# COUNT DATA MODELLING OF INTERNET TRAFFIC

**INDEPENDENT STUDY PROJECT REPORT (CS-010)**

**by**

**REHAM MUZZAMIL**

**Department of Computer and Information Systems Engineering**

**NED University of Engineering & Technology**
**Karachi-75270**

# COUNT DATA MODELLING OF INTERNET TRAFFIC

**INDEPENDENT STUDY PROJECT REPORT (CS-010)**

**by**

**REHAM MUZZAMIL**

**CS-10/2018-19**

**Supervisor:**
**Dr. Muhammad Asad Arfeen**

**Assistant Professor & DIT**

**Department of Computer and Information Systems Engineering**

**NED University of Engineering & Technology**
**Karachi-75270**

# ABSTRACT

*A count model is based on a discrete statistical distribution which models the probabilistic behaviour of a number of events in a fixed interval of time and can be used to predict the behaviour of a real traffic stream and thus, can be used to evaluate performance of networks. In this study, the aim is to analyze count data modelling of Internet traffic using multiple discrete distributions and provide a comparative study. Network traffic models have evolved significantly over the lifetime of the Internet. The earliest models were largely Poisson-based, designed for ease of analysis. The discovery of traffic characteristics like the presence of over dispersion and under dispersion and the presence of outliers in the count data required significant changes to traffic models. In this study, we have examined this evolution from the early models to the modern models on the discrete distributions that account for complexities and properties like over dispersion and under dispersion. We have studied the traffic behaviour on the real data sets and presented a statistical analysis and best fitted distribution model of internet traffic, the data used for the modelling is synthetic due to the limitations of Memory and processing time. The models are evaluated in R Software. Maximum Likelihood Estimator (MLE) technique is used to identify the maximum MLE log-likelihood which characterized as best fitted distribution. Poisson, Negative Binomial, Weibull and Mittag Leffler fitted on multiple datasets are presented. Among the four distributions,  Mittag Leffler is identified as the best traffic characteristic based on MLE maximum log-likelihood. .These results are valuable on modeling future tele-traffic engineering algorithm like policing, shaping, scheduling or queue which is based in real IP-based campus network environment. It is also useful for future prediction of tele-traffic models.*

# ACKNOWLEDGEMENTS

*Starting with the name of Almighty Allah, His mercy and grace. Without the will of Allah this Independent Study Project would not have been possible. I am thankful to Almighty God for blessing me with the strength and courage to endure all the challenges in the conduct of this research.*

*I would like to thank Dr. Muhammad Asad Arfeen for being my mentor and introducing me to this field of research. He has always provided me his valuable suggestions during the research and development of this project. This research has been carried out just because of his motivation and encouragement. Without his constant support and assistance this project would not have been possible.*

*I am grateful to my parents and sister who were always there with me through thick and thin. I would also like to extend gratitude to the community of CAIDA for providing me Internet Traffic Traces.*

*Finally, I would like to express my regards to the research community abroad who are working extensively in the areas of simulation modeling and analysis, count data models and their applications in multiple domains. Kudos to all of you!*

# TABLE OF CONTENTS

# CHAPTER 1

# INTRODUCTION

## 1.1. BACKGROUND STUDY OF THE RESEARCH

For several past years researchers have been looking for a stochastic process which could be used as an accurate and simple model for Internet traffic. A traffic model is a stochastic process which can be used to predict the behaviour of a real traffic stream. Ideally, the traffic model should accurately represent all of the relevant statistical properties of the original traffic, but such a model may become overly complex. A major application of traffic models is in predicting the behaviour of the traffic as it passes through a network.

Network traffic modeling is used as the basis for the design of network applications and for capacity planning of networking systems. Given the impact of poor choices in this arena, it is clear that the validity of the underlying models is of critical importance. The factors used to evaluate a system are taken directly from the underlying traffic model. This use requires that traffic models be both valid, resembling reality closely, and sufficiently simple as to allow queuing analysis models to reach a steady-state.

Understanding the underlying traffic behavior of backbone and edge networks is vital for traffic engineering tasks such as link capacity planning, traffic classification, and anomaly detection. Traffic characterization is typically addressed through statistical analysis of individual link(s) and network-wide traffic volume properties such as counts of bytes and packets, as well as by analyzing the distributional behavior of particular packet header fields.

Thus, Modelling a count variable (the number of events occurring in a given time interval) is a common task to predict the behavior of the traffic stream and to evaluate the performance of the network. The widespread popularity of the Poisson model for count data arises but it fails when the data possess overdispersion and underdispersion.

## 1.2. OVERVIEW

Understanding the underlying traffic behavior of backbone and edge networks is vital for traffic engineering tasks such as link capacity planning, traffic classification, and anomaly detection. Traffic characterization is typically addressed through statistical analysis of individual link(s) and network-wide traffic volume properties such as counts of bytes and packets, as well as by analyzing the distributional behavior of particular packet header fields. Modelling a count variable (the number of events occurring in a given time interval) is a common task to predict the behavior of the traffic stream and to evaluate the performance of the network. The widespread popularity of the Poisson model for count data arises, in part, from its derivation as the number of arrivals in a given time period assuming exponentially distributed interarrival times. But of the thousands of other count models that have been developed over the years (see Wimmer and Altmann 1999 for an excellent synthesis), very few share this straightforward connection between a count model and its timing model equivalent. The standard method of count data modeling is Poisson distribution, which has the assumption of equidispersion, as identified by the same mean and variance values. The modelling of count data frequently causes the emergence of over-dispersion which has a higher variance than mean itself. Many research could be found especially for handling over-dispersion problem such Negative Binomial, Zero Inflated Poisson and Quasi approach. However, a few method in research could be fitted under-dispersion problem, such as Generalized Poisson which able to handle both problems, but has limited range of under-dispersion values. Count data models allow for regression-type analyses when the dependent variable of interest is a numerical count. They can be used to estimate the effect of a policy intervention either on the average rate or on the probability of no event, a single event, or multiple events. The effect can, for example, be identified from a comparison of treatment and non-treatment units while adjusting for confounding variables, or from a difference in-differences comparison, where the effect of the policy is deduced from comparing the pre-post change in the outcome distribution for a treatment group with the pre-post change for a control group.

## 1.3. CORE OBJECTIVE OF THE STUDY

The objective of this study is to discuss the modelling, estimation and testing of count data models from the viewpoint of an application in the domain of traffic engineering and inspect mathematical stochastic models on Internet Traffic. The idea is to study the characteristics possess in the Internet Traffic traces taken from the website of CAIDA and map corresponding count data. Model count data on to multiple discrete distributions and analyse the results to figure out the best possible fit offered by any distribution.

## 1.4. METHODOLOGY FOR ANALYSIS:

The data traces are taken from the CAIDA – The Internet Traffic Archive. Count Data of trace files are obtained using LibTrace library, the count data is than undergone though process of finding the underlying pattern in them and their characteristics by plotting their time aggregation plots. Moreover, the actual simulation and estimation is performed on the synthetic dataset due to the limitation of memory. The dataset are generated using multiple approaches and their fitted distributions behaviors are examined. Model selection and evaluation is done on the basis of AIC, BIC and log-likelihood parameters values. The methodology is discussed in detail in chapter six.

## 1.5. ORGANIZATION OF THE ISP:

Chapter one would be discussing the background study of the research, overview of the research's topic, core objective and the methodology that would be perused in the thesis would be discussed. Further, chapter two is about Literature Review regarding count data modelling and different distributions. Chapter three includes details of count models, what is a count variable and why count models should be used. Later in this chapter different count models are discussed. Time Aggregation plots are also presented here. Another important aspect of index of dispersion is also described inside this chapter. Chapter four discusses about Parameter Estimation and different technique available for the Parameter Estimation. The technique of Maximum Likelihood Estimation is discussed in detail in this chapter. Chapter five describes the existing Internet Traffic Archives platforms. How trace files can be converted into a count data using LibTrace is also a part of this chapter. Modelling and simulation is performed in R, therefore main packages of R is discussed here. Chapter six discusses the details of the practical work done in R and the results of simulations are discussed. Furthermore, tabular and graphical representation of the comparative study is shown here. Last chapter holds the concluding remarks, limitations and gaps faced while carrying out this research and the future enhancements.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1. LITERATURE REVIEW

The design of robust and reliable networks and network services is becoming increasingly difficult in today's world. The only path to achieve this goal is to develop a detailed understanding of the traffic characteristics of the network. Analysis of the traffic provides information like the average load, the bandwidth requirements for different applications, and numerous other details. Traffic models enables network designers to make assumptions about the networks being designed based on past experience and also enable prediction of performance for future requirements. Traffic models are used in two fundamental ways:

(1) As part of an analytical model or

(2) To drive a Discrete Event Simulation (DES).

In case of Internet traffic modelling, two aspects of a **count model** are important for consideration. Namely, the required length of time interval for recording counts, and the probability mass which a model assigns to the quantiles in body and tail (extreme values) parts of the distribution.

**(*Black McShane 1996*)** introduced a generalized model for count data based upon an assumed Weibull interarrival process that nests the Poisson and negative binomial models as special cases. The computational intractability is overcome by deriving the Weibull count model using a polynomial expansion which then allows for closed-form inference (integration term-by-term) when incorporating heterogeneity due to the conjugacy of the expansion and a commonly employed gamma distribution. In addition, we demonstrate that this new Weibull count model can model both over- and underdispersed count data. Weibull regression models provide the best fits, that is, a slight improvement in log-likelihood for the Weibull model without heterogeneity and a significant improvement for the heterogeneous Weibull model, compared to the Poisson and Winkelmann's gamma count model.

**(*Alexander Kasyoki Muoka 2016*)** Statistical simulation technique was used to compare the performance of these count data models. Count data sets with different proportions of zero were simulated. Akaike Information Criterion (AIC) was used in the simulation study to compare how well several count data models fit the simulated datasets. From the results of the study it was concluded that negative binomial model fits better to over-dispersed data which has below 0.3 proportion of zeros and that hurdle model performs better in data with 0.3 and above proportion of zero.

***(Ver Hoef and Boveng (2007)*** made a comparison between quasi-Poisson and negative binomial regressions as two contrasting approaches for dealing with overdispersed count data in ecology. With an example on harbor seal data they showed that the choice of approach can affect the outcome of the analysis. The authors recommended sound scientific reasoning and graphical investigation of the data as the basis for model choice. Yet, different processes underlying overdispersion in ecological data and resulting in various mean–variance relationships have not been thoroughly investigated.

***(Asad Arfeen 2013)*** This paper highlights the important role played by the two parameter Weibull distribution in Internet traffic modeling. They have shown the versatile role played by the simple two parameter Weibull distribution in Internet traffic structural modeling. The Weibull shape parameter can capture traffic inter-arrival (packets, flows and dynamics as it traverses from access to core networks; and, can also be used to zoom in/out between packet-, flow- and session inter-arrivals at a certain tier (access or core) sessions).

***(Asad Arfeen – 2019)*** An extensive literature survey with new developments in Internet traffic count data modelling has been presented. The contributions in this article establish the notion of the "Renewal of Renewal Theory in Internet Traffic Modelling". This is the first study which presents a duplex analysis of all structural components of Internet traffic (packets, flows and sessions) at access and backbone core tiers of Internet. The results have been validated by using real traffic data fitness tests and trace driven queueing performance evaluation. The results of this article will help researchers use simple renewal processes as a better alternate to complex self-similar or modulated stochastic processes for modelling all structural components of Internet traffic at any time scale with physical justifications. A count data model is used to model event counts without recording the timing of individual events. Without doubt it can be stated that the statistical properties of count data are inherited from the statistical properties of the underlying interarrival time distribution (Winkelmann and Baetschmann, 2014). Therefore, the relation between a count model and its timing process (if it is known) can be very useful in augmenting the capabilities of a count model. A comparative study is performed on real internet traffic traces and found out that sum of probability mass assigned by the Weibull count model is closest to the sum of probability mass of session count data in every network.

***(ARISTIDIS K. NIKOLOULOPOULOS)*** They have discussed the most commonly used mixed Poisson distributions, namely the negative binomial, the Poisson inverse Gaussian and the generalized Poisson distributions. The results show that for small mean and overdispersion, all the models are quite the same, whereas for larger means the generalized Poisson and the Poisson inverse Gaussian distributions have larger tails than the negative binomial, and the differences are much larger. In practice, it is not easy to discriminate between them for small counts and small overdispersion, but for large overdispersion discrimination is relatively easy.

***(Wan Fairos – 2010)*** They have reviewed several modeling approach for count data. A good starting point is the basic count model of Poisson regression model. When there is an evidence of overdispersion, an negative binomial regression model is a better choice. The Negative Binomial regression model is more flexible as it allows for the variance to be greater than mean and considers the observed heterogeneity in the model. Another situation exists when overdispersion results from a high frequency of zero counts. If such condition occurs, a modified Poisson models such as Hurdles regression or Zero-Inflated regression model might give a more satisfactory fit to the data.

