**Instituto Tecnológico y de Estudios Superiores de Monterrey**

Campus Ciudad de México

**Project 3**

**Modeling of Internet Traffic**

Random Processes

PhD. César Vargas Rosales


| | |
|---|---|
| Alfredo Zhu Chen | \|A01651980 |
| Luis Arturo Dan Fong | \|A01650672 |
| Juan Pablo Valverde López | \|A01656127 |
| Andrés Islas Bravo | \|A01339391 |


Miércoles 2 de junio de 2021

**Part 1 Fundamentals:**

1. **Provide the equations or formulas of the cumulative distribution functions (CDF) and the probability density functions (PDF), along with mean value (expectation), and the variance of the following random variables.**

   **Gaussian or Normal**

A Gaussian random variable $X$ with mean $m$ and standard deviation $\sigma$ is denoted as $X \sim N\ (m, \sigma)$ and has a PDF, mean value, variance and CDF defined as follows:

**PDF:**

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

**Expectation-mean value:**

$$E(x) = m = \int_{-\infty}^{\infty} x\, f_X(x) dx$$

**Variance:**

$$VAR(X) = E[(X-m)^2] = \sigma^2 = \int_{-\infty}^{\infty} (x-m)^2 f_X(x) dx$$

**CDF**

$$F_X(x) = \int_{-\infty}^{x} f_X(\alpha) d\alpha$$

$$F_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(\alpha-m)^2}{2\sigma^2}} d\alpha$$

By using normalization, it is needed to change the variable $\frac{(x-m)}{\sigma} = z$to obtain.

$$F_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\frac{(x-m)}{\sigma}} e^{-\frac{z^2}{2}} \sigma\, dz$$

$$F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{(x-m)}{\sigma}} e^{-\frac{z^2}{2}} dz$$

$$F_X(x) = \Phi_X(\frac{x-m}{\sigma})$$

To evaluate the probabilities of a Gaussian random variable, the $Q$ function is used, defined as:

$$Q(x) = 1 - \Phi_X = P(X > x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-\frac{z^2}{2}} dz$$

**Lognormal**

A Lognormal random variable $X$ with mean $m$ and standard deviation $\sigma$ is denoted as $X \sim log\ N\ (m, \sigma)$ and has a PDF, mean value, variance and CDF defined as follows:

**PDF**

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(ln(x)-m)^2}{2\sigma^2}}$$

**Expectation-mean value:**

$$E(x) = e^{m+\frac{\sigma^2}{2}}$$

**Variance**

$$VAR(x) = (e^{\sigma^2} - 1)e^{2m+\sigma^2}$$

**CDF**

$$F_X(x) = \int_{-\infty}^{x} f_X(\alpha) d\alpha$$

$$F_X(x) = P(X \geq x) = 1 - P(X < x)$$

$$F_X(x) = \Phi(\frac{ln(x)-m}{\sigma})$$

Where $\sigma$ is the scale parameter and $\Phi$ is the cumulative distribution function of the Normal Distribution.

**Uniform**

A random variable *X* is uniformly distributed in the interval *[a, b)*, such that $X \sim U[a, b)$. Its PDF, mean value, variance and CDF are defined as follows:

**PDF**:

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & x < a \; and \; x > b \end{cases}$$

**Expectation-mean value:**

$$E(X) = \frac{a+b}{\boldsymbol{2}}$$

**Variance:**

$$VAR(X) = \frac{(b-a)^2}{\boldsymbol{12}}$$

**CDF:**

$$F_X(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \le x \le b \\ 1, & x > b \end{cases}$$

**Exponential**

A random variable *X* is exponentially distributed in the interval $X \sim exp(\lambda)$. Where λ is a positive integer parameter. Its PDF, mean value, variance and CDF are given by:

**PDF:1**

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x < 0 \\ 0, & x \ge 0 \end{cases}$$

**Expectation-mean value:**

$$E(X) = \frac{\boldsymbol{1}}{\lambda}$$

**Variance:**

$$VAR(X) = \frac{\boldsymbol{1}}{\lambda^2}$$

**1CDF:**

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\lambda x}, & x \ge 0 \end{cases}$$

**Pareto** A random variable $X$ is pareto distributed if $X \sim Pareto(\alpha, x_m)$. Where $x_m$ (scale parameter) is the minimum possible value of $X$, being necessarily positive and $\alpha$is the shape parameter also positive.

**PDF:**

$$f_X(x) = \begin{cases} \frac{\alpha x_m^{\alpha}}{x^{\alpha+1}}, & x \geq x_m \\ 0, & x < x_m \end{cases}$$

**Expectation-mean value:**

$$E(X) = \frac{\alpha x_m}{\alpha - 1}, \ \alpha > 1$$

**Variance**

$$E(X) = \left(\frac{x_m}{\alpha - 1}\right)^2 \frac{\alpha}{\alpha - 2}, \ \alpha > 2$$

**CDF:**

$$F_X(x) = \begin{cases} 1 - \left(\frac{x_m}{x}\right)^{\alpha}, & x \geq x_m \\ 0, & x < x_m \end{cases}$$

**Weibull**

**PDF:**

$$f_X(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Where $k > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter of the distribution.

**Expectation-mean value:**

$$E(X) = \lambda \, \Gamma(1 + \frac{1}{k})$$

Where $\Gamma$ is the gamma function defined as follows

$$\Gamma(\beta) = \int_0^{\infty} t^{\beta-1} e^{-t} dt = (\beta - 1)\Gamma(\beta - 1)$$

**Variance:**

$$VAR(X) = \lambda^2 \left[ \Gamma(1 + \frac{2}{k}) - (\Gamma(1 + \frac{1}{k}))^2 \right]$$

**CDF:**

$$F_X(x) = \begin{cases} 1 - e^{-\left(\frac{x}{\lambda}\right)^k}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

2. **Write the fundamental ideas about the mean excess value and the survival functions, and obtain or provide the formulas for the same random variables and provide plots of the mean excess value and survival functions of the random variables given.**

**Mean excess value:**

The distribution of excess over a threshold u for a random variable *X* with distribution function *F* is defined as follows [1]:

$$F_u(x) = P[X - u \leq x | X > u]$$

It is used for peak over threshold(POT) modeling which fits appropriate distributions to data on excesses. POT modeling is based on the generalized Pareto class of distributions and it is appropriate for it to be used to describe statistical properties of excess. A random variable *X* is said to have a Generalized Pareto Distribution(GPD) if its CDF has the following form [1]:

$$G_{\xi,\beta}(x) = \begin{cases} 1 - (1 + \xi x/\beta)^{-1/\xi} & \text{if } \xi \neq 0 \\ 1 - \exp(-x/\beta) & \text{if } \xi = 0 \end{cases}$$

where $\beta > 0$, and $x \geq 0$ when $\xi \geq 0$ and $0 \leq x \leq -\beta/\xi$ if $\xi < 0$. The parameters $\xi$ and $\beta$ are referred to as the shape and scale parameters respectively. For a Pareto distribution, the tail index $\alpha$ is just the reciprocal of $\xi$ when $\xi > 0$. A special case is when $\xi = 0$ and in this case the GPD is the same as the exponential distribution with mean β[1].

