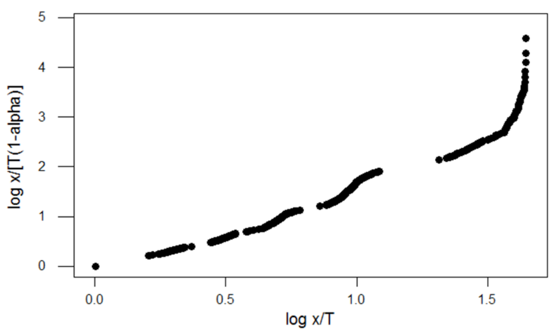
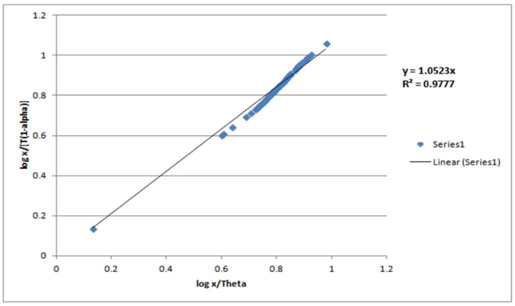


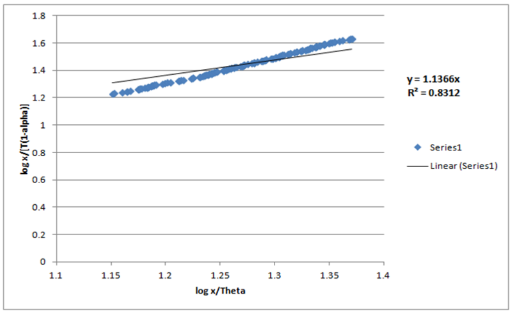
Figure 3. λα Clusters
Scatterplot of logx/[T(1-alpha)] & log x/T

Figure 3, is a scatterplot of the numerator and the denominator of $\boldsymbol{\lambda}_{\alpha}$, on (11), showing the clusters at different locations, and from where data can be segmented by cluster to regress the linear equation of (12). The slope of the regressed line will be the estimated **λ** for the cluster as shown below on Figure 4. The mean, standard deviation and the minimum and maximum λ are descriptive statistical values from each cluster, these are presented to verify the accuracy of the regressed **λ**. For first cluster, figure 4a, the regressed line is ***y = 1.0523x;*** and ***λ = 1.0523.*** The computed descriptive statistical mean for the cluster $\boldsymbol{\lambda}_{mean}$ = ***1.04826*** with ***σ = 0.024612.***



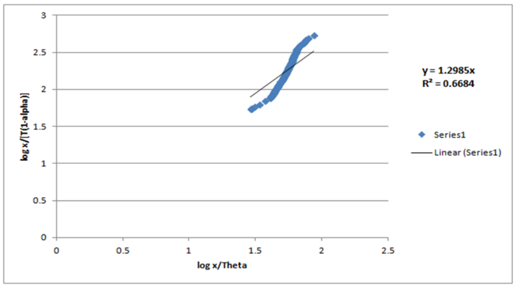
mean:	1.04826	std deviation:	0.024612
min. value:	1.00387	max. value:	1.08466

Figure 4a



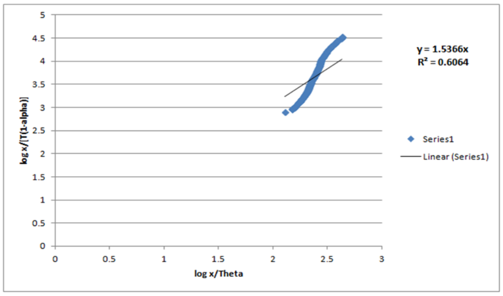
mean:	1.133241	std deviation:	0.036752
min. value:	1.07029	max. value:	1.19486

Figure 4b



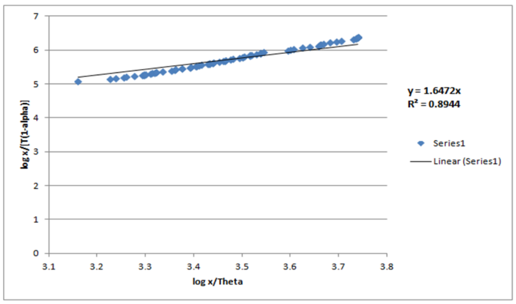
mean:	1.292121	std deviation:	0.075555
min. value:	1.17494	max. value:	1.41933