***(Cameron – 1986)*** They have discussed the modelling strategy based on some simple tests, in which one can proceed to increasingly flexible and data-coherent models, beginning with the basic Poisson model. We have provided a detailed application of this modelling strategy to illustrate both its simplicity and feasibility. Our results are broadly supportive of the QGPMLE procedure advocated by GMT but we also advocate further exploration of suitable categorical variable mnodels as an alternative approach.

**(EL Plan – 2014)** Modeling and Simulation is a powerful tool to characterize, quantify, understand, predict, power, optimize, and rationalize (pre)clinical trial data and studies. The Poisson model, a close relative of the survival model, is the basis for all count data models. Adaptations for handling overdispersion, underdispersion, autocorrelation, or inhomogeneity were proposed in the literature and presented here.

**(Emilio Gómez-Déniz – 2010)** They have introduced a new probability mass function by discretizing the continuous failure model of the Lindley distribution. The model obtained is over-dispersed and competitive with the Poisson distribution to fit automobile claim frequency data. After revising some of its properties a compound discrete Lindley distribution is obtained in closed form. This model is suitable to be applied in the collective risk model when both number of claims and size of a single claim are implemented into the model. The new compound distribution fades away to zero much more slowly than the classical compound Poisson distribution, being therefore suitable for modelling extreme data.

# CHAPTER 3

# COUNT MODELS FOR INTERNET TRAFFIC

### 3.1. Why use Count Models?

There are two main uses of count data models in policy evaluation. Often, the focus is on determining the effect of a policy change on the average count. Other applications exploit the fact that count data models yield predictions for the entire probability distribution. In a policy context, one can therefore determine the effect of the policy for each value of the outcome.

### 3.2. Count Variable

A count variable is a variable that takes on discrete values (0,1,2, ...) representing the number of occurrences of an event in a constant period of time. Substance-using days, number of cigarettes smoked a day, number of arrests; number of hospital admissions and insurance claims are some examples. A count variable can only take on positive integer values or zero because an event cannot happen a negative number of times. Because of this situation, count data are inherently positively skewed with a high proportion of zeros. Analyzing this type of data poses an obstacle. They are not optimally modeled with a normal distribution, especially if the variable of interest is sparse (Cameron and Trivedi 2007; Winkelmann 2008; Hilbe 2007)

### 3.3. COUNT DATA MODELS:

### 3.3.1. Poisson Count Model:

A random variable Y is said to have a Poisson distribution with parameter μ if it takes integer values y = 0, 1, 2, . . . with probability

$$\Pr\{Y = y\} = \frac{e^{-\mu}\mu^y}{y!}$$

for μ > 0. The mean and variance of this distribution can be shown to be

$$\mathbf{E(Y) = var(Y) = \mu}$$

Since the mean is equal to the variance, any factor that affects one will also affect the other. Thus, the usual assumption of homoscedasticity would not be appropriate for Poisson data.

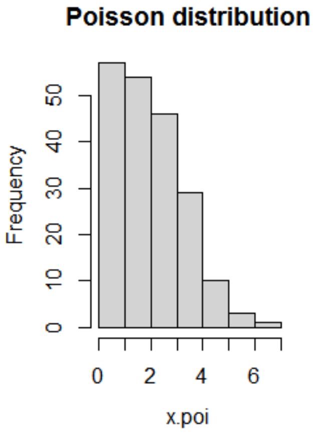
A useful property of the Poisson distribution is that the sum of independent Poisson random variables is also Poisson. Specifically, if Y1 and Y2 are independent with Yi ~ P(μi) for i = 1, 2 then

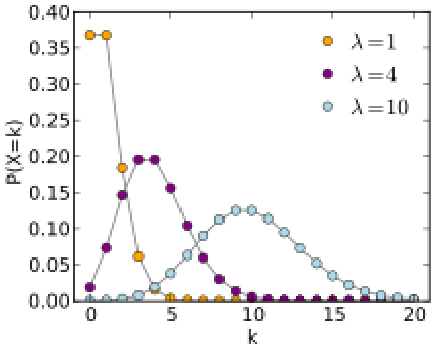
$$\mathbf{Y1 + Y2 \sim P(\mu1 + \mu2).}$$

This result generalizes in an obvious way to the sum of more than two Poisson observations.

The Poisson count model has the following features:

- The Poisson count model generates stationary counts with no trends, that is, event probabilities do not change with time.
- The Poisson count model implies an exponential distribution for the underlying interarrival process and the cumulative interarrival process can be represented by the Erlang distribution.
- The counts resulting from the Poisson count model are independent or uncorrelated.
- The counts resulting from the Poisson count model are equidispersed, that is, the variance of counts is equal to their mean.





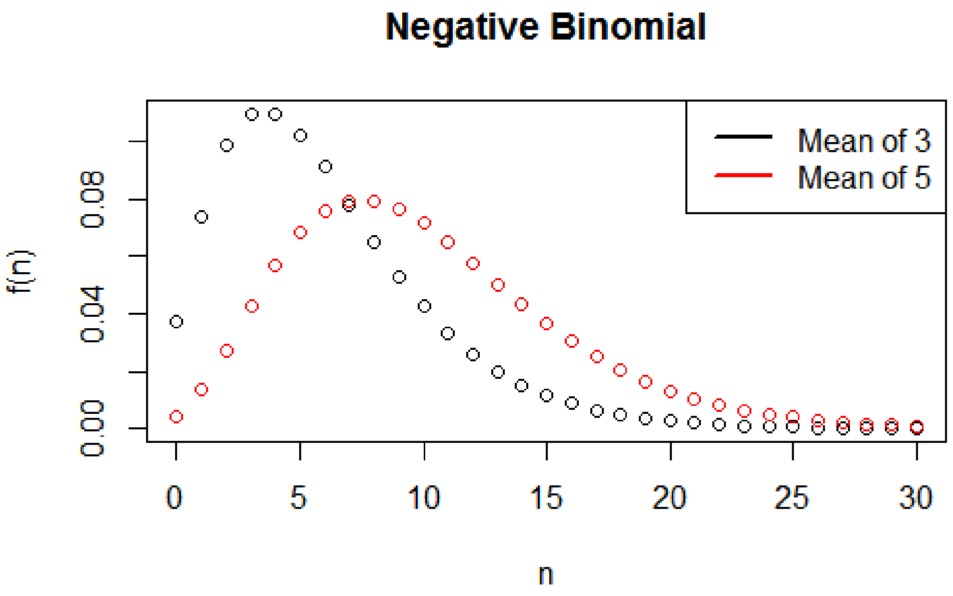
### 3.3.2. Negative Binomial Count Model

The negative binomial count model introduces an additional parameter d and can be defined as :

$$\mathbb{P}(N=k)=\frac{\Gamma(k+d)}{k!\Gamma(d)}\left(\frac{\mu}{\mu+d}\right)^{k}\left(\frac{1}{1+\mu d^{-1}}\right)^{d}, \qquad k=0,1,2,\ldots, \qquad ($$

where µ is the mean and d > 0 is the dispersion parameter which controls the variance to mean relation of the data produced by the negative binomial count model.

Negative binomial regression can be used for over-dispersed count data,  that is when the conditional variance exceeds the conditional mean. It can be considered as a generalization of Poisson regression since it has the same mean structure as Poisson regression and it has an extra parameter to model the over-dispersion. If the conditional distribution of the outcome variable is over-dispersed, the confidence intervals for the Negative binomial regression are likely to be narrower as compared to those from a Poisson regression model.



Negative binomial count model has the following features:

- The model allows arrival dependence based on positive contagion, that is, an arrival (non-arrival) of an event increases (decreases) the probability of the next arrival.
- The model can handle overdispersed data.
- The model allows likelihood ratio and other standard maximum likelihood tests to be implemented.
- The convolution of the negative binomial random variables with the same overdispersion is also negative binomial, irrespective of the mean of the component random variables (see page 459 in [Wilkinson, 1956]). The analytical form of the convolution of negative binomial random variables has been derived in [Furman, 2007].

### 3.3.3. Weibull Count Model

The probability density function of the Weibull distribution is given as

$$f_X(x) = \lambda c x^{c-1} e^{-\lambda x^c}, \qquad x \geqslant 0, c > 0.$$

where c is the shape parameter and λ is the rate parameter.

The hazard rate of Weibull distribution admits a closed form expression as follows:

$$h(t) = \frac{f(t)}{1 - F(t)} = \lambda c t^{c-1}.$$

If N(t) denotes the number of arrivals in time interval (0,t] with interarrival times being Weibull distributed, then the Weibull count model is given as

$$\mathbb{P}[N(t) = n] = \sum_{j=n}^{\infty} \frac{(-1)^{j+n} (\lambda t^c)^j \alpha_j^n}{\Gamma(cj+1)}, \qquad n = 0,1,2,\ldots,$$

The following features of the Weibull count model:

- The model allows overdispersed, equidispersed and underdispersed count data.
- The model is directly connected to the continuous time Weibull distribution for all values of the shape parameter.
- The model is computationally better than the iterative algorithms to calculate probability of counts resulting from the Weibull interarrival times; see [Rinne, 2008], for such algorithms.

### 3.3.4. Mittag Leffler Count Model

The Mittag-Leffler function distribution (MLFD) belongs to the generalized hypergeometric and generalized power series families and also arises as weighted Poisson distributions. MLFD is a flexible distribution which falls under the domain of attraction of stable laws with varying shapes and has a unique mode at zero or it is unimodal with one/two non-zero modes. It can be under-, equi- or over- dispersed.

It’s probability density function is given by:

$$f(x;\alpha,k) = \sum_{k=1}^{\infty} \frac{(-1)^{k-1} k\alpha x^{k\alpha-1}}{\Gamma(1+k\alpha)}, \qquad x > 0, 0 < \alpha \leqslant 1,$$

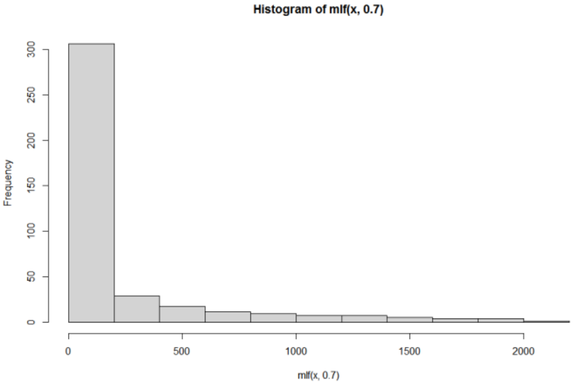
where α is the shape parameter of this distribution.

The hazard rate of the Mittag-Leffler distribution can be written as

$$h(t) = \frac{\sum_{k=1}^{\infty} \left[ \frac{(-1)^{k-1} k\alpha x^{k\alpha-1}}{\Gamma(1+k\alpha)} \right]}{\sum_{k=0}^{\infty} \left[ \frac{(-1)^{k-1} x^{k\alpha}}{\Gamma(1+k\alpha)} \right]}.$$

An expression for probability of counts is given by :

$$\mathbb{P}[N(t)=n]=\sum_{j=1}^{\infty}\left[\frac{\binom{j}{n}(-1)^{j-n}t^{j\alpha}}{\Gamma(1+j\alpha)}\right], \qquad n=0,1,2,\ldots.$$



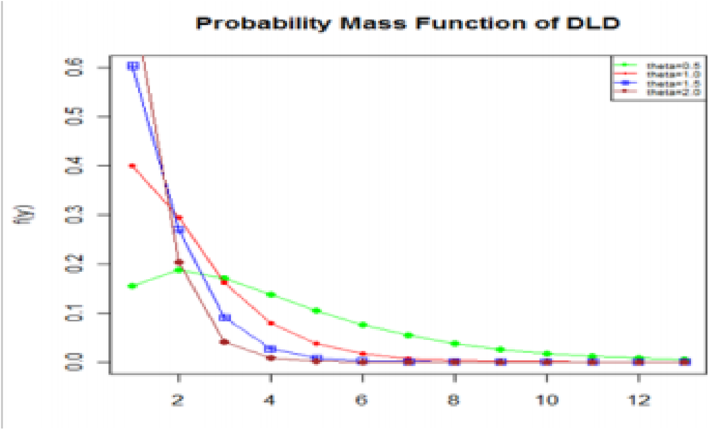
The Mittag-Leffler count model has the following properties:

- ➢ Poisson count model results as a special case for α = 1.
- ➢ All moments of Mittag-Leffler count model are finite for any α.
- ➢ The distribution has been found to fare well when compared with the hyper-Poisson and COM-Poisson type negative binomial distributions in its suitability in empirical modeling of differently dispersed count data.
- ➢ For the range 0 < α < 1, the hazard rate of Mittag-Leffler count model is a decreasing function of time. Therefore, the distribution exhibits negative duration dependence which causes overdispersion in the count data.