For a large class of distributions, the excess distribution $F_u$ is asymptotically equivalent to a GPD law $G_{\xi,\beta}(u)$, as the threshold $u$ approaches the right end point of the distribution *F*. The threshold $u$, where the GPD model provides a suitable approximation to $F_u$, is critical in applications. Once having this said, the mean excess(ME) function is a tool popularly used to determine the threshold $u$ and the adequate GPD model in practice. For a random variable *X* satisfying $EX^+ < \infty$with CDF of *F(x)* with right endpoint $x_F$ and tail $\underline{F}(x) = 1 - F(x)$, the mean function is [1]:

$$M(u) := E[X - u | X > u] = \frac{\int_u^{X_F} \underline{F}(s)ds}{\underline{F}(u)} \quad u < x_F$$

Given an independent and identically distributed set of samples $X_1, \ldots, X_n$ from *F(x)*, a natural estimate of $M(u)$ is the empirical mean excess function $\widehat{M}(u)$is defined as follows [1]:

$$\widehat{M}(u) = \frac{\sum_{i=1}^{n} (X_i - u) I_{[X_i > u]}}{\sum_{i=1}^{n} I_{[X_i > u]}} \quad u \geq 0$$

The mean excess plot is an empirical graphical plot of the mean excess function $M(u)$, which is very helpful in the study of risk, insurance and extreme values. It is also a very common tool to understand the right tail behavior of a data set. If the mean exists, it assists in distinguishing light-tailed data sets from heavy-tailed ones [1][2][3].

**Survival functions**

Given the distribution function $F_X(x)$ of the random variable $X$, the probability that $X$ survives time duration $t$ is often called the survivor function, or the survival function in reliability theory. The survival function is equivalent to the complementary distribution function, and it can be shown as follows, which also holds for a discrete random variable as well.

$$S_X(x) = \underline{F}_X(x) = F^c_X(x) = 1 - F_X(x) \qquad x \geq 0$$

**Gaussian or Normal**

**Mean excess value**

$$M(u) = \frac{\int_u^{\infty} 1 - \Phi_X(\frac{s-m}{\sigma})ds}{1 - \Phi_X(\frac{u-m}{\sigma})}$$

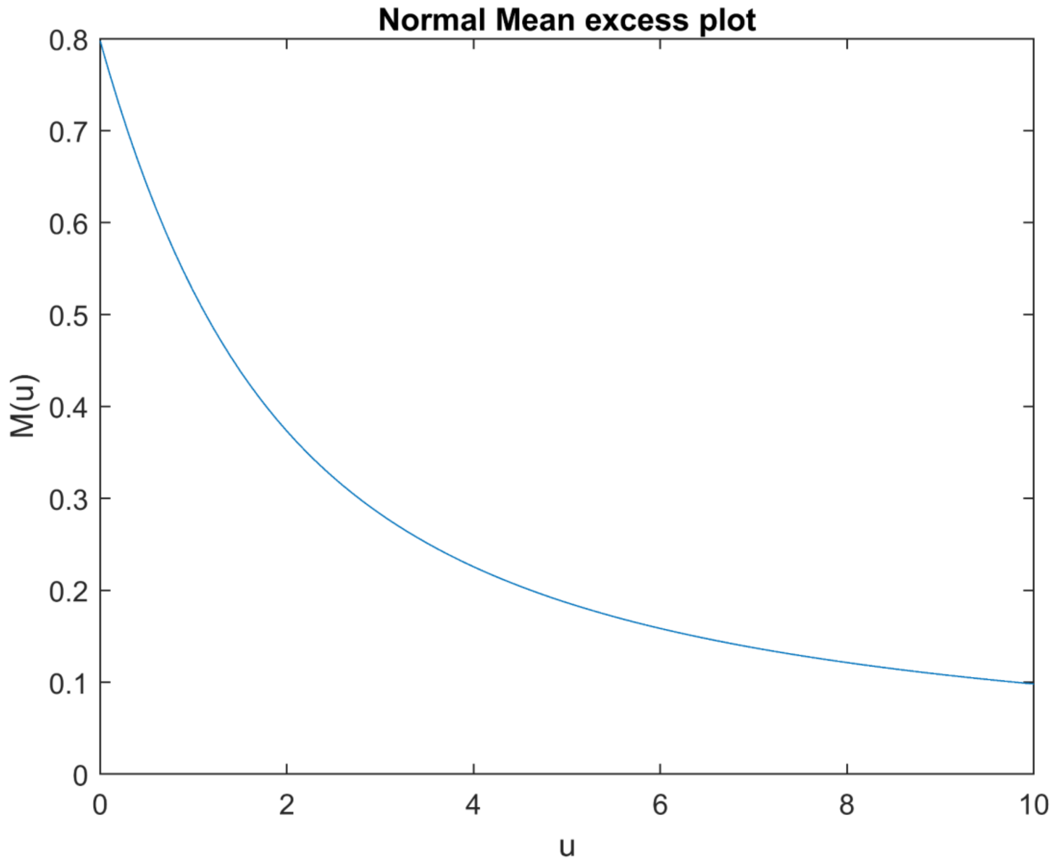


Figure. 1 ME Plot of Normal Distribution
For $m = 0$ and $\sigma = 1$

**Survival function**

$$S_X(x) = \underline{F}_X(x) = 1 - F_X(x) = 1 - \frac{1}{\sqrt{2\pi}\sigma}\int_x^{\infty} e^{-\frac{(\alpha-m)^2}{2\sigma^2}} d\alpha$$