Figure 4c



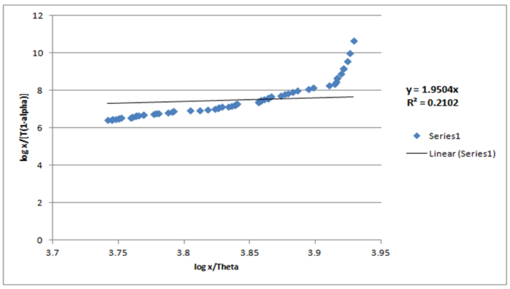
mean:	1.527407	std deviation:	0.10895
min. value:	1.36942	max. value:	1.7242

Figure 4d



mean:	1.644286	std deviation:	0.034452
min. value:	1.59579	max. value:	1.71043

Figure 4e



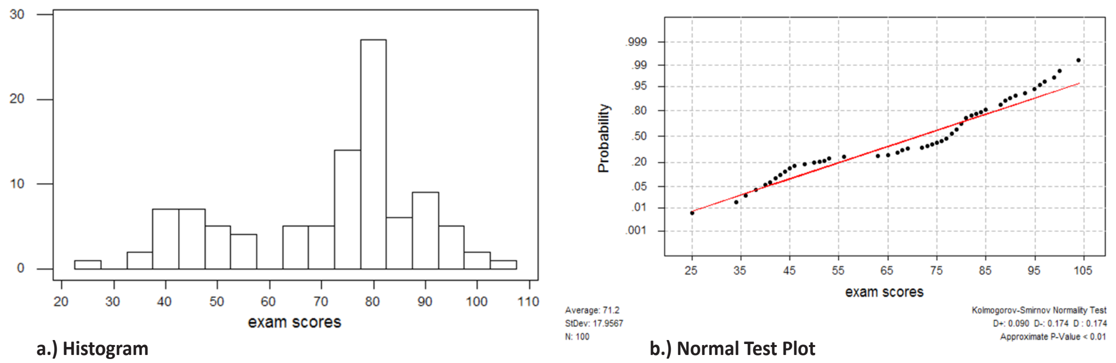
mean:	1.944344	std deviation:	0.218662
min. value:	1.71672	max. value:	2.72039

Figure 4f

2) Algebra Scores

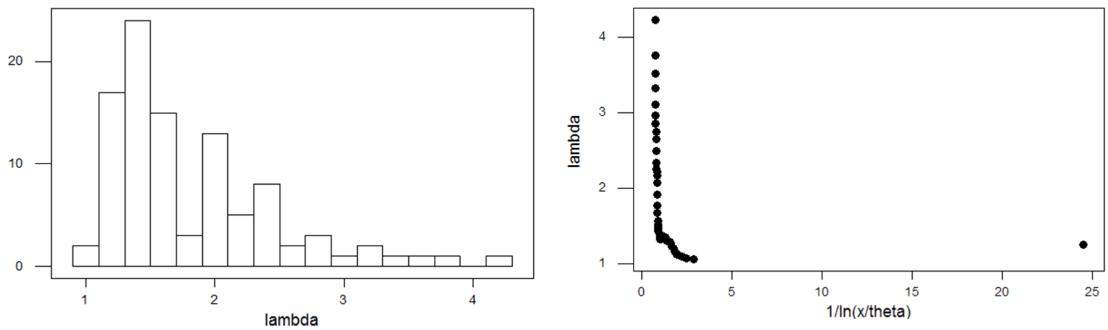
Algebra is one of the important subjects for the engineering and sciences. Often the subject is also a major hurdle that must be overcome by most students who choose such courses. Students performance in this subject as per observations are found not to follow the normal distribution, this is also true to other relatively difficult subjects. In this example, there are more low scores than high scores. Below is the distribution of test scores given to 100 students.