### 3.3.5. Discrete Lindley Distribution

*Bakouch et al. (2014)* proposed the discrete Lindley (DL) distribution as a discrete version of the continuous Lindley distribution. The pmf of the discrete random variable Y corresponding to a continuous random variable X following Lindley distribution

$$P_1(Y=y)=P_1(y;\theta)=\frac{\left(e^{\theta}-1\right)^2}{e^{2\theta}}(1+y)e^{-\theta y};\ y=0,1,2,\ldots,\ \theta>0$$



The survival function and CDF are as under:

$$S(y;\theta)=\left[\frac{\left(e^{\theta}-1\right)y+\left(2e^{\theta}-1\right)}{e^{2\theta}}\right]e^{-\theta y};y=0,1,2,\ldots,\ \theta>0$$

$$F_2(y;\theta)=1-\left[\frac{\left(e^{\theta}-1\right)y+\left(2e^{\theta}-1\right)}{e^{2\theta}}\right]e^{-\theta y};y=0,1,2,\ldots,\ \theta>0$$

It is easy to see that limx→∞r(x;θ)=θ. Hence, the parameter θ can be interpreted as a strict upperbound on the failure rate function, an important characteristic for lifetime models.

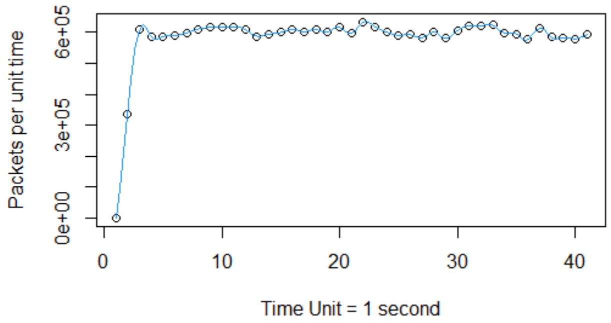
The properties of Discrete Lindley are as under:

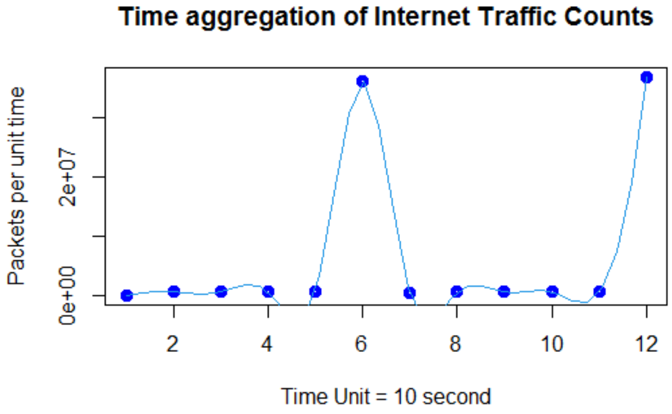
- P y( ;θ ) is log-concave and therefore, the DLD has an increasing hazard rate
- DLD is over-dispersed and hence it can be applied to model over-dispersed data.
- Features include the uni-modality

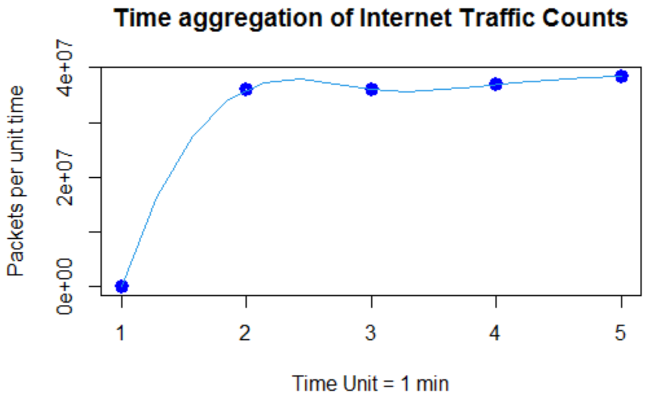
### 3.4. Time Aggregation plots of Internet Traffic

A graphical representation of Internet traffic count data as a function of increasing time interval for traffic aggregation has been introduced in [Paxson & Floyd, 1995; Willinger & Paxson, 1998]. These plots display the count data at various increasing time scales to assess the influence of time aggregation on traffic fluctuations. The multiple time scale view of fluctuations in Internet traffic is qualitative in nature, nevertheless, it can serve as the first step to understand the behaviour of Internet traffic count data.

Following are the time aggregation plots for the Real Data *(Taken few sample points only)*




**Time aggregation of Internet Traffic Counts**


**Time aggregation of Internet Traffic Counts**

### 3.5. Index of Dispersion for Counts:

It is important to clearly define the context of dispersion due to its essential role in modeling count data, and distributions for modeling these data should take into account the data’s dispersion. Generally, the dispersion for any data can be described as the variability or spread of the data. In other words, dispersion refers to the stretch or the squeeze of a data’s distribution. Specifically, dispersion in count data is formally defined in relation to a specified model being fitted to the data (Cameron and Trivedi

(2013) and Hilbe (2014)). In this context, the variance ratio (VR) can be defined as the ratio between the observed variance from the data and the theoretical variance from the model fit, as:

**VR = observed variance / theoretical variance**

Accordingly, modeling any count data might exhibit three types of dispersion; namely, over-dispersion, under-dispersion and equi-dispersion. Over-dispersion refers to the case when the observed variance of the count data is greater than the expected variance specified by the fitted model. Under-dispersion describes the opposite case, where the observed variance is less than that theorized by the model. Equi-dispersion refers to the case of equal variances. Then, a model that fails to capture the over- or under-dispersion in the data and shows different variance than that observed is called an over- or under-dispersed model. Therefore, the definition of dispersion through the VR can be helpful in studying the dispersion of a model.

Moreover, the dispersion of count data can be defined in relation to the Poisson model. Hence, it is common with these data to refer to the dispersion as being relative to Poisson. In such a case, the variance of the model is estimated by the sample mean. Thus, over-, equi- or under- dispersion relative to Poisson refers to cases where the sample variance (observed variance) is greater, equal or smaller than the sample mean (theoretical variance), respectively. Therefore, the dispersion of a dataset, under this definition, can be identified with regard to the Dip, or the dispersion coefficient, which is defined as the ratio of the variance to the mean (variance-to-mean relation):

$$Dip = \frac{\sigma^2}{\mu}$$

# CHAPTER 4

# PARAMETER ESTIMATION AND TESTING

## 4.1. Parameter Estimation

Parameter estimation mainly consists in characterizing a parameter set consistent with measurements, the model and the equation error description. The problem to be solved is that of finding the set of admissible parameter values corresponding to an admissible error. The uncertainties must be treated by a global analysis of the problem: both the equation error and the parameter set are considered unknown. Then, a solution is given as a domain of time-variant parameters and a bounded set of the error. This procedure consists in explaining the measurements performed at all time by optimising a precision criterion based on the poly tope theory.

The essential choice in estimation is between maximum likelihood methods on the one hand, based on strong distributional assumptions, and on pseudo-maximum-likelihood or maximum quasi-likelihood methods on the other, based on weaker assumptions. If the probability distribution of the variable y, is known to belong to a specified parametric family, that is, the data generation process is known, and the likelihood function is well-behaved, maximum likelihood (ML) is the obvious estimation procedure.

### 4.1.1. Maximum Likelihood Estimation:

Maximum likelihood estimation is a method that determines values for the parameters of a model. The parameter values are found such that they maximise the likelihood that the process described by the model produced the data that were actually observed.

Maximum likelihood is a very general technique for parameter estimation and inference in statistics. Suppose we have a density function f(y;θ), characterized

by some unknown but fixed parameters θ, which could be a parameter θ or a vector of parameters θ = (θ1,θ2,...,θP), where P is the number of parameters to be estimated. Then, the maximum likelihood method estimates these parameters by finding the values of θ that maximize the likelihood of Y and θ. Due to the fact that the likelihoods are all positive and the logarithm is an increasing function, the log-likelihood is equivalent to the likelihood, and they have their maximum at the same point. Therefore, it would be easier to maximize the log-likelihood instead of the likelihood since the summation is easier than the product. In other words, this method of estimation can be briefly applied according to the following three steps:

**1. Likelihood Function**:

The likelihood function for an observed sample (y1,y2,...,yn) of size n, which is identically independent distributed (iid) as f(y;θ) and regarded as a function of θ given the sample data, can be defined to be the joint probability function, as follows:

$$L(\underline{\theta}; y) = \prod_{i=1}^{n} f(y_i; \underline{\theta})$$

2. **Log-Likelihood Function**:

   The log-likelihood function is the natural logarithm of the likelihood function which is defined as follows:

$$\ell(\underline{\theta}; y) = \log(L(\underline{\theta}; y)) = \sum_{i=1}^{n} \log(f(y_i; \underline{\theta}))$$

3. **Maximum Likelihood Estimator**:

**A**n MLE ˆ θML of θ maximizes the likelihood, L(θ;y), or typically, the log-likelihood `(θ;y):

$$\hat{\theta}_{ML} = \arg\max_{\theta} \ell(\theta; y)$$

Optimizing the likelihood (or equivalently log-likelihood) functions can be found analytically by differentiating the log-likelihood function `(θ;y) with respect to the parameter θ and setting the results equal to zero. How ever, for some complicated cases this may result in non-linear equations, which might require the application of numerical solutions using several algorithms. The complexity of the MLEs depends on the form of the probability function f(y;θ). The maximum likelihood method is the most commonly applied method of classical inference. This is due to its useful standard large sample properties, such as consistency and asymptotic normality.

Numerous studies have applied this technique to estimate parameters, especially for coefficient regression. In other words, the maximum likelihood approach can be applied to the traditional normal linear regression to estimate its parameters. Moreover, the maximum likelihood approach is used to fit most of the GLMs and Generalized additive models for location, scale and shape.

## 4.2. Model Evaluation and Validation

### 4.2.1 Goodness of Fit Tests

In this section, several goodness-of-fit measures will be briefly discussed, including the Pearson chi-squares, deviance, likelihood ratio test, Akaike Information Criteria (AIC) and Bayesian Schwartz Criteria (BSC).

#### 4.2.1.1. Pearson chi-squares

Two of the most frequently used measures for goodness-of-fit in the Generalized Linear Models are the Pearson chi-squares and the deviance. The Pearson chi-squares statistic is equivalent to,

$$\sum_i \frac{(y_i - \mu_i)^2}{Var(Y_i)}.$$

For an adequate model, the statistic has an asymptotic chi-squares distribution with n - p degrees of freedom, where n denotes the number of rating classes and p the number of parameters.

#### 4.2.1.2. Deviance

The deviance is given by :

$$D = 2(\ell(\mathbf{y};\mathbf{y}) - \ell(\boldsymbol{\mu};\mathbf{y})),$$

where g(lu;y) and g(y;y) are the model's log likelihood evaluated respectively under p and y. For an adequate model, D also has an asymptotic chi-squares distribution with n - p degrees of freedom. Therefore, if the values for both Pearson chi-squares and D are close to the degrees of freedom, the model may be considered as adequate.

The deviance could also be used to compare between two nested models, one of which is a simplified version of the other.

#### 4.2.1.3. Likelihood ratio

An LRT was used to compare the nested models (i.e., NB vs. Poisson, ZINB vs. ZIP, and NBH vs. PH) in order to test whether the over-dispersion parameter would be necessary. In LRT, the null hypothesis is for their stricted or constrained model(null model)and the alternative hypothesis is for the unrestricted or unconstrained model (alternative or full model)(Hilbe2007).

The advantage of using the maximum likelihood method is that the likelihood ratio test may be employed to assess the adequacy of the Negative Binomial I (MLE) or the Generalized Poisson I (MLE) over the Poisson because both Negative Binomial I (MLE) and Generalized Poisson I (NILE) will reduce to the Poisson when the dispersion parameter, a, equals zero.

### 4.2.1.4. AIC

The Akaike information criterion was used to evaluate the goodness-of-fit of the six models and is defined as follows:

$$\mathbf{AIC=-2logL+2p}$$

where log L is the maximum of the likelihood function for a fitted model and p is the number of parameters in the fitted model. The preferred model is the one with the minimum AIC value. (BurnhamandAnderson2004).

### 4.2.1.5. BIC

The BIC is defined as (Schwartz),

$$\mathbf{BIC = -2L + p\ Log(n)}$$

where L denotes the log likelihood evaluated under u, p the number of parameters and n the number of rating classes. For this measure, the smaller the BIC, the better the model is.