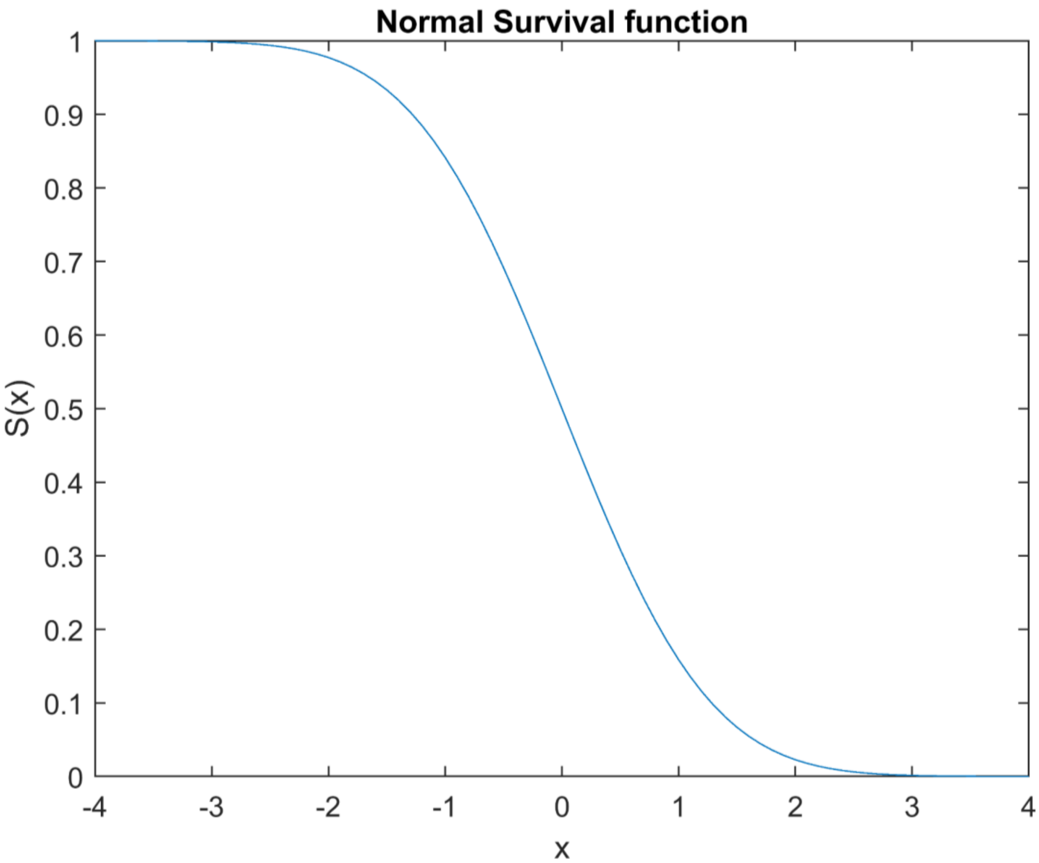


Figure. 2 Normal survival function

For $m = 0$ and $\sigma = 1$

**Lognormal**

**Mean excess value**

$$M(u) = \frac{\sigma^2 u}{ln(u) - m}(1 + o(1))$$

where o(1) stands for a term which tends to zero as $x \to \infty$

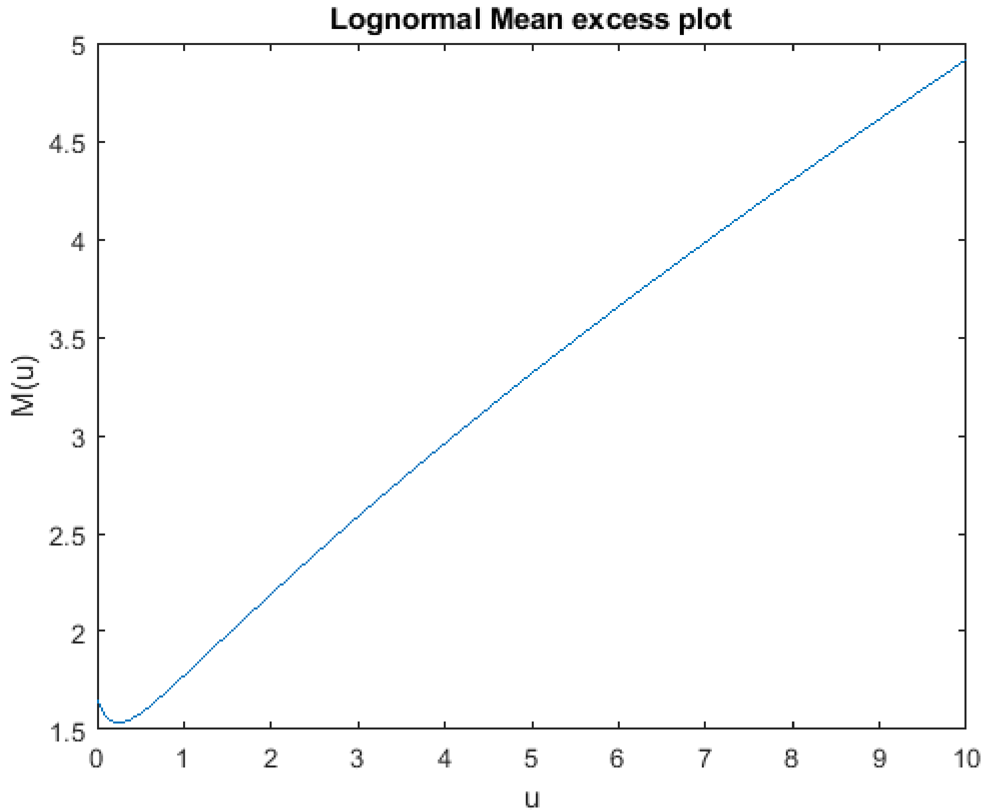


Figure. 3 ME Function of Lognormal distribution
For $m = 0$ and $\sigma = 1$

**Survival function**

$$S_X(x) = \underline{F}_X(x) = \boldsymbol{1} - F_X(x) = \boldsymbol{1} - \Phi\left(\frac{ln(x) - m}{\sigma}\right)$$

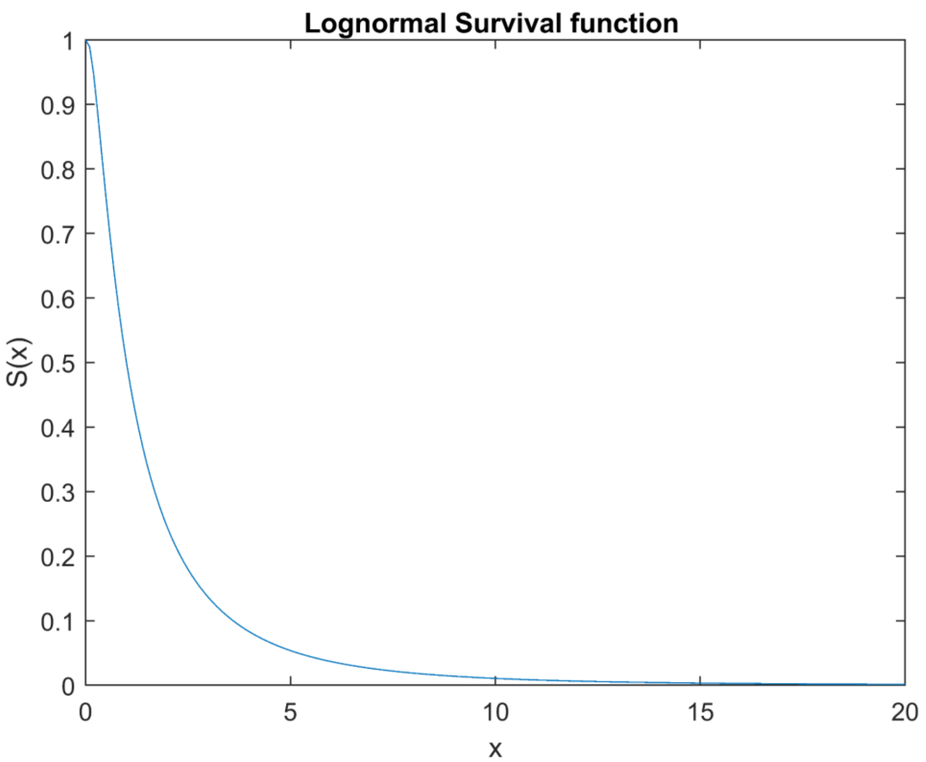


Figure 4. Survival Function of Lognormal Distribution
For $m = 0$ and $\sigma = 1$

