International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 15 (2018) pp. 11942-11946

Analysing The Homogeneity of Means of Data Distributed Over Best Suited Model Using Anom

Mrs. K.Sowmya[1] and Dr. R.Satya Prasad[2]

[1]Research Scholar,

[2]Professor, Department of Computer Science and Engineering,
Acharya Nagarjuna University, Guntur, Andhra Pradesh, India.

Abstract

Data can be analyzed by observing the homogeneity of means with the aid of well-known technique –Analysis of Means (ANOM) that decides that the given live data identified to follow the best suitable model out of proposed competing models. The competing models are tested for their suitability for each dataset depending on the structure of dataset using a more rational technique based on Correlation coefficient rather than graphical technique namely Quantile-Quantile (QQ) Plot. The best identified model and the corresponding dataset are then paired to develop the ANOM calculation procedure using R.

Keywords: ANalysis Of Means, subgrouping, Maximum Likliehood Estimation (MLE), Half Logistic Disstribution (HLD), Correlation, Reliabilit, Visualization of Data

INTRODUCTION

Data plays a vital role in understanding the nature of the domain in which it is generated. Data generated is said to be different types based on its nature like descriptive, inferential, grouped, ungrouped etc. We can broadly categorize into quantitative and qualitative. Qualitative data emboss the nature and help in understanding behaviour of the process which generated the data. Generally Qualitative data when is in the form the counting or capturing the measure during a period we consider it as grouped data. All such data are best analysed using Poisson Models. Poisson Process models are again categorized into homogeneous and non- homogeneous based on the interval during which the data is captured. The time intervals in between the event occurrences play a vital role in assimilating and visualizing the data [1].

Data analysis is critical in portraying the revelations from various origins of data. It acts like a filter in acquiring meaningful insights out of huge datasets. Analysing the data using statistical approach aids in reaching the research conclusion removing the human bias factor.

To inspect the data for its insight and behaviour various techniques are available namely central tendencies, dispersion measures, analysis of means & variances, distribution models, etc. [1]. Data diagnosis is carried out by various techniques but usage of distribution models has a greater impact and gives a truthful insight on data

In this paper, the main focus is to identify the nature and behaviour of means of the data using Analysis of Means (ANOM) over the selected model. The following sections

The following sections are portrayed to detail the distribution model in consideration for the data - Half Logistic Distribution (HLD) model is considered for analysing data.The unknown parameters of the models are estimated using Maximum Likelihood Estimation (MLE) process in Section II.

In section III the justification of the selected model is further investigated to clinch to the best suited model by using correlation coefficient instead of traditional graphhical approach of QQ-Plots. In section IV the model is applied and data is deeply inspected to attain a detail picture on the homogeneity of means using ANOM under various measures. All the process mentioned is coded and implemented in statistical programming language R.

Section V is presented with the experimentation datasets, results of MLE, correlaton coefficient comparisions, ANOM subgrouping and corresponding graphs and the detailed analysis of the obtained results and the conclusion follows in section VI.

HLD PARAMETER ESTIMATION USING MLE PROCESS.

Assessment of parameters is very influential in predicting the software reliability. Upon concluding the analytical solution for the mean value function m(t) for the specific model, the MLE technique is enforced for attaining the parameter estimation. The crucial intention of Maximum Likelihood parameter Estimation is to resolve the parameters that magnify the probability of the fragment data. The MLE is deliberated as vigorous, robustious and mathematically fierce. They yield estimators with good statistical factors. In the

outline analysis, MLE methods are resilient, versatile and can be employed to distinct models and data categories [1][2][3][5]. Accomplishing to present day's computer capability, the mathematical intensity is not a considerable hurdle.

The constants 'a', 'b' surfacing in the mean value function also appear in NHPP, through the intensity function to materialize error detection rate and in various other expressions are treated as parameters of the model. To assess the software reliability, the unknown parameters 'a' and' b' are to be treasured and they are to be predicted using the software fragment data [6].