### 4.2.1.6. Rootogram

Rootogram or Regression models and showed that this is especially useful for diagnosing and treating issues such as over dispersion and/or excess zeros in count data models. The rootogram compares the observed and expected values graphically by plotting histogram-like rectangles or bars for the observed frequencies and a curve for the fitted frequencies, all on a square-root scale. There are three types of rootogram: standing, hanging and suspended. The basic version, the standing rootogram,is the least useful among the three. It simply plots rectangles/bars and a curve representing the model, but the fit is not easily assessed. Hanging rootograms emphasize the fitted values and suspended rootograms, the corresponding residuals.

#### 4.2.1.7. P-Value

The probability of obtaining a chi-square value greater than the above. This is the significance level of the test. If this value is less than some predefined alpha level, say 0.05, the variable is said to be statistically significant.

# CHAPTER 5

# TOOLS, DATASET AND INTERNET TRAFFIC ARCHIVE

## 5.1. Internet Traffic Archive:

### 5.1.1 THE CENTER FOR APPLIED INTERNET DATA ANALYSIS (CAIDA)

The Center for Applied Internet Data Analysis (CAIDA) manages network research and builds research infrastructure to support large-scale data collection, curation, and data distribution to the scientific research community. It maintains a growing number of computational and data analysis services. As internet network has reached its prescribed limits CAIDA is helping in routing, security, testbeds management. Its data archive is based upon sampling of internet traffic as opposed to flow or packet archiving. It does not store external data but of its own reaching 32 TB. The storage of meta-data as only and indexing of it help in quick access and does not add any overhead burden on CAIDA servers.

### 5.1.2 WAIKATO INTERNET TRAFFIC STORAGE (WITS)

The Waikato Internet Traffic (WITS) Storage project maintains and document the internet traffic traces for researchers and scientists. It provides only some traces for public user because of legal constraints. The too it uses is a library Libtrace, written in C language and Java. The library mainly works on packet capturing methods of archiving. This library is used by WITS for multiple types of inputs without any losing any information. The packet captured by it is archived along with its meta-data in the indexing based upon the time and date of capture. However, it is possible to browse the WITS archive using but attempts to download the trace files require IPv6 hosts. Wits has also mirrored its trace on a repository. These repositories apply certain restriction for the usage of their data and required to create accounts for that purpose.

### 5.1.3 WIDE PROJECT (MAWI GROUP)

WIDE (MAWI GROUP) project is an initiative to facilitate researchers in the domain of internet traffic networks. It provides a data repository of backbone traffic. Traffic traces are collected by tcpdump and, after removing privacy information and anonymity, removal of TCP and UDP payloads and IP masking, the traces are made open to the public. Tcpdpriv is used to remove user data while tcpdstat is used to get summary of a tcpdump file in pcap format. It archives packets by tcpdump library. The backbone trace of fifteen minutes is capture and archived on daily basis. The trace is available for public in zipped format. The sampling points are: first one is trans-pacific 1.5Mbps T1 line, from U.S. to Japan link and second is 6Bone is located on a FastEthernet segment connected to NXPIXP-6 (An IPv6 internet exchange point in Tokyo).

### 5.1.4 INTERNET TRAFFIC STATISTICS ARCHIVE (ITSA):

Internet traffic statistics archive works on flow-level traffic measurement by similar to NetFlow or IPFIX from multiple sources. As flow-enabled devices are ubiquitous in networks therefore, flow data is the most suitable for the traffic measurement. It computes pertinent traffic statistics and then uploads those public accessible repositories in the World Wide Web.

The archive begin its route from the router the forwards the traffic towards archive which afterwards captured by NetFlow – a tool for flow capture. Afterwards that captured data is processed and repots through JSON are made.

### 4.2. LibTrace

Libtrace is a library for trace processing. It supports multiple input methods, including device capture, raw and gz-compressed trace, and sockets; and multiple input formats, including pcap and DAG. Libtrace comes bundled with a series of tools that perform most common trace manipulation tasks.

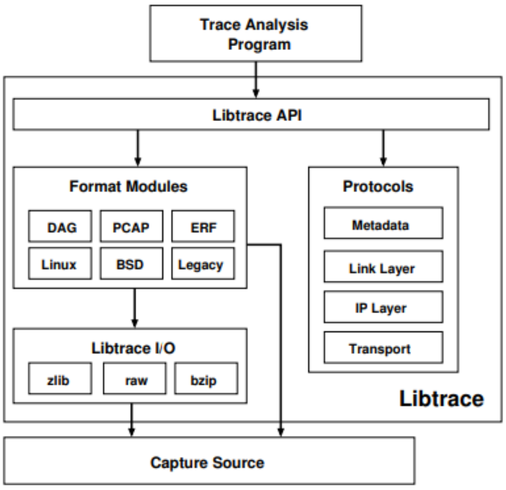


Figure 1. The architecture of libtrace.

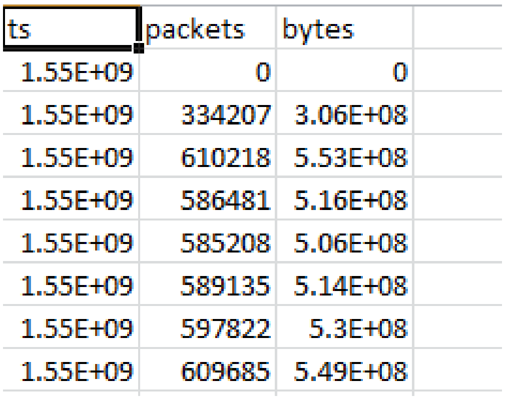
These include:

- traceanon - anonymises trace files
- traceconvert - converts a trace from one format to another
- tracediff - reports differences between two trace files
- traceends - summarises traffic sent and received by endpoints
- tracefilter - applies a BPF filter to a trace
- tracemerge - merges multiple trace inputs into a single trace
- tracepktdump - displays packet contents in a readable format, similar to tcpdump
- tracereplay - replays a trace file using original timing
- tracereport - produces a variety of reports on a trace
- tracertstats - produces stats about an input trace in real time
- tracesplit - splits trace files
- tracesplit_dir - splits trace files based on packet direction
- tracestats - summarises number of bytes and packets matching BPF filters
- tracesummary - summarises the basic stats for a trace
- tracetop - reports the busiest flows over time, similar to ntop
- tracetopends - reports the busiest endpoints in a trace

In this research study, tracertstats tool is used to get packet and byte counts. An example of using tracertstats is :

```
tracertstats -i 1 pcap: equinix-nyc.dirA.20190117-125910.UTC.pcap.gz > logFile.csv
```

*Example of the results produced is as under:*



| ts | packets | bytes |
|---|---|---|
| 1.55E+09 | 0 | 0 |
| 1.55E+09 | 334207 | 3.06E+08 |
| 1.55E+09 | 610218 | 5.53E+08 |
| 1.55E+09 | 586481 | 5.16E+08 |
| 1.55E+09 | 585208 | 5.06E+08 |
| 1.55E+09 | 589135 | 5.14E+08 |
| 1.55E+09 | 597822 | 5.3E+08 |
| 1.55E+09 | 609685 | 5.49E+08 |

Where ts = time in milliseconds

Packets = packets count

Bytes = bytes count

### 4.3. R Software

The models are fitted and estimated in R software where all models used could be done by already existing functions.

# CHAPTER 6

# SIMULATION AND COMPARATIVE STUDY

## 6.1. Simulation

Simulation is carried out on synthetic datasets. Datasets are generated and then passed into a model to get fitted graphs.

## 6.1. Modelling and Estimation on Synthetic Datasets

We first validate the reliability of our model selection by applying it to synthetic data.

### 6.1.1. Approach – 1

**Using Negative Binomial and Uniform Distribution with zeroes and outliers proportion**

- The simulation study for generating synthetic data of counts was carried out by considering a proportion of zeroes and outliers.
- Numbers are generated using Negative Binomial and Uniform Distribution for a sample size of 200.

Steps to generate Synthetic Data using this approach and performing modelling are as under:

**Step 1:** n = 200 is set to sample size.

**Step 2:** The zero-inflation is defined as 25%.

**Step 3:** The outlier ratio is defined as 1%.

**Step 4:** The integers between 20 and 39 are defined as the set of outliers.

**Step 5:** The number of observations without outliers is determined by the formula

n1 = n – (outlier ratio)

That is, n1 = 200 – (0.01x 200) = 198 is obtained.

**Step 6:** n1 = 198 numbers are generated from Negative Binomial distribution and Uniform distribution to obtain count data at the desired zero-inflation ratio (%25).

**Step 7:** The remaining 2 observations are randomly selected from the set of outliers specified in step 4.

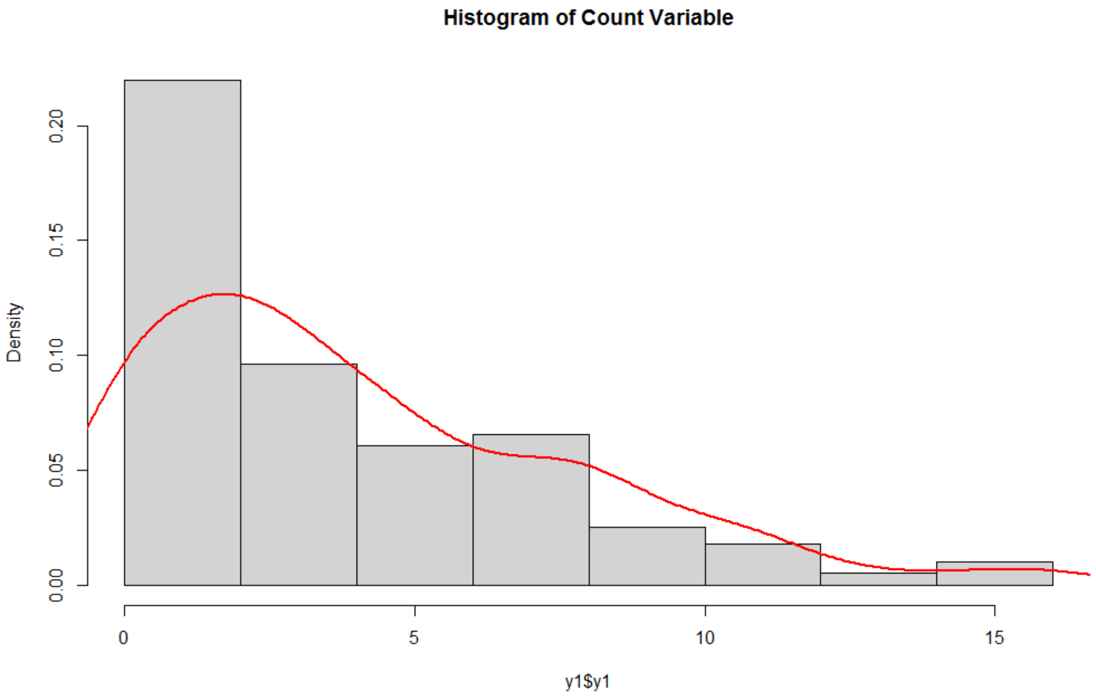
**Step 8:** By combining the observations obtained in steps 6 and 7, the number generation for the dependent variable (Y) is completed with n = 200 numbers.

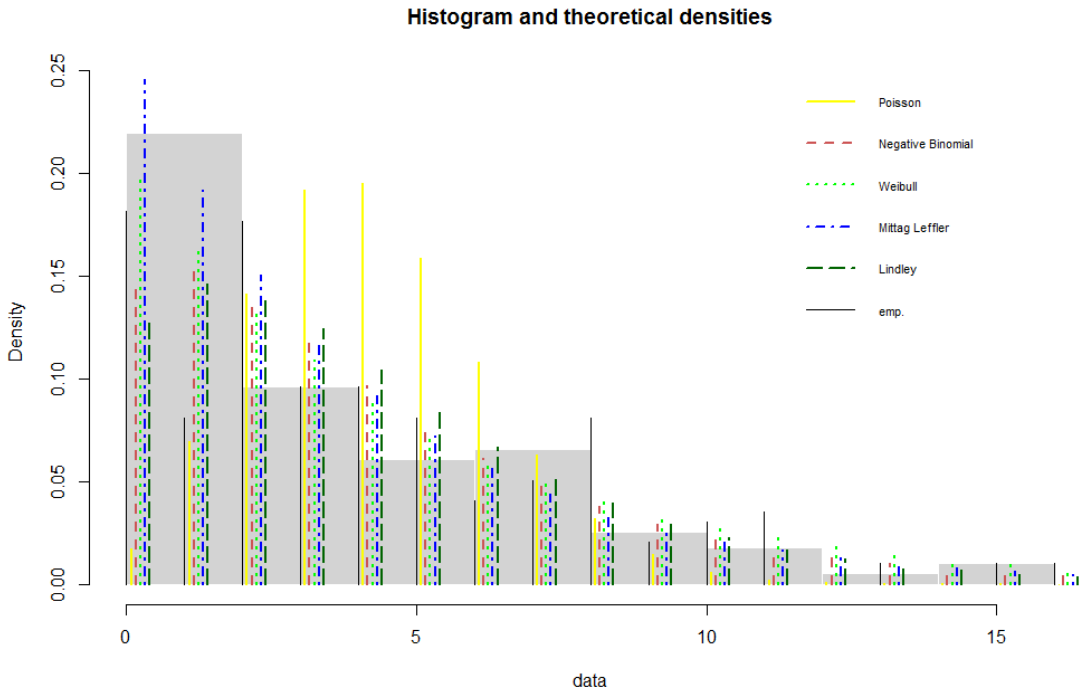
**Step 9:** Once synthetic count data is generated, it is passed to all the four distributions which are considered in this research scope to get their fitted counts.