## Uniform

### Mean excess value

$$M(u) = \frac{1}{2}(a + b) - u \quad u < a$$

$$M(u) = a - u - \frac{(a - b)^{2}}{2(u - b)} \; a \leq u \leq b$$

$$M(u) = 0 \; u > b$$

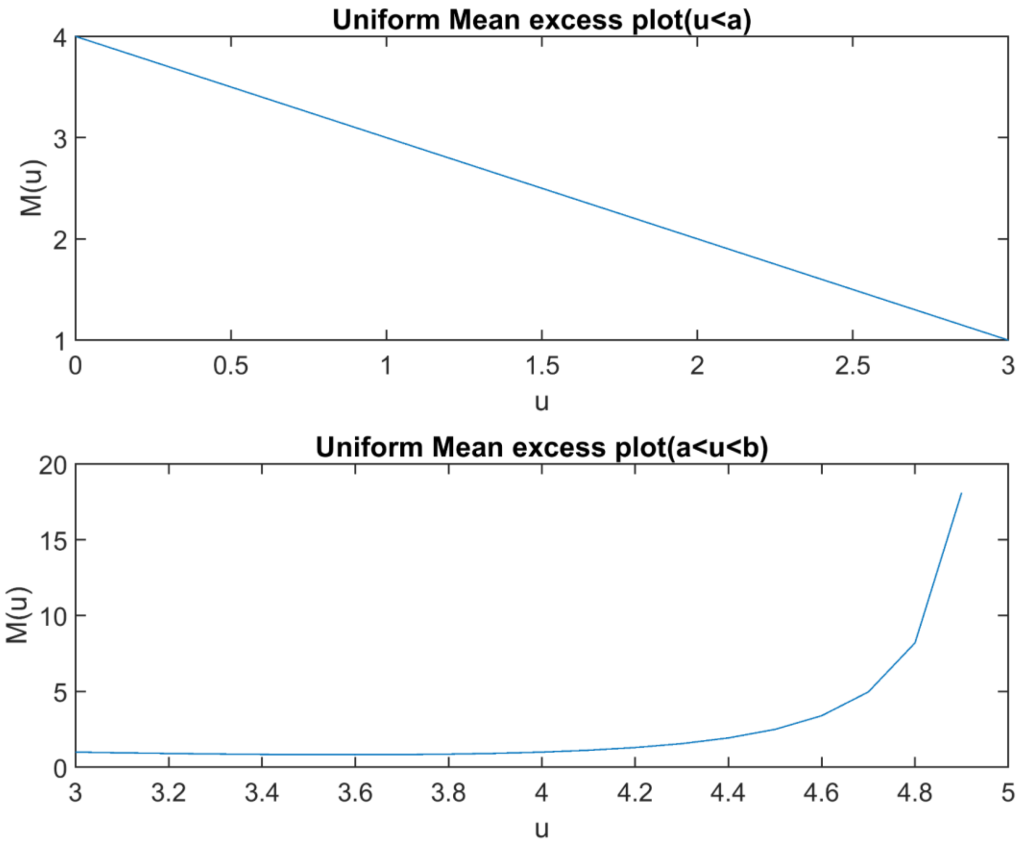


Figure 5. Uniform ME Function.
For $b = 5$ and $a = 3$

### Survival function

$$S_X(x) = \underline{F}_X(x) = 0 \quad x < a$$

$$S_X(x) = \underline{F}_X(x) = \frac{b - x}{b - a} \; a \leq x \leq b$$

$$S_X(x) = \underline{F}_X(x) = 1 \; x < b$$

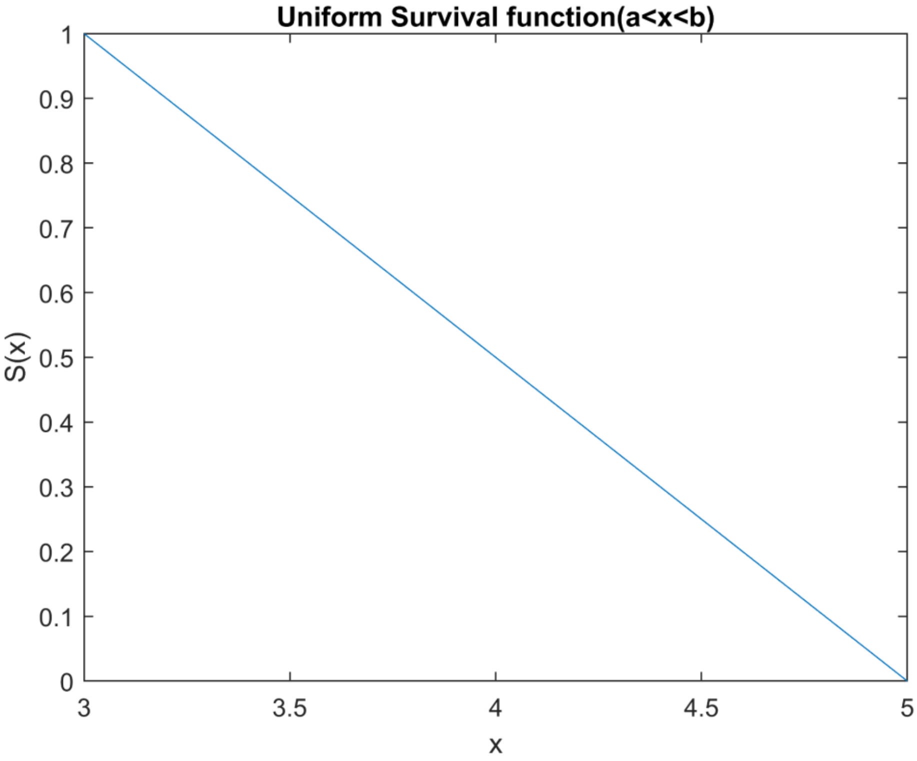


Figure 6. Uniform survival function
For $b = 5$ and $a = 3$

## Exponential

### Mean excess value

$$M(u) = \frac{1}{\lambda}$$

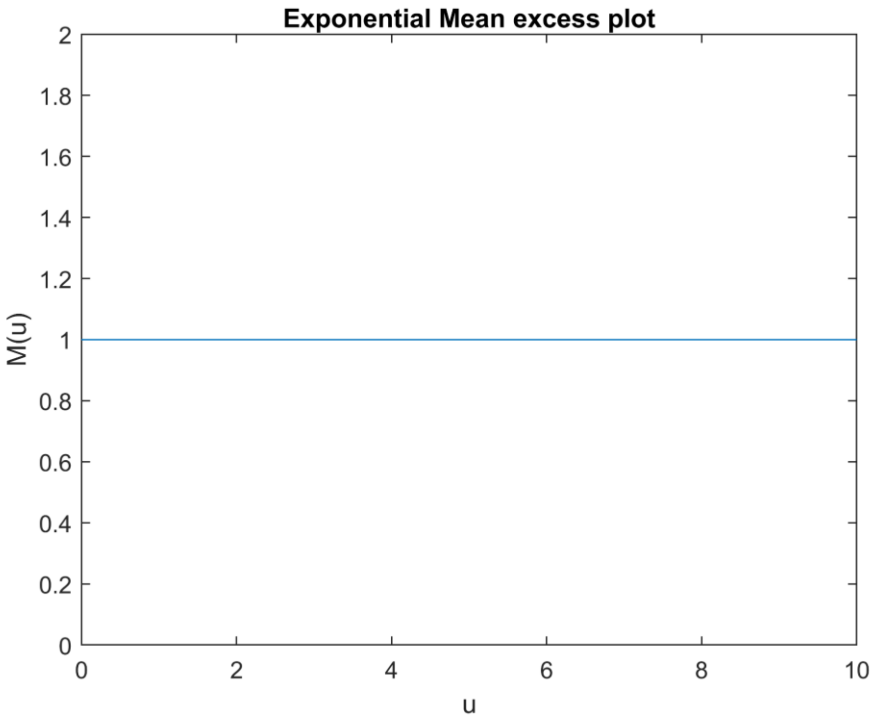