Figure 5a shows that the data is not normally distributed and confirmed by figure 5b. Figure 6 shows the corresponding λ distribution and the multifractal spectrum, indicating the multifractal nature of the data.



a.) Histogram b.) Normal Test Plot

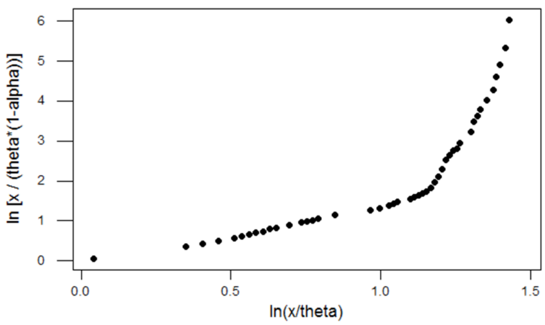
Figure 5. Exam Scores for Algebra



a.) Histogram b.) Scatter-Plot of λ and the scale

Figure 6. Computed λα and the Multifractal Spectrum

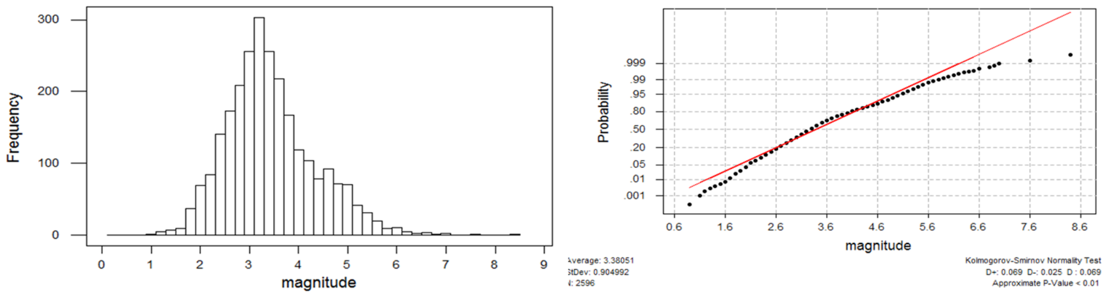
Figure 7. λα Fractal Dimension Clusters



3) Earthquakes Magnitude and the Inter-event Times

The following data are the magnitudes and the inter-event times of earthquakes that happen around the Philippine peninsula between the dates January 1, 2011 and April 23 2013. The magnitude refers to the intensity of the earthquakes while inter-event time is the calculated time interval between the succeeding occurrence. Figure 8a, is the magnitude distribution, and figure 8b, the normality test plot. Figure 9b is the multifractal spectrum of the λ.

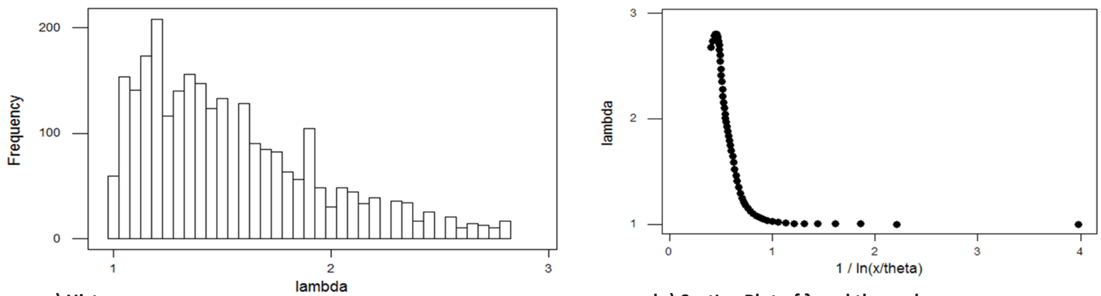
Figure 10, shows the distribution of the inter-event time and the normality test of data, showing more of an exponential distribution rather the normal.



a.) Histogram

b.) Normality Test Plot of the Magnitude

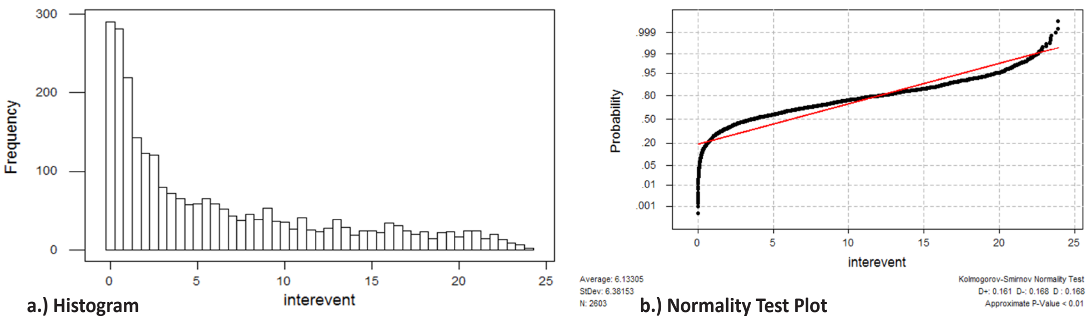
Figure 8. Distribution of the Magnitude and the Normality Test Plot



a.) Histogram

b.) Scatter-Plot of λ and the scale

Figure 9. Distribution of the fractal dimension of the magnitude and the Multifractal Spectrum