**Step 10:** Plot their graphs to visualize the result of the comparison study.

#### 6.1.1.1. Without Outliers:

Following are the histogram, plots of fitted distributions formed by using the above mentioned steps but outliers were not generated in the following approach:





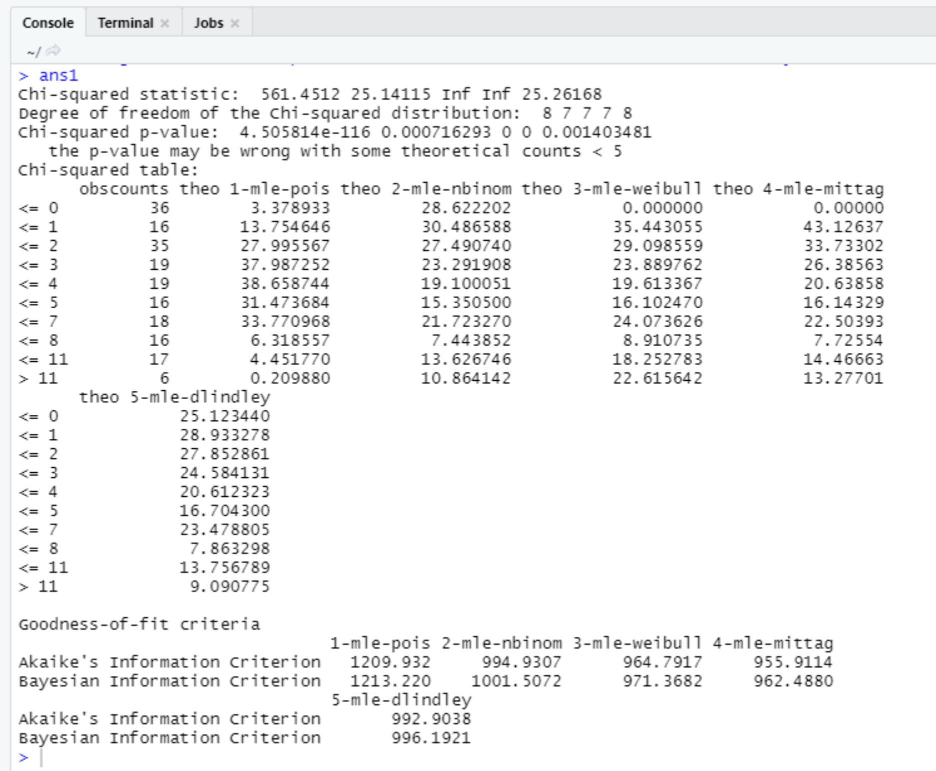
**Summary of the data**:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

  0.000   1.000   3.000   4.071   6.000  16.000
```

```
Mean = 4.070707

Variance = 13.31477
```

In the above comparison, **Mittag Leffler** shows a promising result as it assigns low mass to the values at the tail end.

Goodness of fit test snapshot is as under:



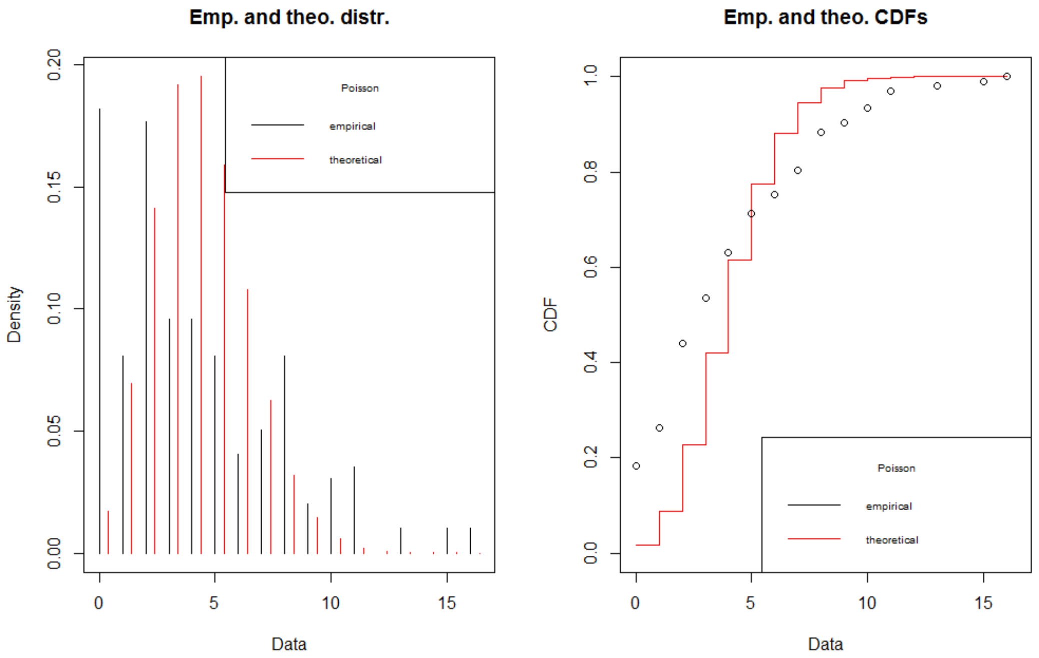
```
> ans1
Chi-squared statistic:  561.4512 25.14115 Inf Inf 25.26168
Degree of freedom of the Chi-squared distribution:  8 7 7 7 8
Chi-squared p-value:  4.505814e-116 0.000716293 0 0 0.001403481
   the p-value may be wrong with some theoretical counts < 5
Chi-squared table:
      obscounts theo 1-mle-pois theo 2-mle-nbinom theo 3-mle-weibull theo 4-mle-mittag
<= 0         36        3.378933         28.622202          0.000000          0.00000
<= 1         16       13.754646         30.486588         35.443055         43.12637
<= 2         35       27.995567         27.490740         29.098559         33.73302
<= 3         19       37.987252         23.291908         23.889762         26.38563
<= 4         19       38.658744         19.100051         19.613367         20.63858
<= 5         16       31.473684         15.350500         16.102470         16.14329
<= 7         18       33.770968         21.723270         24.073626         22.50393
<= 8         16        6.318557          7.443852          8.910735          7.72554
<= 11        17        4.451770         13.626746         18.252783         14.46663
> 11          6        0.209880         10.864142         22.615642         13.27701
      theo 5-mle-dlindley
<= 0            25.123440
<= 1            28.933278
<= 2            27.852861
<= 3            24.584131
<= 4            20.612323
<= 5            16.704300
<= 7            23.478805
<= 8             7.863298
<= 11           13.756789
> 11             9.090775