Figure 7. ME Function Exponential Distribution.
For $\lambda = 1$

### Survival function

$$S_X(x) = \underline{F}_X(x) = 1 - F_X(x) = e^{-\lambda x}, \qquad x \geq 0$$

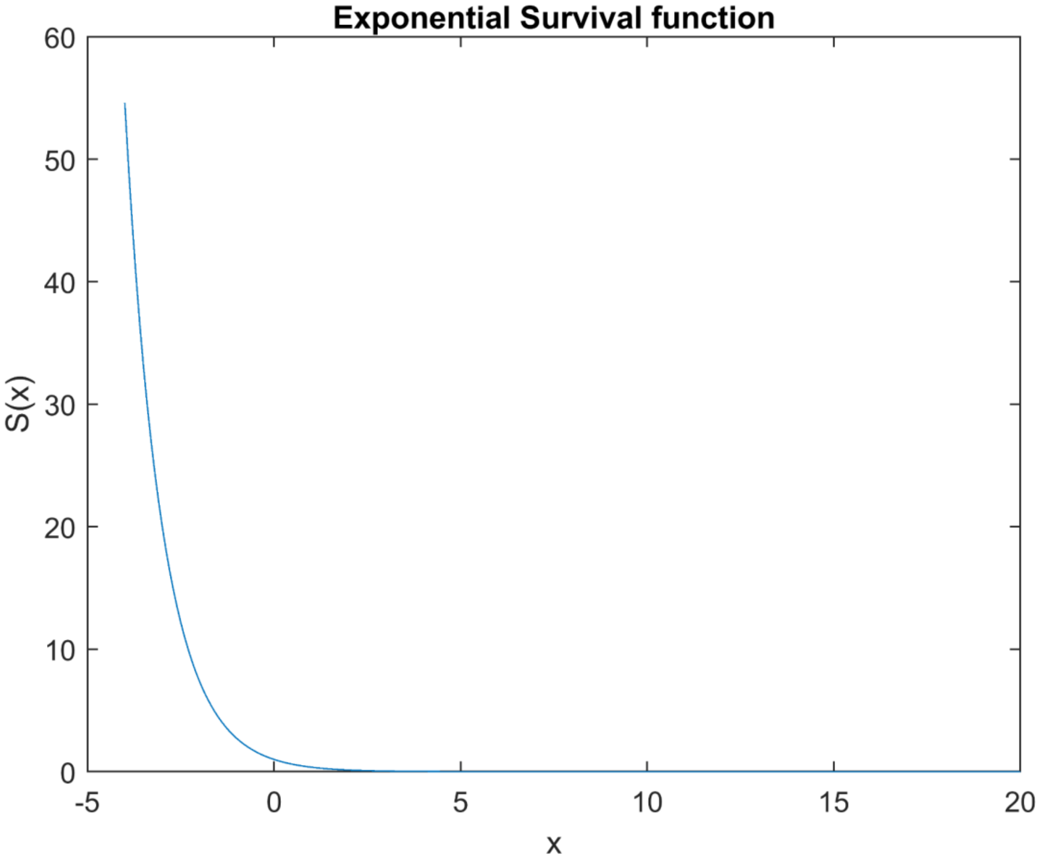


Figure 8. Survival Function Exponential Distribution.
For $\lambda = 1$

## Pareto

### Mean excess value

$$M(u) = \frac{\int_u^{\infty} 1 - S_X(s)ds}{1 - S_X(u)} \quad \alpha > 1$$

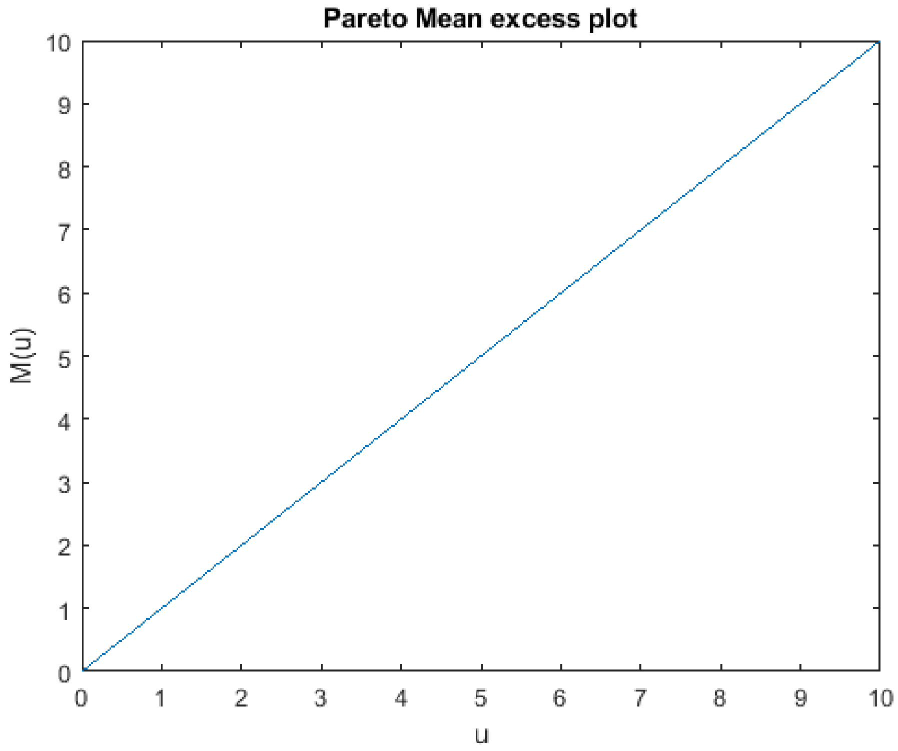


Figure 9. ME Function Pareto Distribution.
For $\alpha = 2$ and $x_m = 4$

**Survival function**

$$\bar{F}_X(x) = P(X > x) = \begin{cases} \left(\frac{x_m}{x}\right)^{\alpha}, & x \geq x_m \\ 1, & x < x_m \end{cases}$$