For a detail, let 'n' be the time instances where the first, second, third..., kth faults in the software are encountered. We can consolidate it as, if T_k is the total time to the kth failure, 'tk' is an observation of random variable Tk and 'n' such similar failures are successively recorded. The combined probability of such failure time gaps $t_1, t_2, \ldots, t_n$ is given by the Likelihood function as

$$L = e^{-m(t_n)} . \prod_{K=1}^{n} m'(t_k) \quad (1)$$

The logarithmic application on the equation (1) would result a log likelihood function and is given in equation (2).

$$LogL = \sum_{i=1}^{k} \lambda(t_i) - m(t_k) \quad (2)$$

The mean value function m(t) and intensity function $\lambda(t)$ of HLD is given as

$$m(t) = a\frac{(1-e^{-bt})}{(1+e^{-bt})}$$

$$\lambda(t) = \frac{2abe^{-bt}}{(1+e^{-bt})^2}$$

The Maximum Likelihood Estimators (MLEs) are featured to maximize L and estimate the values of 'a' and 'b'. The process to maximize is by applying partial derivation with respective to the unknown variables and equate to zero to obtain a close form for the required variable. If the closed form is not destined, then the variable can be estimated using Newton Raphson Method. Subsequently 'a' and 'b' would be solutions of the equations. The section proceeds with Half Logistic Distribution (HLD) model.

$$\frac{\partial \log L}{\partial a} = 0 , \frac{\partial \log L}{\partial b} = 0 , \frac{\partial^2 \log L}{\partial b^2} = 0$$

MODEL SELECTION USING CORRELATION COEFFICIENT

The qq-plot is a graphical way of approach for determining whether the given two datasets follow common distribution. The quantiles from the first dataset are treated as ideal quantiles. The other dataset in question is marked in accordance to the same distribution which is used for the first ones. If both the quantiles belong to the same distribution, then the lines plotted will be same for the ideal case and a maximal linear closeness otherwise. Here to decide the distribution which best suits the data, multiple qq-plots must be plotted and these qq-plots now act as probability plots. We replace the first dataset with the quantiles of theoretical distribution. The graphical plotting is visually good and clearly gives the details to decide whether the distribution model best suits or not when the linear difference to distribution is considerably visually clear and identifiable. However there are possible cases where a data is best suited to (say) two distributions and the plots are considerably same but the difference is not visually recognizable.

Correlation Factor can be used as an alternative approach instead of plotting the quantiles.The qq-plot requires the theoretical quantiles along with the distribution quantiles and the quantiles need to be plotted and there is no quantified measure on the plot to decide the best suited distribution, only the visual observation is the deciding factor. The process of using correlation factor not only gives a quantified measure, but also resolves the possible tie in identifying one of many seemingly equally close distributions for a dataset through qq-plot [4].

To use the correlation factor, the inputs required are same as that of qq-plots, i.e., distribution and theoretical quantiles. But the burden of plotting is overcomed and quantified measure is achieved. The quantified measures achieved for different models in question are compared and the solution is easily obtained [4].

The following are common to determine best fit model through qq-plots or through correlation factor.

Consider the data for which the model must be decided and check whether there are any missing values; if so fill them with appropriate mean values. Then list the models in question and repeat the following steps for each model over the same data in focus. Firstly estimate the unknown parameters of the considered grouped data using Maximum Likelihood Estimation (MLE) method, and then generate theoretical quantiles and distribution quantiles.

Once both the quantiles are determined, the qq-plot considers plotting of the quantiles and obtains graphical measure whereas the correlation factor process determines the correlation between the quantiles and thus obtains the quantified measure.

APPLYING ANOM OVER SELECTED MODEL DATA DISTRIBUTION

The ANOM is the best graphical statistical techniques for comparing group of treatment means.

Statistically significant divergent groups form the comprehensive groups are noticed using ANOM methodology by contrasting with the mean of each group to overall. Apart from comparison of group means it can be used to compare rates, proportions and variances. Here this prime extends Analysis to median, extreme values, mid- range, range of subgroups using ANOM[8].In the narration of the section, after observing the tradition of various earlier researchers in ANOM, the word group and sub group are used in the same sense without any technical difference between them.

The prime lime lights analysis of the data using subgrouping of ANOM over the data to be visualized and is distributed over HLD model. The HLD model is opted and maximum likelihood estimation mechanism is enforced to gauge the exotic factors of HLD. The data which is to be analysed is considered and fitted to the model. The distribution which is generated from the model is divided into groups of equal size say 'gs', this 'gs' is optimal if it ranges between 2 to 10, $2 \leq gs \leq 10$.

Say n groups of each size 'gs' are obtained. The groups are numbered and are treated as group number. For each group all the different measures are calculated. These measures involve mean, median, extreme values maximum, minimum, range, midrange for each subgroup.

Data to be analyzed is first applied on the best suited -model which generates a distribution. The obtained distribution is divided into groups of equal size. These groups are given numbering and every group is processed to obtain group mean, median, range, mid-range, extreme values- maximum and minimum.