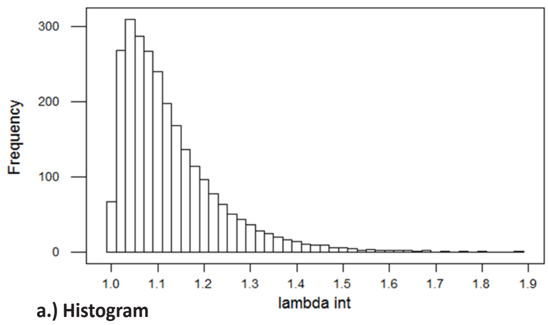


a.) Histogram

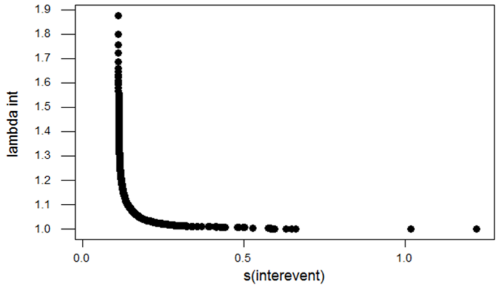
b.) Normality Test Plot

Figure 10. Distribution of the Inter-event time and the Normality Test Plot

Figure 11b is the multifractal spectrum of the inter-event fractal dimension λ.



a.) Histogram



b.) Scatter-Plot of λ

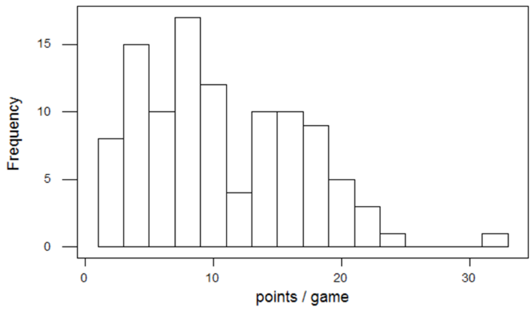
Figure 11. Computed λα and the Multifractal Spectrum of the Inter-event Time

4) Points per Game of 105 NBA players

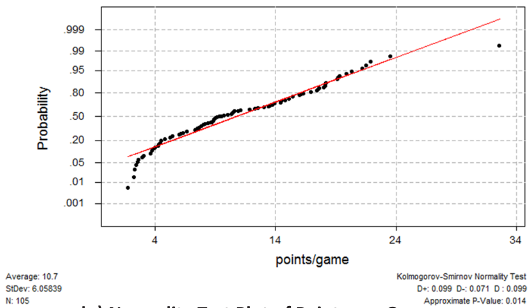
The game of basketball is one sport that highly utilized statistics to determine the performance of the players in every game. The example below is an example of such statistical data which when tested is multifractal, figure 12 and figure 13.

5) Grain Yield

The following example is the data of yield of grain of a farm with 125 plots. Figure 14 is the distribution and the normality test plot. In Figure 15, the λ distribution and the multifractal spectrum showing the multifractal nature of the data.