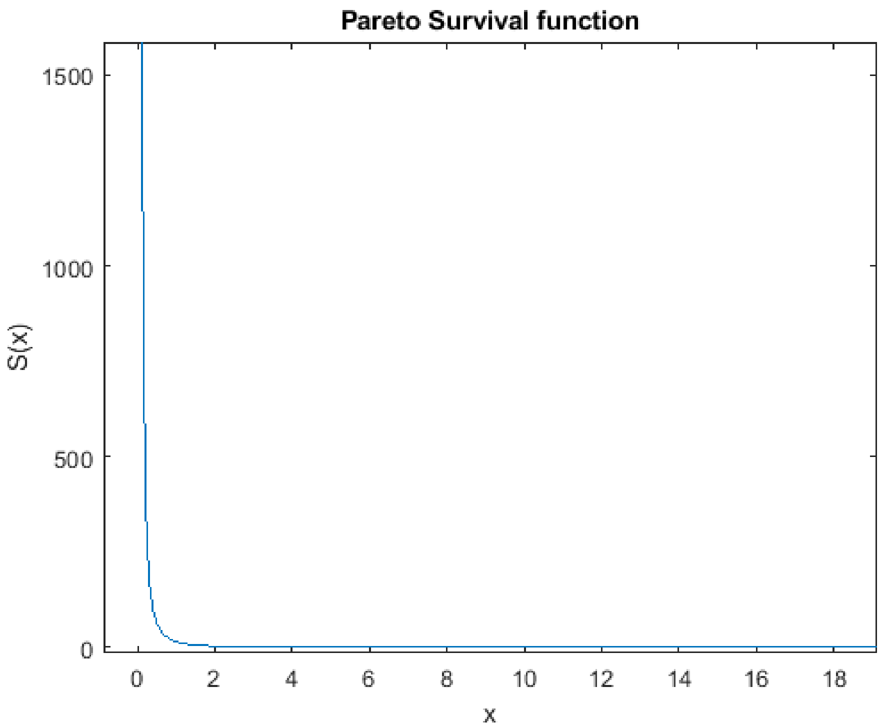


Figure 10. Survival Function Pareto Distribution
For $\alpha = 2$ and $x_m = 4$

**Weibull**

**Mean excess value**

$$M(u) = \frac{\lambda^{k} u^{1-k}}{k}(1 + o(1))$$

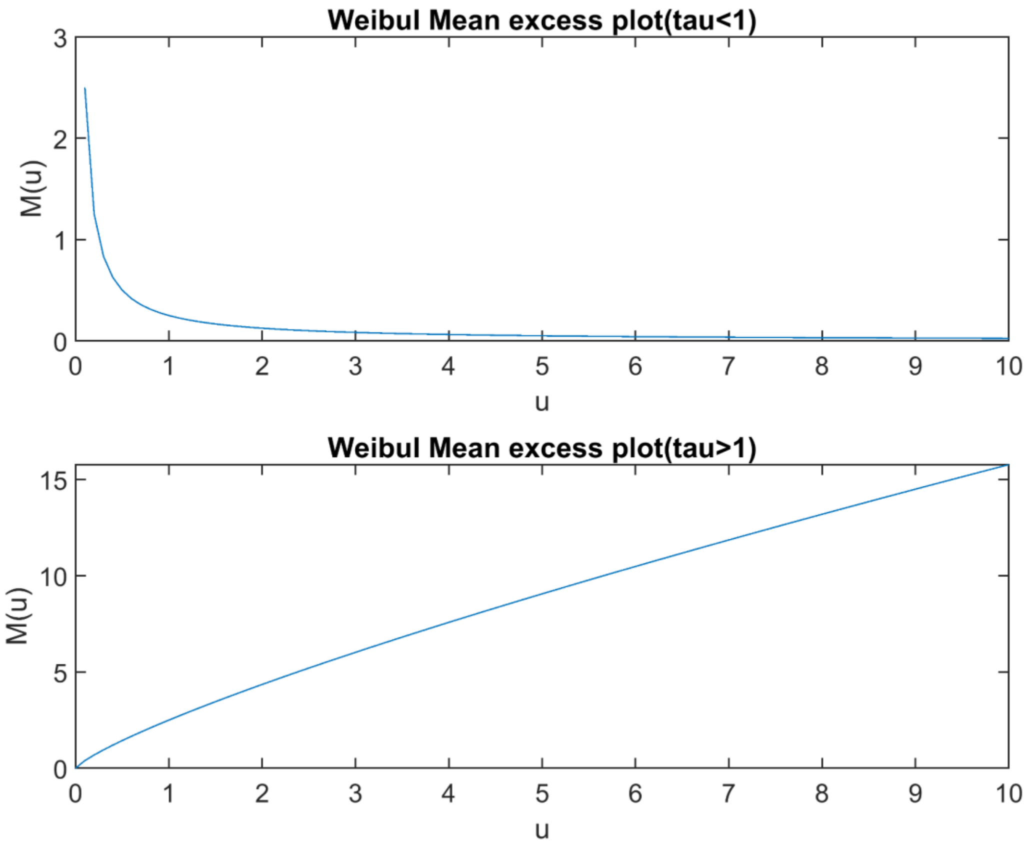


Figure 11. ME Function Weibull Distribution.
For $\frac{k}{\lambda} = 2$ and $k = tau$

**Survival function** Its complementary cumulative distribution function is a stretched exponential function.

$$\underline{F}_X(x) = e^{-x\beta}$$

Stretching exponent $\beta$ between 0 and 1.

$$S_X(x) = \underline{F}_X(x) = 1 - F_X(x) = e^{-(\frac{x}{\lambda})^{k}}$$

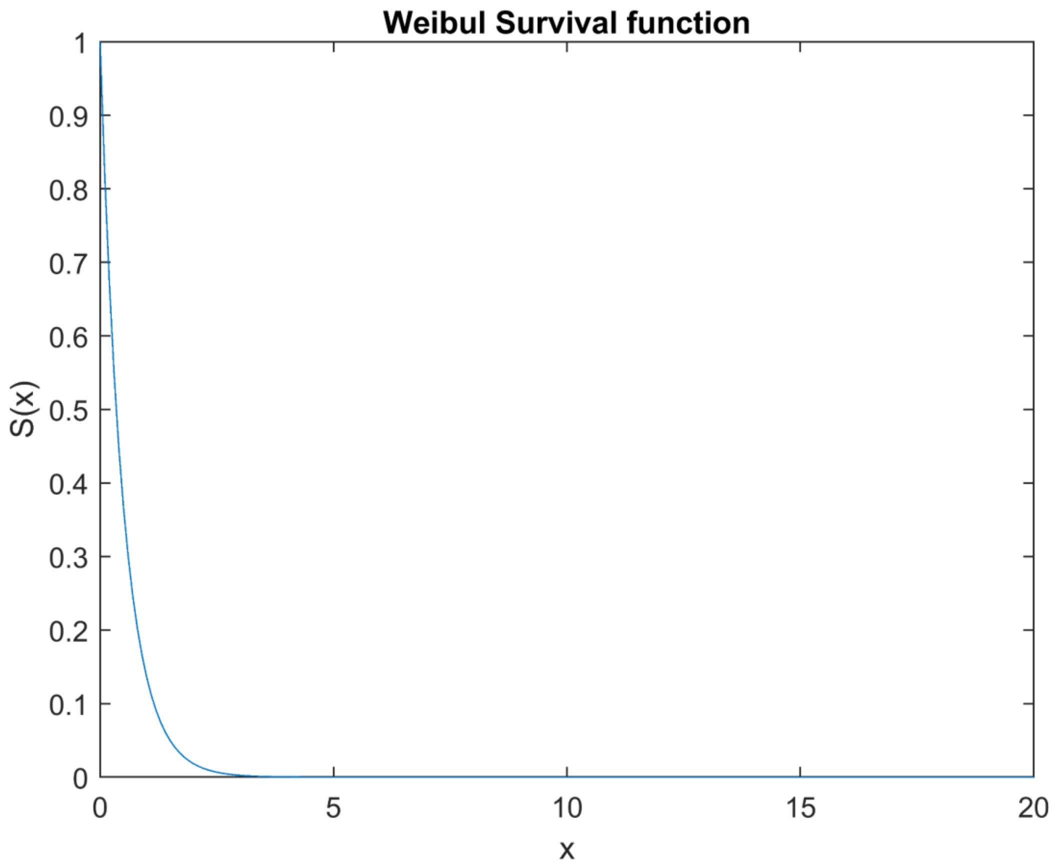


Figure. 12. Survival Function Weibull Distribution.
For $\frac{k}{\lambda} = 2$ and $k = tau$