Goodness-of-fit criteria
                               1-mle-pois 2-mle-nbinom 3-mle-weibull 4-mle-mittag
Akaike's Information Criterion   1209.932     994.9307      964.7917     955.9114
Bayesian Information Criterion   1213.220    1001.5072      971.3682     962.4880
                               5-mle-dlindley
Akaike's Information Criterion      992.9038
Bayesian Information Criterion      996.1921
> 
```

| VARIABLE | POISSON | NEGATIVE BINOMIAL | WEIBULL | MITTAG LEFFLER | DISCRETE LINDLEY |
|---|---|---|---|---|---|
| **Log Likelihood of Parameters** | -603.966 | -495.4653 | -480.3958 | -475.9557 | -495.4519 |
| **AIC** | 1209.932 | 994.9307 | 964.7917 | 955.9114 | 992.9038 |
| **BIC** | 1213.220 | 1001.5072 | 971.3682 | 962.4880 | 996.1921 |
| **Estimates** | lambda<br>4.070707 | size<br>1.442637<br>mu<br>4.070488 | Scale 5.07<br>Shape 1 | tail<br>1.000000<br>scale<br>4.070707 | theta<br>0.3781611 |
| **Standard Error** | 0.1433845 | 0.2186501<br>0.2802790 | NA<br>NA | NA<br>NA | 0.01943115 |

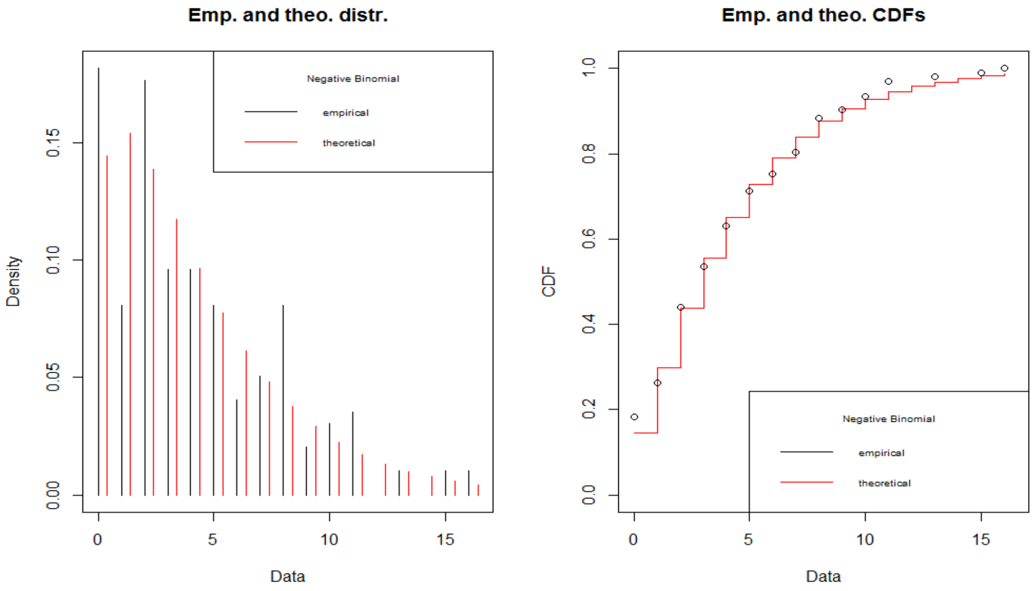
**Table 1 : Summary of Results with Data set 1 (Without Outliers)**

Plots summary for each fit are as under:

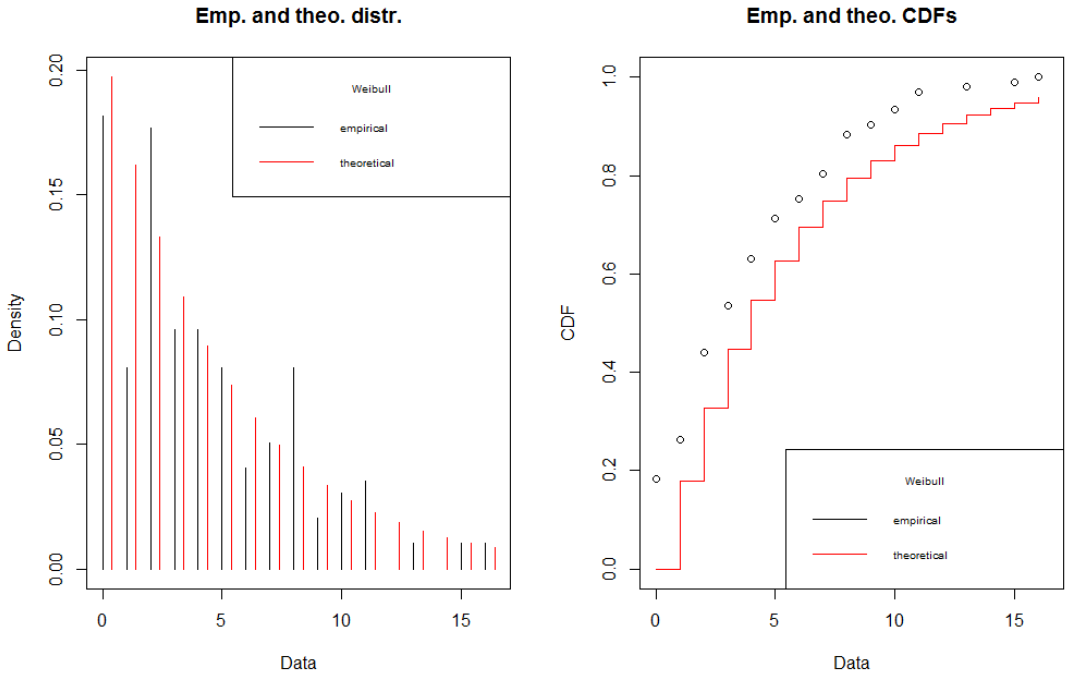
### Poisson Fit:



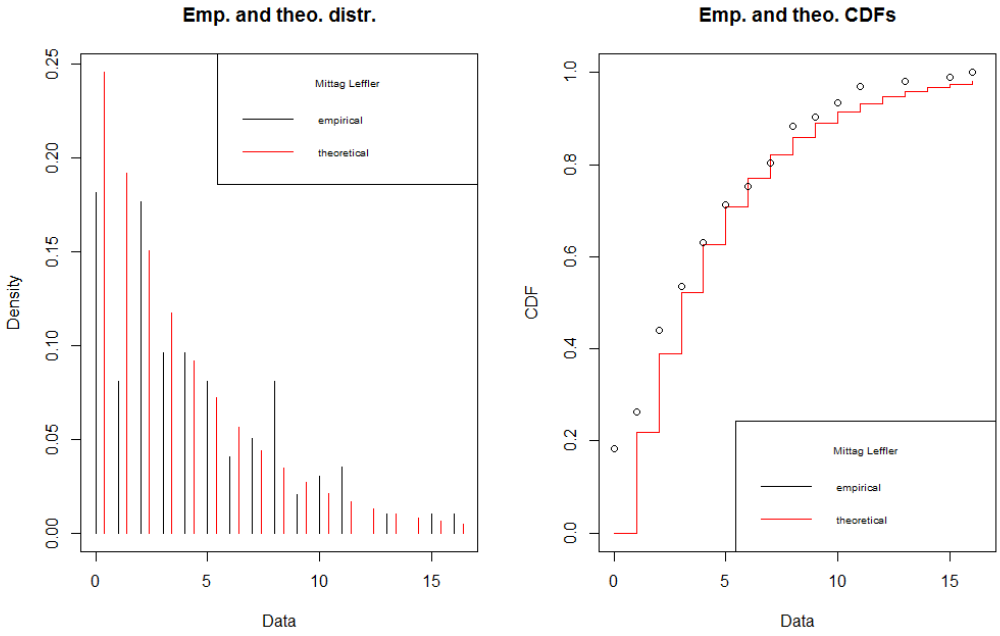
### Negative Binomial Fit:



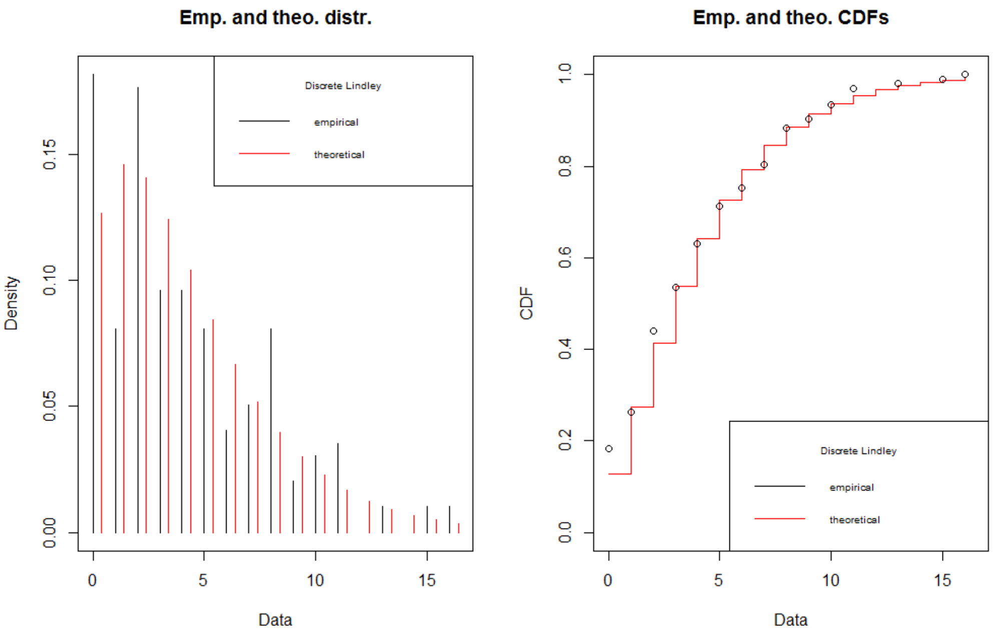
## Weibull Fit:



## Mittag Leffler Fit:

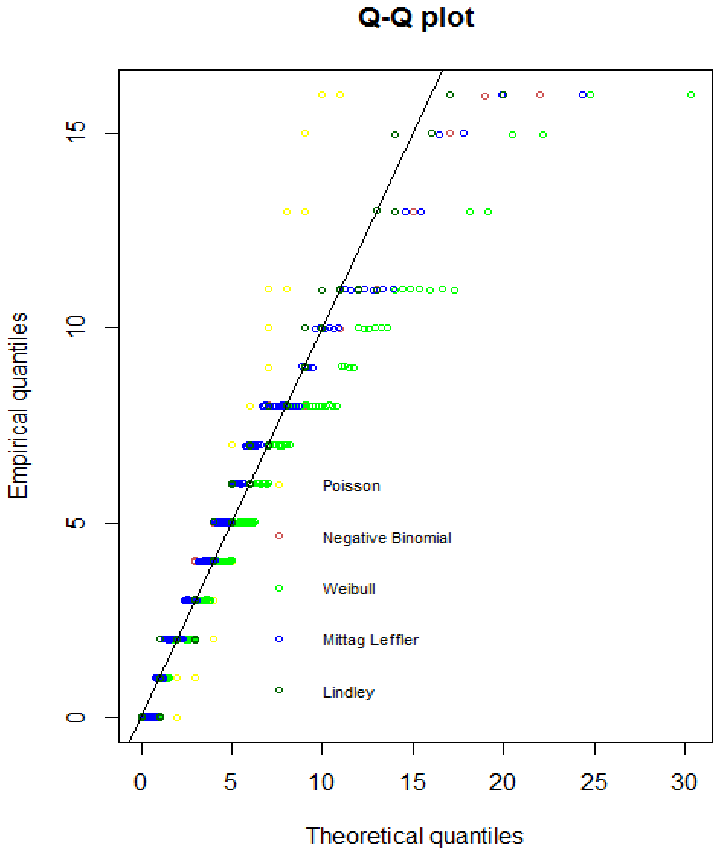


### Discrete Lindley:



### Q-Q plot:

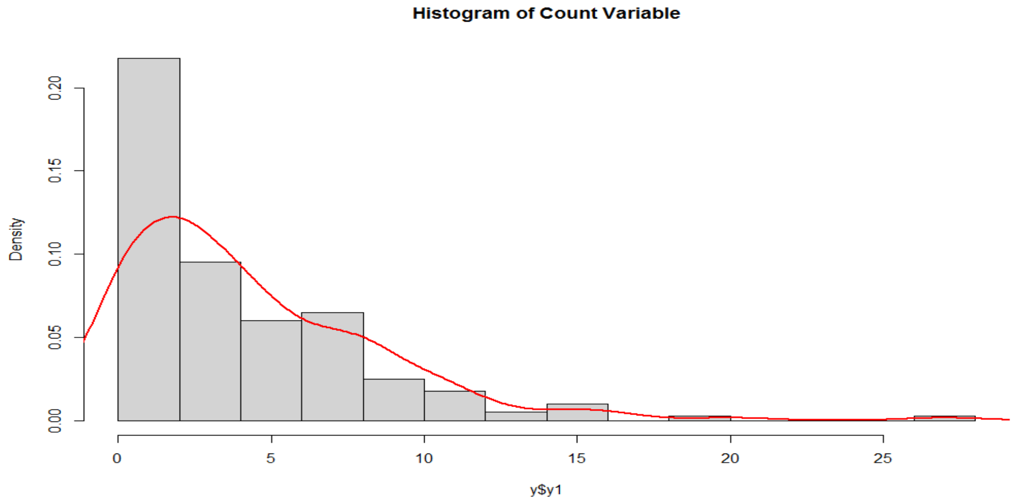
The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. A 45-degree reference line is also plotted.



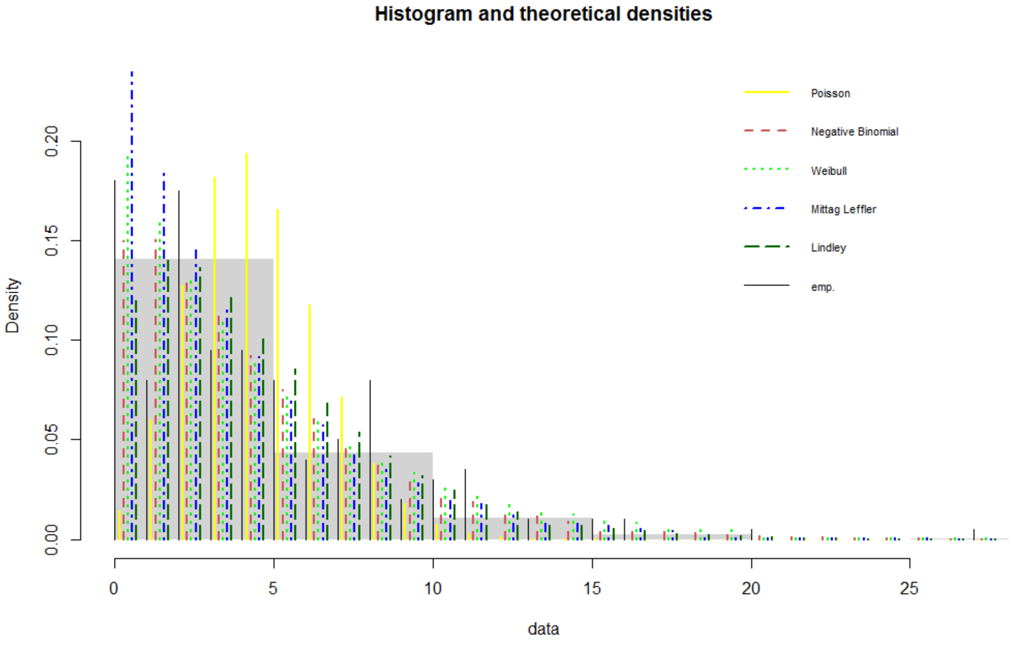
### 6.1.1.2. With Outliers

The algorithm used for the generation of synthetic counts are described in section 6.1.1.

Histogram of the count is shown below:



Fitted Distributions are shown here:



**Summary of the Data:**

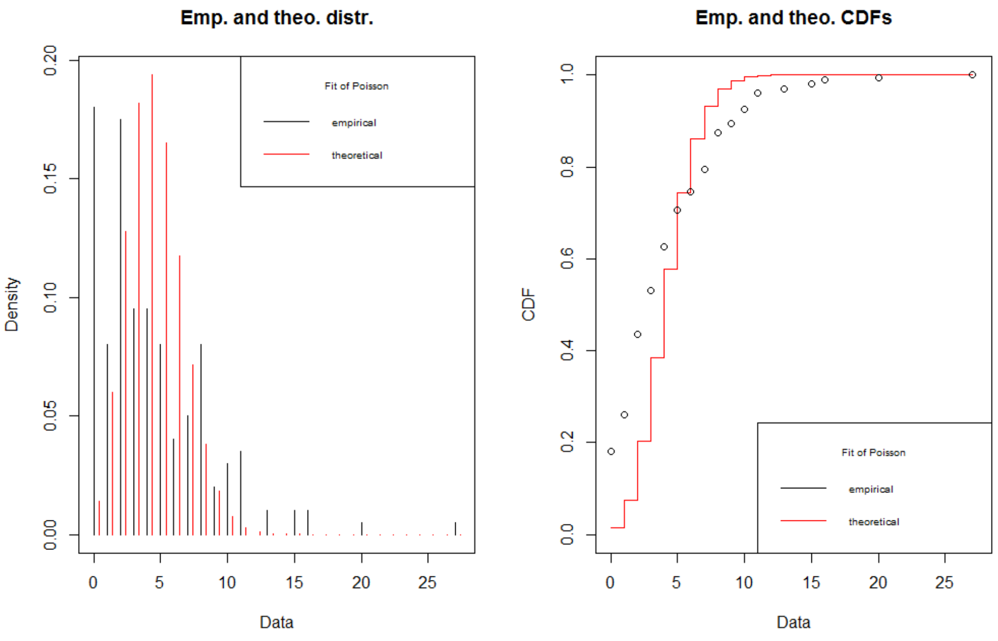
```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