Each subgroup is processed to obtain subgroup mean by obtaining the mean of all the distribution data in the subgroup. Each subgroup is sorted and median is obtained and this process is repeated for every subgroup. For every subgroup extreme values - maximum and minimum values are noted and thus range (maximum-minimum) of the subgroup is obtained, using these three maximum, minimum and range, mid-range of each subgroup is obtained.

The mean, median, extreme values – minimum & maximum, range and mid-range for all subgroups are framed along with their group number.

The framed data measure is each considered separately and sorted. The limits are obtained by using ANOM measure in other words three sigma standard limits. This gives the limit probability. The probability is used and applied on sorted framed data measure, to obtain the lower limit and upper limit lines. The average of the limits is considered as central line.

The unsorted framed data of the measure for which the limits are calculated is taken and a scatter plot is plotted with the lower, upper and central limit lines. This gives scatter points equal to the number of groups on the plot, each point indicates a subgroup. The subgroups which are below and above the lower and upper limits respectively are most vulnerable to failure.

This process is repeated for each measure (mean, median, extreme values-minimum & maximum, range, mid-range) for subgroup size 5. Further the process is repeated for each measure for all subgroup sizes ranging from 2 to 10.

RESULTS , GRAPHS& ANALYSIS OF RESULTS

The process can be applied on dutiable data minuting the comprehensive visualization of the data in all dimensions. The process is carried out on analysis of the network intrusions and its visualization over all possible measures.

The intrusion data is first filtered out biased on connection type. The parameters which showcase the intrusions are observed, processed to depurate the data. The obtained data is staged for the HLD model is quantified and justified with the correlation factor. The obtained means are grouped, analysed and visualized using ANOM.

The data in Table 1 contains time interval between intrusions on network serviced using transmission control protocol (TCP) connection. HLD stands as the best suited model since the correlation factor seems to be the maximum. The HLD is now applied on the data distribution for ANOM and subgroups are obtained for mean, median, range, mid-range and extreme values and the groups can be analysed for homogeneity.

These computations initiated with the seed value of 0.1 for Newton Raphson method. The comparative study focuses on unknown exotic parameters of the distribution through MLE process and then obtaining correlation factor with the distribution quantiles and data to compare and pinpoint the best suited model i.e; HLD.

Table 1 showing the intrusion data for processing.

Duration	IntrusionTime	ServiceType
2.954396	00:00:01	tcp
2.8906	00:00:03	tcp
2.9755	00:00:03	tcp
2.991582	00:00:04	tcp
2.864921	00:00:05	tcp
2.841049	00:00:09	tcp
3.019667	00:00:10	tcp
3.054436	00:00:12	tcp
3.017788	00:00:13	tcp
2.935167	00:00:17	tcp
2.966738	00:00:21	tcp
2.998744	00:00:24	tcp
2.980539	00:00:25	tcp
1.233717	00:00:28	tcp
3.009168	00:00:29	tcp
1.004381	00:00:31	tcp
1.168256	00:00:33	tcp
2.962536	00:00:34	tcp
2.966704	00:00:34	tcp
2.967952	00:00:34	tcp
3.015965	00:00:48	tcp
0.994583	00:00:51	tcp

Showing 1 to 23 of 78,970 entries

The required data fields is tailored according to the needs of the analysis which made the raw data for processing as shown in Table 1. The obtained data is only a single day network data. The data contains the duration of the connection, the time of the day at which the intrusion occurred and type of the connection here the connection type in focus is TCP.

After the applying the process discussed earlier, the analysis of various measures are obtained. The means are plotted over a scatter plot. The plot on close observation shows they are intrusions which are very seriously to be considered since they are placed out of the decision lines. By inspecting the visualization of the scatter points it gives the clear picture that the point many times are cohesioning indicating multiple intrusions are occurring at the given time in the same connection which is again a seriously considerable phenomenon. By grasping the distance between the scatter points from its neighbours, the intrusions are continuous and are following a pattern. All these observations give a visualization on the intrusions of the network which otherwise couldn't be that simple.