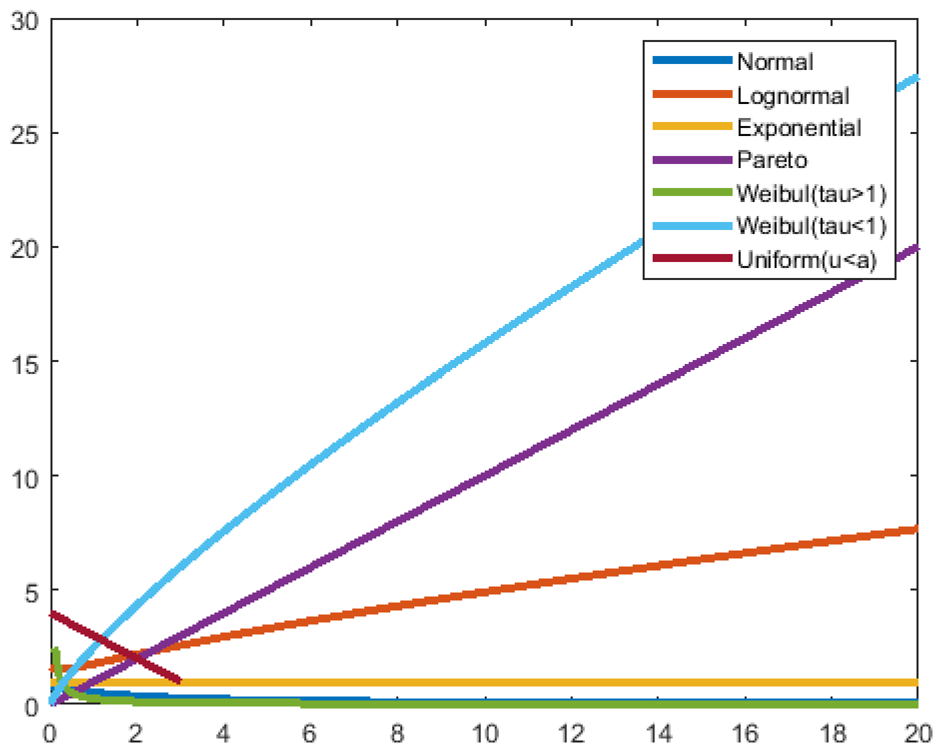


Figure 13. ME Functions of Normal, Lognormal, Exponential, Pareto, Weibull and Uniform Distributions.

3. **Provide a description of two of the models of the list. Your report in this part must not be longer than four pages for each model.**

**Poisson regression**

Poisson regression is similar to the regular multiple regression, except that the dependent variable ($X$) is an observed count that follows the Poisson distribution. Meaning that the possible values of $X$are the positive integers. One example of the usage of a Poisson regression, is the number of failures a machine can suffer at various operating conditions, therefore, it has become a popular method for developing regression models for counts.

Using the Poisson regression model, requires making assumptions such as:

1. Poisson response: The response variable is a count per unit of time or space, described by a Poisson distribution.
2. Independence: The observations must be independent of one another.
3. Mean: By definition, the mean of a Poisson random variable must be equal to its variance.
4. Linearity: The log of the mean rate, log($\lambda$) must be a linear function of x.

The Poisson regression uses as its core the Probability Mass Function (PMF) of the Poisson Distribution.

$$P_X(k) = P(X = k|\lambda) = \frac{e^{-(\lambda t)}(\lambda t)^k}{k!}$$

If for the Poisson regression, the variable $\lambda$ is considered to be a constant, a modified Mean Model can be used for predicting future counts of events while using the previous equation. In case the Poisson regression model, uses a $\lambda$which is not a constant and therefore can vary from one observation to another. Usually the variable $\lambda$ is given by a vector of explanatory variables also known as regressors or regression variables. The objective of the regression model is to fit the observed counts y to the matrix of regression values X via a link-function that expresses the rate vector $\lambda$ as a function of the regression coefficients $\beta$ and the regression matrix X.

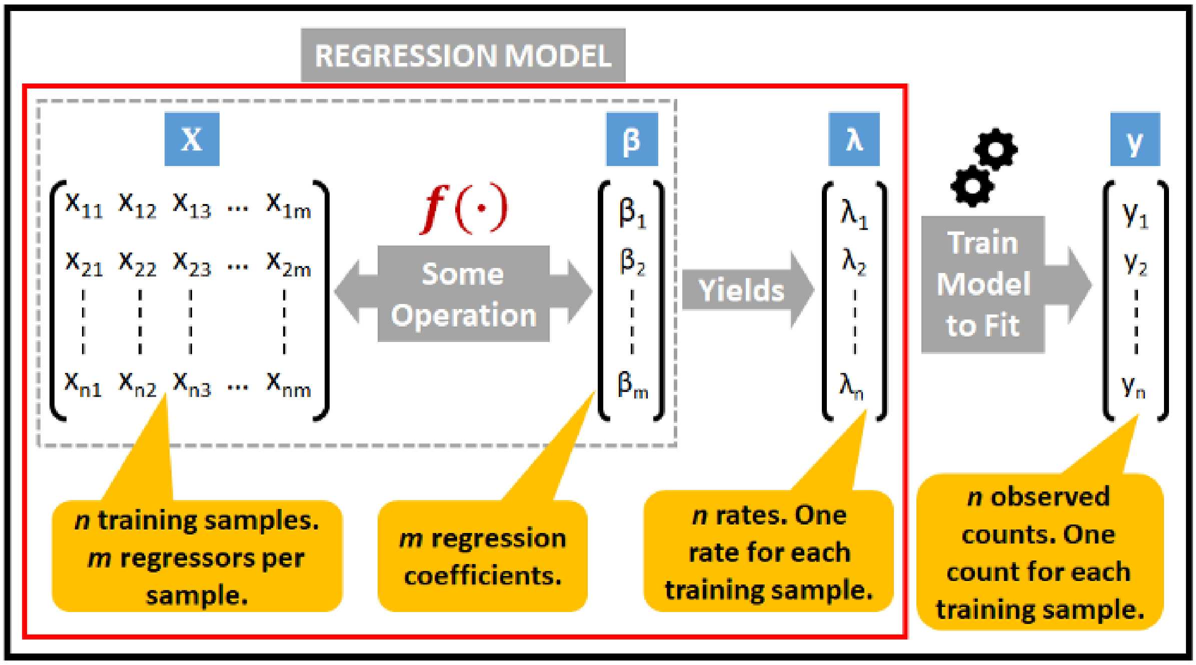

Figure 14 Poisson Regression Model [6]

For each observation $y_i$ with index $i$ of the dataset corresponding to the regression variable $X_i$, the probability of observing the count $y_i$ is as the following PMF:

$$PMF\ (y_i|x_i) = \frac{e^{-(\lambda_i)}(\lambda_i)^{y_i}}{y_i!}$$

The link-function keeps the parameter $\lambda$non-negative even when the regression values X or the coefficients $\beta$ have negative values, this is important to mention as it is a requirement for count based data.