  0.000   1.000   3.000   4.265   7.000  27.000
```

```
Mean  = 4.265

Variance = 17.06008
```

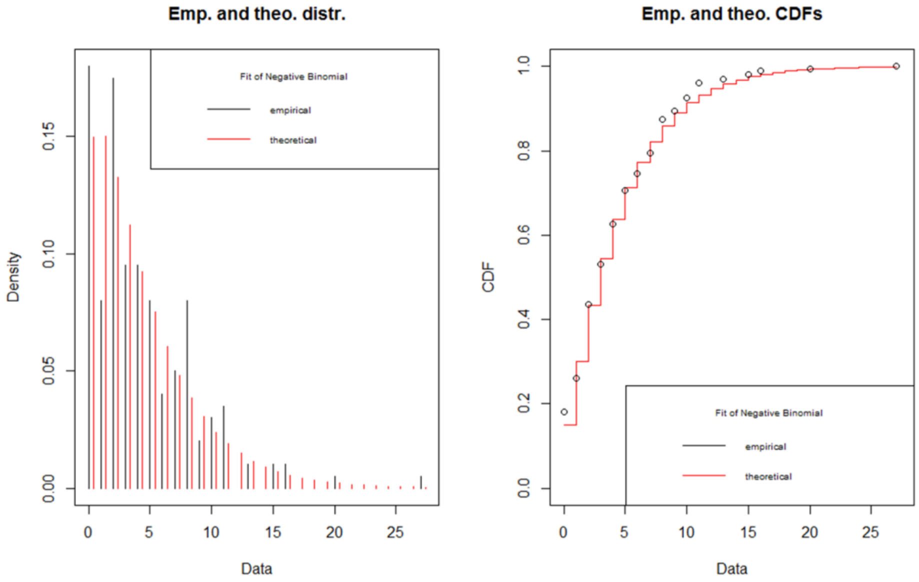
Overdispersion is present since mean and variance is not equal.

Separate plots are shown below:

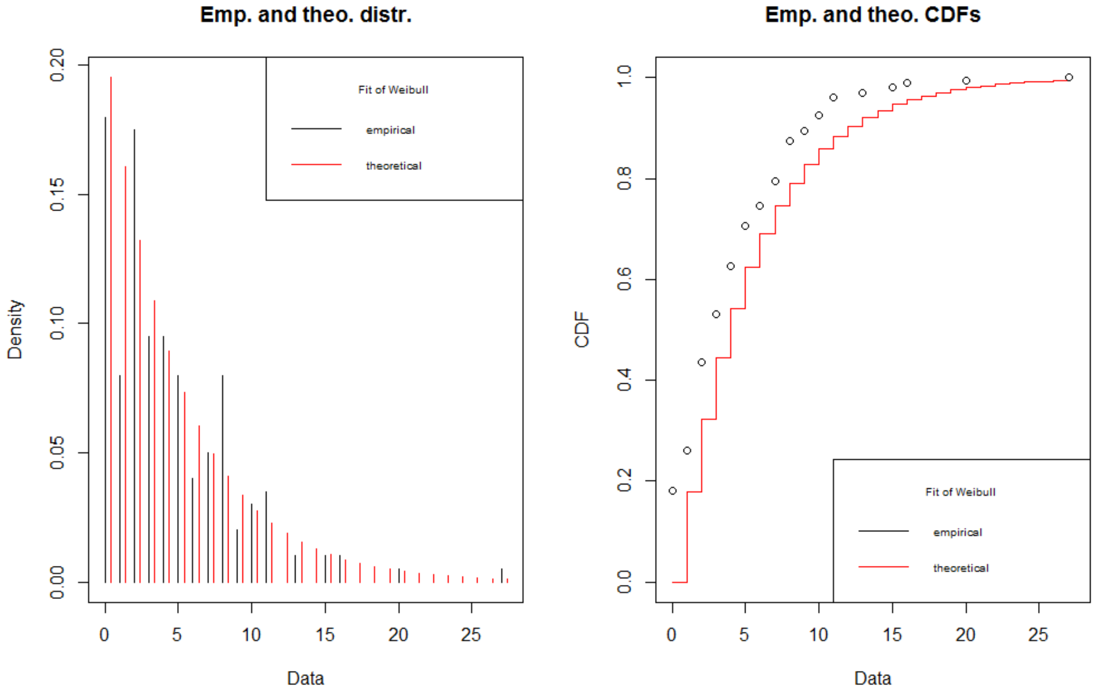
**Poisson fit:**



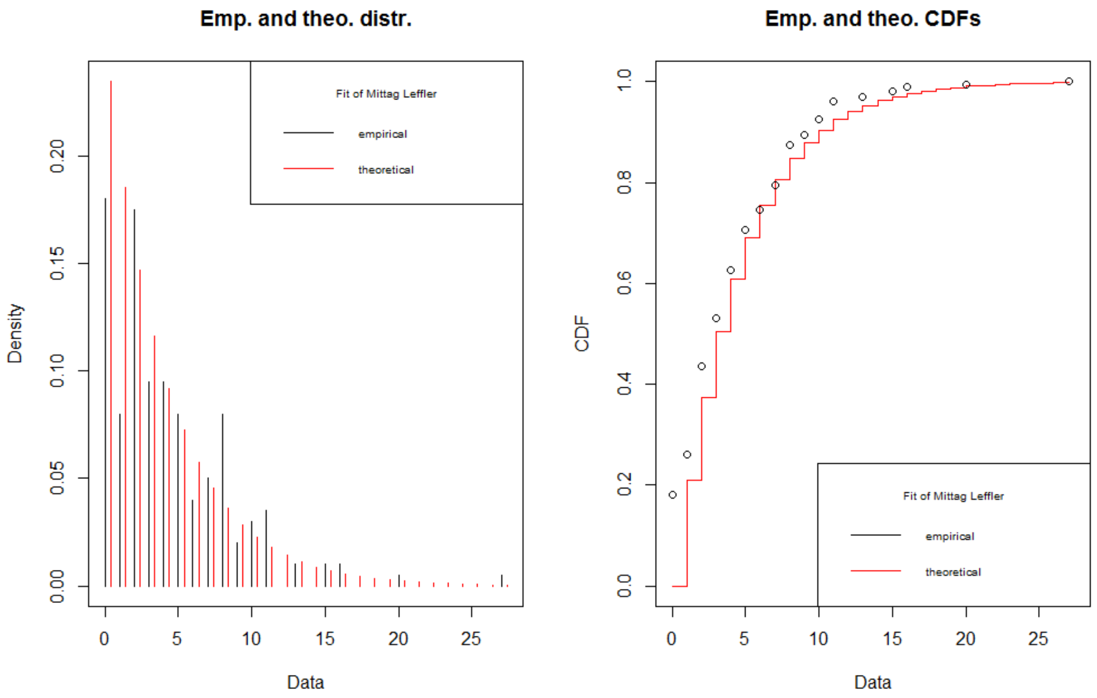
## Negative Binomial Fit:



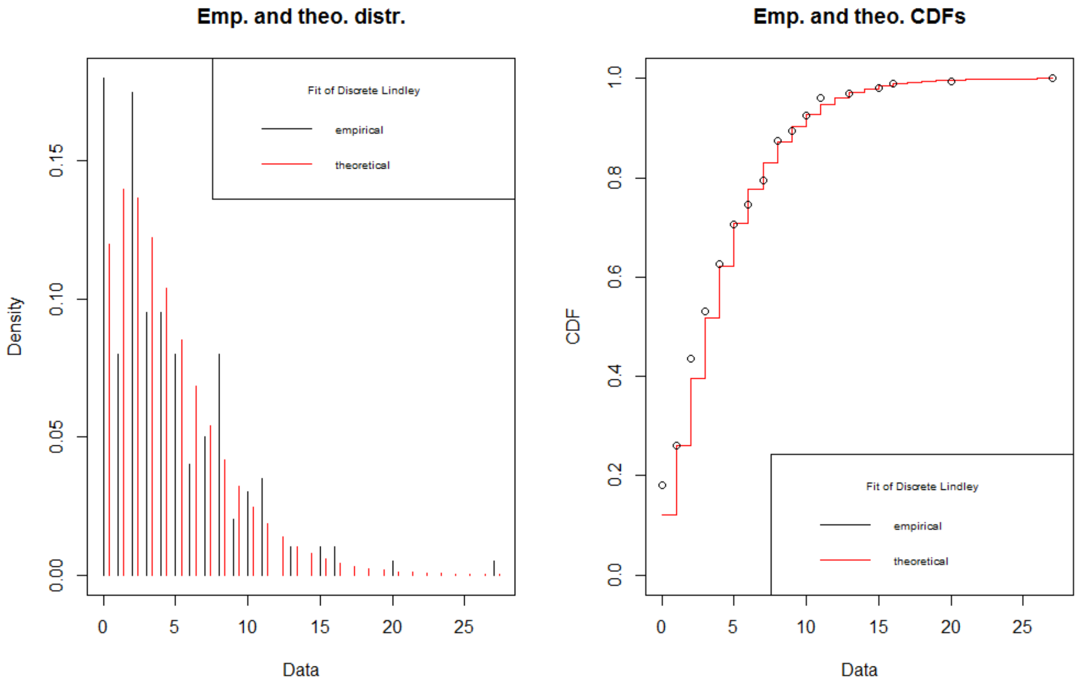
## Weibull Fit:



## Mittag Leffler Fit:



## Discrete Lindley Fit:



**GOODNESS OF FIT TEST:**

| VARIABLE | POISSON | NEGATIVE BINOMIAL | WEIBULL | MITTAG LEFFLER | DISCRETE LINDLEY |
|---|---|---|---|---|---|
| **Log Likelihood of Parameters** | -625.1082 | -510.0775 | -493.2325 | -490.0884 | -511.4333 |
| **AIC** | 1306.216 | 1024.155 | 990.4649 | 984.1769 | 1024.867 |
| **BIC** | 1309.515 | 1030.752 | 997.0615 | 990.7735 | 1028.165 |
| **Estimates** | lambda<br>4.265 | size<br>1.310571<br>mu<br>4.264671 | scale<br>5.12<br>shape<br>1.00 | tail<br>1.000000<br>scale 4.265 | theta<br>0.364205 |
| **Standard Error** | 0.1460308 | 0.1878656<br>0.3011617 | NA<br>NA | NA<br>NA | Std. Error<br>0.01859654 |

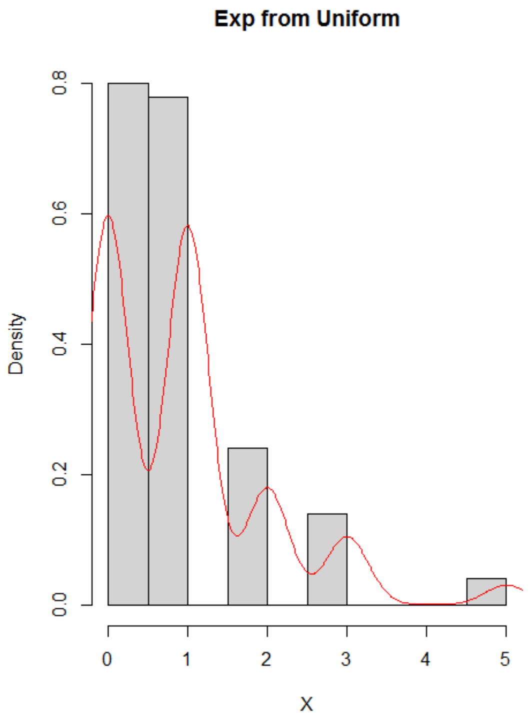
**Table 1 : Summary of Results with Data set 1 (With Outliers)**

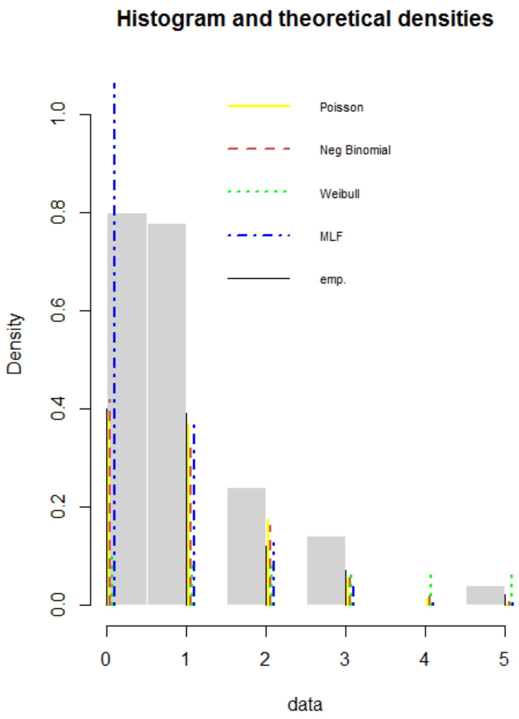
### 6.1.2. Approach – 2

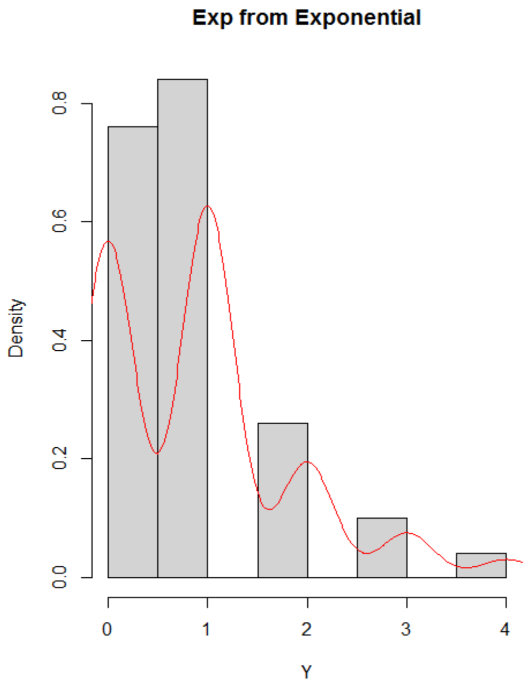
**Using Uniform and Exponential Distribution and Round – off values**

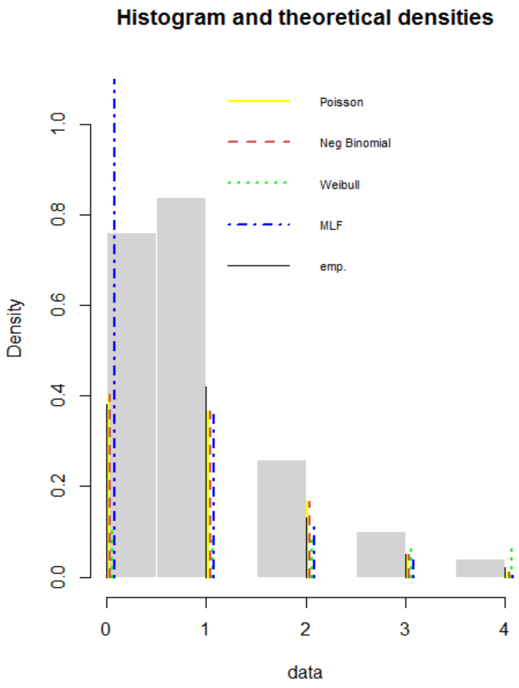
- Using Uniform and Exponential Distribution for random numbers generations and then rounding it off to get discrete data.

- Nsim=10^2 #number of random variables
- U= runif(Nsim)
- X= round(-log(U)) #transforms of uniforms
- Y= round(rexp(Nsim, 1)) #exponentials from R


Exp from Uniform
Density
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
X


Histogram and theoretical densities
Density
0.0
0.2
0.4
0.6
0.8
1.0
Poisson
Neg Binomial
Weibull
MLF
emp.
0
1
2
3
4
5
data


Exp from Exponential
Density
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
Y


Histogram and theoretical densities
Density
0.0
0.2
0.4
0.6
0.8
1.0
Poisson
Neg Binomial
Weibull
MLF
emp.
0
1
2
3
4
data

## 6.3. Results and Discussion

It is important to note that the results of this work were limited to the assumptions that the count data has at least some zero count and that the zeros have an importance attached to them. In the context of Internet Traffic Modelling, we can map it to the assumption of no packet arrived within that particular time frame.

A sample of 200 count data points composed of a fixed proportion of zeros was simulated. The algorithm is mentioned in *section 6.1.1*. The two datasets were generated using the same algorithm with a only difference of outliers. In the second dataset, two outliers were also added. A regression was performed for each of the simulated data set. The average AIC, BIC based on each of the five models (Poisson, Negative binomial, Weibull, Mittag Leffler, Discrete Lindley) was obtained. The mean and the variance of the simulated response variable were also noted.

From table 1, for 0.25 proportion of zeros in the simulated count data set, the average AIC for Poisson model (1209.932) was the highest while that for Mittag Leffler (955.114) was the least. While BIC for Poisson model (1213.220) was the highest and for the Mittag Leffler (962.4880) it was the least. AIC and BIC are usually interpreted in the lower is better fashion. The mean of the response variable was less than the variance. These results are interpreted to mean that if the response variable is made of about 25% zeros and it is also over-dispersed (variance > mean), then the best model to use is the Mittag Leffler as opposed to other four models.

Poisson model fits badly to over-dispersed count data with about 25% proportion of zeros. Other three models are better compared to Poisson as depicted by their lower

AICs. Of all the models under consideration, Mittag Leffler scores the best in terms of AIC and BIC values at this level (0.25) of zero proportions in the count data.

Compared with the empirical probability mass of traffic counts (synthetically generated), it can be observed that the Mittag Leffler count model has much better performance than other count models. In the body part of the traffic count data distribution (lower values), Poisson and negative binomial count models provide better fits, but they cannot capture the probability mass of the higher count values in the tail part of Internet traffic count data distribution. On the other hand, the Mittag Leffler, Weibull and Lindley model performs best in the tail part and assigns higher probability mass to the higher quantiles but among these Mittag Leffler seems the most promising one. The Weibull Count Model gives the next best fits and can be considered as a close competitor of the Mittag Leffler count model.

# CHAPTER 7

# CONCLUSION AND FUTURE WORK

## 7.1. Limitations and Gaps

The reliability of conclusions on policy effects depends on the validity of the assumptions underpinning count data modeling. Several specification tests are discussed in for example, for the Poisson assumption of equality between variance and mean against the alternative of overdispersion. In practice, it is hard to defend any count data model as being exactly true. Rather, such models should be regarded as approximations of the truth, the results being approximate effects. And for statistical inference, it is always good practice to report robust standard errors.

The main limitation we have faced here is the limitation of memory and cpu intensive processes, since when we used real data set with large numbers, it requires a large amount of memory to process and perform iterations and develop a realistic error prone model.

## 7.2. Concluding Remarks

In this research we have examined five different discrete distributions to use as count data model for Internet traffic. Discrete probability distributions play an important role in modeling the counts. The count data sets are generally overdispersed. This study introduces a flexible discrete distribution to model these kinds of data sets and we have found Mittag Leffler to be very promising in this role. We have shown that the Mittag Leffler meets our criteria for a simple and accurate traffic model, since it provides support for both over and under dispersion. It can be used to predict future multiplexing and link efficiency levels. The main advantage of the MLF distribution against the existing ones is that the statistical properties of this distribution are in explicit forms which are important in statistical inference. The importance of the MLF distribution is demonstrated via two synthetic data sets and compared with four other competitive models.

### 7.3. Future Work

- Use of count data model is capacity planning of future networks is also an important area both for researchers and network practitioners and service providers