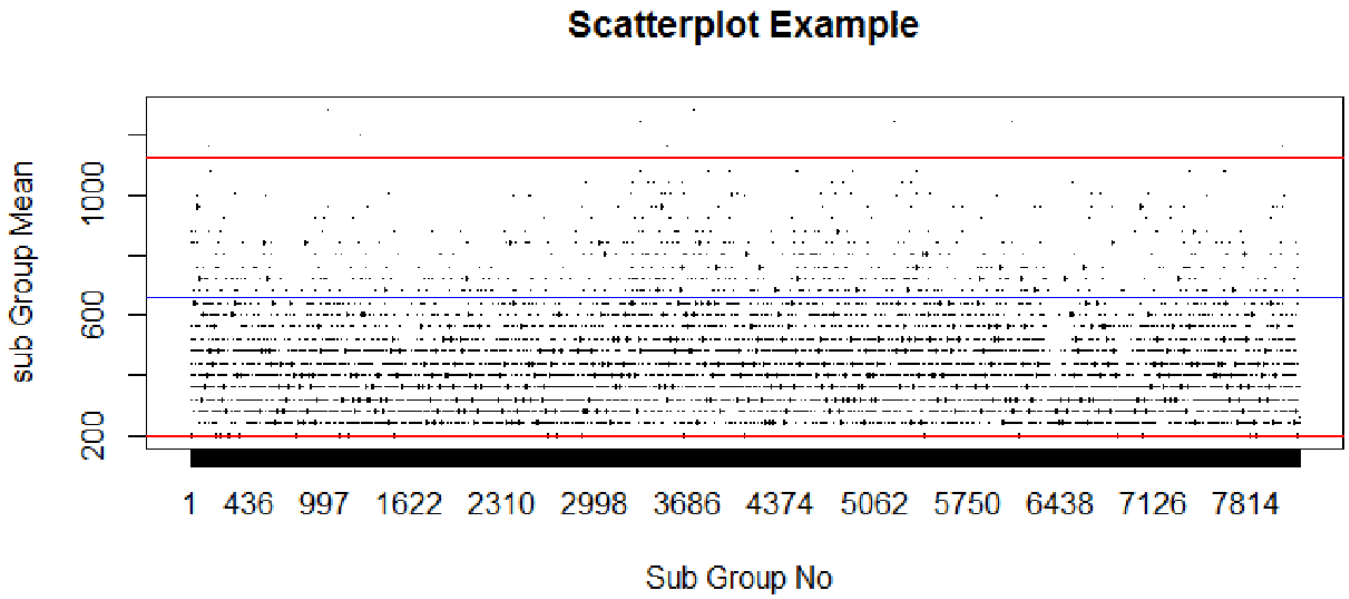

Figure 1. Scatter plot of subgroup means of analysis of intrusion data using ANOM and distributed over HLD model

CONCLUSION

This paper suggests R implementation for the following data analysis exemplified with the help of two competition software reliability models.

The best suitability of the model for a given failure data among various competitors through QQ-plot.

The non-resolvability of a best model to a failure data is addressed through a more admissible procedure called maximum correlation coefficient.

The data when categorized into a number of groups/subgroups of the same size, the group characteristics like mean, median, range, mid-range, extreme values-maximum and minimum are made useful to identify homogeneous/heterogeneous characteristics called analysis of means (ANOM).

For a better understanding, we have considered the well-known Goel-Okumoto and Half Logistic Distribution models based SRGM. Users can use any finite failure SRGM borrowing the respective mathematical expression changes or revisions.

REFERENCES

[1] H. Pham, System software reliability, Springer, 2006.

[2] R Satya Prasad, K Ramchand H Rao and R.R. L Kantham, “Software Reliability Measuring using Modified Maximum Likelihood Estimation and SPC” International Journal of Computer Applications, vol-21, Number 7, pp. 1-5 Article1, May 2011.

[3] R.satyaprasad, “Half Logistic Software Reliability Growth Model”, Ph.D. Thesis, 2007, http://hdl.handle.net/10603/126989

[4] Donald J. Wheeler, “Advanced Topics in Statistical Process Control: The Power of Shewhart's Charts” Handbook. Statistical Process Controls, Inc., SPC Press, Knoxville, TN, 2004

[5] R. Satya Prasad, K Sowmya and R Mahesh, “Monitoring Software Failure Process using Half Logistic Distribution” International Journal of Computer Applications, vol-145(4), pp.1-8, July 2016

[6] A.L.Goel and K. Okumoto, "Time-Dependent Error-Detection Rate Model for Software Reliability and Other Performance Measures". IEEE Transl. on Reliability vol-R–28 Issue. 3 pp. 206– 211, Aug, 1979.

[7] Kyoto Data, (2015). “Benchmark Data Description v5”. http://nsl.cs.unb.ca/NSL-KDD