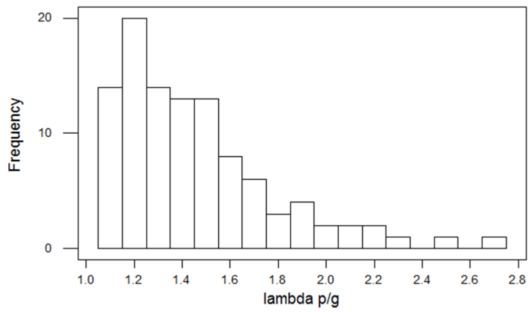


a.) Histogram

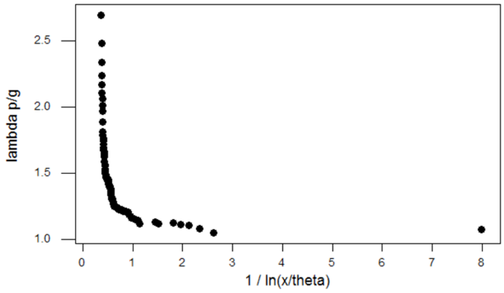


b.) Normality Test Plot of Points per Game

Figure 12. Distribution of the Points per Game and the Normality Test Plot

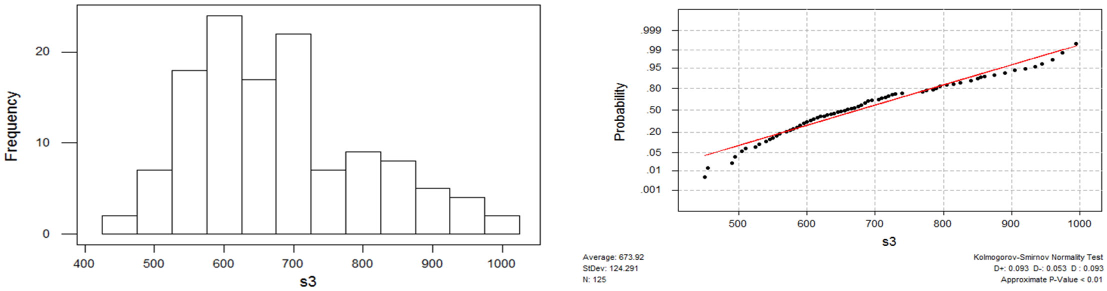


a.) Histogram



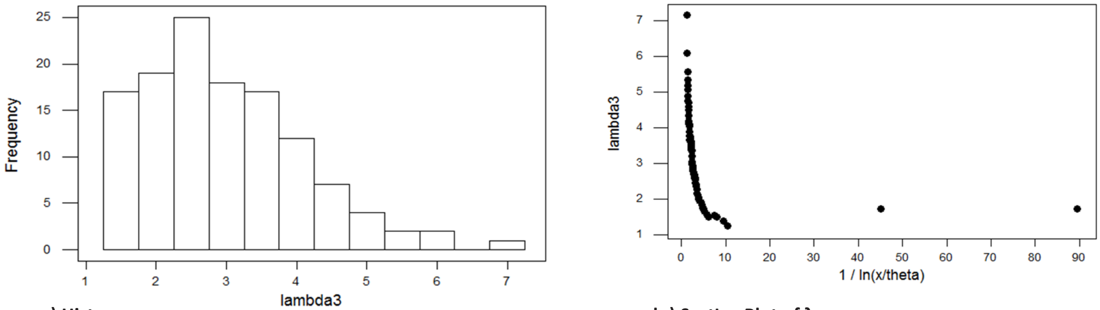
b.) Scatter-Plot of λ

Figure 13. Computed λα and the Multifractal Spectrum



a.) Histogram **b.) Normality Test Plot of Grain Yield**

Figure 14. Distribution of the Grain Yield and the Normality Test Plot

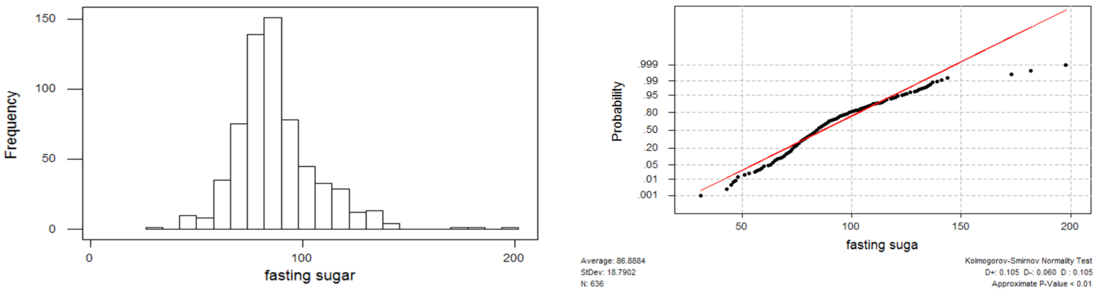


a.) Histogram **b.) Scatter-Plot of λ**

Figure 15. Computed λα and the Multifractal Spectrum of Grain Yield

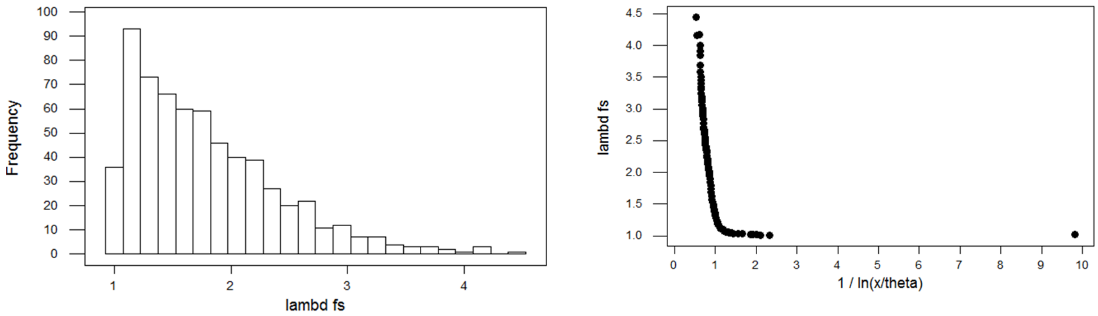
3.6) Fasting Blood Sugars

Fasting blood sugar (FBS) is a measure of the blood glucose. It is often the first test done to check for pre-diabetes and diabetes. (Healthwise Inc., 2012). The succeeding data were recorded



a.) Histogram **b.) Normality Test Plot of FBS**

Figure 16. Distribution of Fasting Blood Sugar and the Normality Test Plot



a.) Histogram b.) Scatter-Plot of λ

Figure 17. Computed λα and the Multifractal Spectrum of FBS

4.0 Discussion

Data in various aspects of the real world: income, test scores, seismic occurrences, sports and agricultural gains, appear to behave like fractal observations. These did not follow the expected normal distribution pattern. In fact, in such real-world data, the persistence of low values over large values is the rule rather than the exception.

Recognizing the fractality of most real-world data, as opposed to data normality has far-reaching implications. First, fractal analysis, acknowledge the inherent irregularities and fluctuations of real data at lower scales. The reliance on the means as an "average" data behavior is not grounded on reality. Second, irregularities occur regularly at every scale (self – similarity) so that what happens at a larger scale also happen at smaller scales.