- Inferring full traffic characteristics from partial measurements (Netflow etc) has always been a challenging area of research especially for large scale network operators. It is expected that further research on Weibull count data modelling can help solve this issue
- Sequential or real-time estimation of the parameters is another challenging task which can be very useful in traffic monitoring and anomaly detection.
- Perform a comparative study on some newly developed discrete distributions.
- Perform fitting and modelling on real internet traffic traces.

# REFERENCES

[1] A. A. M. Saleh and J. M. Simmons, “Statistical Models for Count Data”, Science Journal of Applied Mathematics and Statistics Volume 4, Issue 6, December 2016, Pages: 256-262

[2] Cameron, A., & Trivedi, P. (1999). Regression analysis of count data. Cambridge University Press.

[3] V. Paxson and S. Floyd, “Wide area traffic: the failure of poisson modeling”, IEEE/ACM Transactions on Networking, 3(3):226–244, June 1995.

[4] Winkelmann, R. (1997), Econometric Analysis of Count Data, Berlin, Springer-Verlag

[5] Muhammad Asad Arfeen, K. Pawlikowski, D. McNickle, A. Willig (2013) The Role of the Weibull Distribution in Internet Traffic Modeling

[6] Asad Arfeen a,∗, Krzysztof Pawlikowski b, Don McNickle c, Andreas Willig (2019) The role of the Weibull distribution in modelling traffic in Internet access and backbone core networks” Journal of Network and Computer Applications Volume 141, 1 September 2019, Pages 1-22

[7] Aristidis K. Nikoloulopoulos a & Dimitris Karlis (2008) On modeling count data: a comparison of some well-known discrete distributions

[8] Wan Fairos Wan Yaacob1 Mohamad Alias Lazim1 Yap Bee Wah1 (2010) “A Practical Approach in Modelling Count Data” Proceedings of the Regional Conference on Statistical Sciences 2010 (RCSS’10)

[9] Cameron, A.C. & Trivedi, P.K. (1986). Econometric Models on Count Data: Comparisons and Applications of Some Estimatos and Tests. Journal of Applied Econometrics. 1(1): 29-53.

[10] EL Plan Modeling and Simulation of Count Data CPT Pharmacometrics Syst. Pharmacol. (2014)

[11] Review of Probability Distributions for Modeling Count Data

F. William Townes F. William Townes Department of Computer Science, Princeton University, Princeton, NJ January 14, 2020

[12] Handling Overdispersion with Negative Binomial and Generalized Poisson Regression Models Noriszura Ismail and Abdul Aziz Jemain

[13] G. Schwartz, "Estimating the Dimension of a Model", Annals of Statistics, 1978, Vol. 6, 461-464.

[14] Nakagawa, T. and Osaki, S. The discrete Weibull distribution. IEEE Transactions on Reliability, 24(5):300–301, 1975.

[15] Lambert, D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, 34(1):1–14, 1992.

16] Lawless, J. F. Negative binomial and mixed Poisson regression. Canadian Journal of Statistics, 15(3):209–225, 1987.

[17] *CHARACTERISATION AND MODELLING OF INTERNET TRAFFIC STREAMS Timothy Neame*

[18] The discrete Lindley distribution: properties and applications Emilio Gómez-Déniz a & Enrique Calderín-Ojeda a a Department of Quantitative Methods in Economics, University of Las Palmas de Gran Canaria, 35017, Las Palmas de G.C., Spain

[19] H.S. Bakouch, M.A. Jazi, S. Nadarajah

A new discrete distribution

Statistics, 48 (1) (2014), pp. 200-240

[20] A discrete lindley distribution with applications in biological sciences

Berhane Abebe, Rama Shanker Department of Statistics, College of Science, Eritrea Institute of Technology, Eritrea

[21] Matrix Mittag–Leffler distributions and modeling heavy-tailed risks Hansjörg Albrecher, Martin Bladt & Mogens Bladt

[22] Discrete Weibull regression model for count data Hadeel Saleh Kalktawi

[23] Why the Mittag-Leffler Function Can Be Considered the Queen Function of the Fractional Calculus? Francesco Mainardi Dipartimento di Fisica e Astronomia, Università di Bologna

[24] A COUNT MODEL BASED ON MITTAG-LEFFLER INTERARRIVAL TIMES K.K. Jose, B. Abraham

# APPENDIX

## DEFINITIONS OF TERMS USED IN REPORT

- **Mean**

  The **mean** is the average or the most common value in a collection of numbers. In **statistics**, it is a measure of central tendency of a probability distribution along median and mode. It is also referred to as an expected value.

- **Variance**

  Variance measures variability from the average or mean. It is calculated by taking the differences between each number in the data set and the mean, then squaring the differences to make them positive, and finally dividing the sum of the squares by the number of values in the data set.

- **Standard Deviation**

  A standard deviation is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance. The standard deviation is calculated as the square root of variance by determining each data point's deviation relative to the mean. If the data points are further from the mean, there is a higher deviation within the data set; thus, the more spread out the data, the higher the standard deviation.

- **Co-Variates**

  Covariates are characteristics (excluding the actual treatment) of the participants in an experiment. A covariate can be an independent variable (i.e. of direct interest) or it can be an unwanted, confounding variable. Adding a covariate to a model can increase the accuracy of your results.

- **Co-variance**

**Covariance** measures the total **variation** of two random variables from their expected values.

- **Intercept and Intercept only models**

The intercept in a multiple regression model is the mean for the response when all of the explanatory variables take on the value 0.

The regression constant is also known as the **intercept** thus, regression **models** without predictors are also known as **intercepting only models.**

- **Quantile**

A **quantile** defines a particular part of a data set, i.e. a **quantile** determines how many values in a distribution are above or below a certain limit. Special **quantiles** are the quartile (quarter), the quintile (fifth) and percentiles (hundredth).

- **Outlier**

An **outlier** is an observation that lies an abnormal distance from other values **in** a random sample from a population.

- **Probability Mass Function**

A **probability mass function** (pmf) is a **function** over the sample space of a discrete random variable X which gives the **probability** that X is equal to a certain value.

- **Hazard Function**

The hazard function (also known as the failure rate, hazard rate, or force of mortality) $h(x)$ is the ratio of the probability density function $P(x)$ to the survival function $S(x)$, given by

$$h(x) = \frac{P(x)}{S(x)} \tag{1}$$

$$= \frac{P(x)}{1 - D(x)}, \tag{2}$$

where $D(x)$ is the distribution function .

- **Closed Form**

  An equation is said to be a closed-form solution if it solves a given problem in terms of functions and mathematical operations from a given generally-accepted set. For example, an infinite sum would generally not be considered closed-form.