The data above exhibited fractal behavior with multiple fractal dimensions, Padua et al., 2013, Test Algorithm for fractals. These also demonstrated the innate occurrence of fractals among the many datasets that may often be taken as a normal distribution. The varied sources of the samples, illustrate the pervasiveness of the fractals in the real-world, in the natural events, agricultural produce, medical, social and economics.

5.0 Conclusion

Data at a glimpse often times are simply treated to be normal, and may often times have caused imprudent judgment. The above data sets, allowed us a peek of the fractal nature of real-world data. The pervasiveness of fractal data in the real-world is valid. Data sets where smaller figures are more, than the larger ones are most likely to be fractal. Fractals will also exhibit self-similar characteristics. Multifractal dimension happen when statistically, several fractal dimensions exists within a scale.

References

Huber, P. J. (1981) Robust Statistics. New York: John Wiley and Sons.

Johnson, R. & Wichern W. (2000) Applied Multivariate Statistics, Wiley Series: New York

Mandelbrot, B. (Ed.), The Fractal Geometry of Nature, Freeman, New York, 1983, 14-57

Stigler, S.M (1977). Do Robust Estimators Work with Real Data?,The Annals of Statistics, 5(6), 1055-1098.

Box, G.E.P., & Cox, D.R. (1964). An Analysis of Transformations, Journal of the Royal Statistical Society, B, 26, 211-246.

Padua, R., Adanza, J.G. & Mirasol, J.M. (2013). Fractal Statistical Inference. The Threshold, CHED-JAS Category A Journal, 36- 44.

Padua, R., Barabat, E. & Regalado, D.Y. (2013). A Simple Test for Mono and Multifractality of Statistical Observations, NORSU Prism Journal of Higher Education, CHED-JAS Category B, 23-24.

Family Income by Class. National Statistic Office. (Retrieved February 22, 2013 from http://census.gov. ph/sites/ default/ files)

Philippine Seismic Data: 2011 – 2013. (Retrieved May 9, 2013, from http://www.phivolcs.dost.gov.ph/html/update_SOEPD/EQLatest.html)

Selvam, A.M. (2008), Fractal Fluctuations and Statistical Normal Distribution.

Six Annual Fasting and One Hour Post Glucose Blood Sugars. (Retrieved April 16, 2013 from http://lib.stat.cmu.edu/datasets/Andrews)

Yield of Grain from Each of 1,500 Fifteen-foot Rows of Wheat. (Retrieved April 16, 2013 from http://lib.stat.cmu.edu/datasets/Andrews)

Similarity Index: 3%
Grammarly: 95%