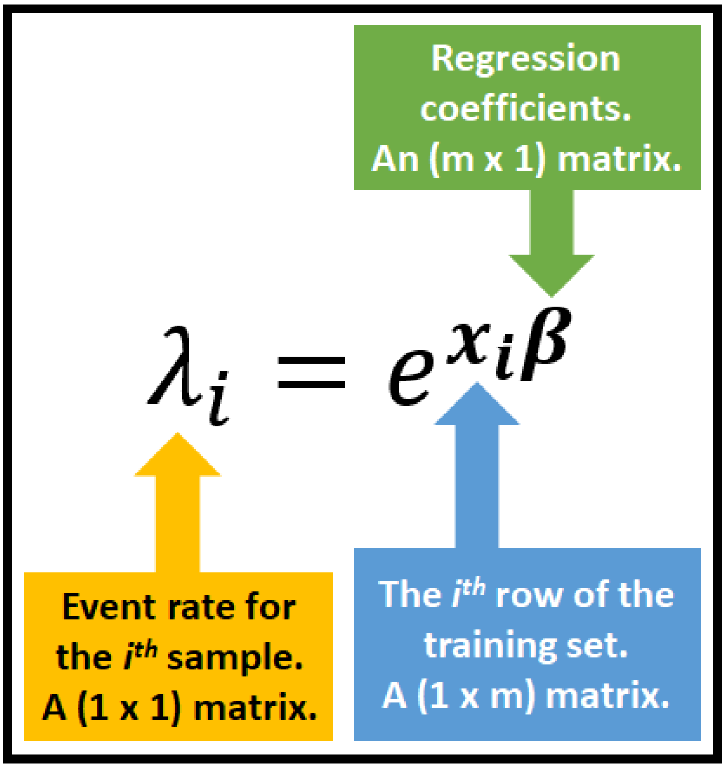


Figure 15. Poisson Regression Model [6]

In order to obtain the coefficients $\beta$, a technique such as Maximum Likelihood Estimation(MLE), which yields to the following equation that gives the coefficients $\beta$An Iterative Reweighted Least Squares(IRLS) is commonly used to solve this equation instead of solving by hand.

$$\sum_{i=1}^{n} (y_i - e^{x_i\beta})x_i = 0$$

**Leaky bucket**

The leaky bucket algorithm is used mainly as a tool for regulating the traffic flow of data, even in the presence of a fast burst of data. It uses the data rate for controlling the maximum speed of data coming from a specific source. In case the data rate $\lambda_a$ is lower than the one specified in the algorithm $\lambda_b$ ,the data is accepted; however, if the data rate coming from the source is higher than the one specified in the algorithm, the data is transferred at the maximum speed that the algorithm is allowed to transmit while the remaining data is stored momentarily in the buffer; in case that the buffer is full, the excess data is discarded.

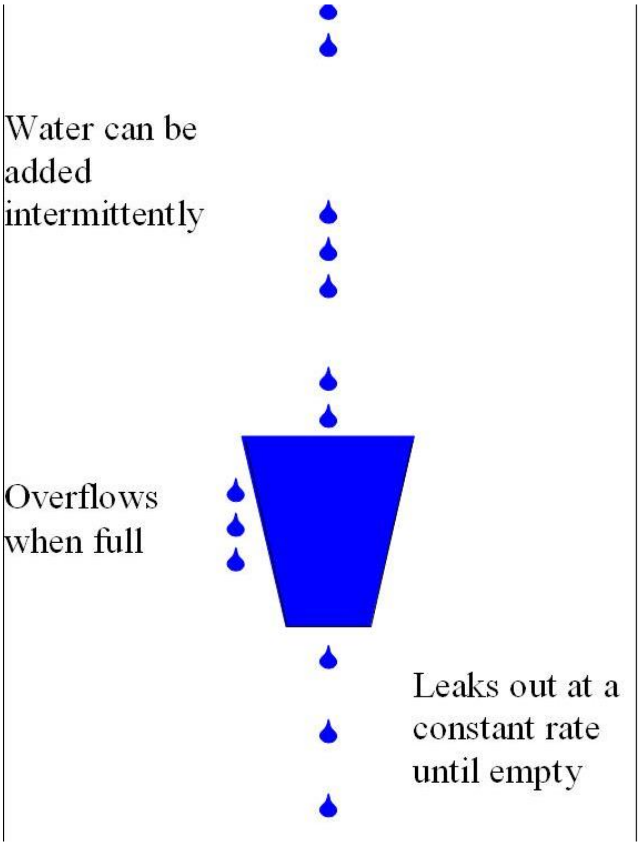


Figure 16. Leaky Bucket Model [5]

The algorithm uses the following variables:

- $\lambda_a$: *Input data rate*
- $\lambda_a = N/t$
- $N$: *Number of packets*
- $t$: *time*
- $\lambda_b$: *Maximum data rate or speed*

The algorithm can present the following scenarios according to the $\lambda_a$ and $\lambda_b$:

a) $\lambda_a < \lambda_b$: the input data rate is lower than the maximum data rate or speed permitted in the algorithm, therefore the data is accepted.
b) $\lambda_a > \lambda_b$: the input data rate is higher than the maximum data rate or speed permitted in the algorithm, therefore the data is transmitted at a $\lambda_b$ rate, while the rest of the data is stored in the buffer.
c) $\lambda_a > \lambda_b$: the buffer is full and the input data rate is higher than the maximum data rate or speed permitted in the algorithm, therefore the data is transmitted at a $\lambda_b$ rate, while the rest of the data is stored in the buffer.

## Part 2: Experimental Work

## OBJECTIVES

1. To find the best distributions that fit data of inter-packet arrival times in Internet traffic.
2. To understand the concepts of mean excess delay, survival function and heavy tails.

## MATERIAL AND METHODS

- 1 PC or laptop with at least 8GB RAM.
- MATLAB R2019a or later versions.
- Software (e.g., Wireshark) or similar.
- Excel to store the data.

Firstly it was required to install WireShark and use the packet capturing software to obtain bytes for us to turn them into bits per second (bps) using Matlab and the inter packet time (IPT) or interarrivals which is the time between each packet.

Once the desired information was gathered, the data was plotted in Matlab using *plot* function, providing a clearer and more graphical representation of both bps and IPT. Likewise, the CDF and PDF were plotted using the same function as in the case of the raw data, however both CDF and PDF were calculated before plotting using different tools. In the case of the CDF it was obtained using the function provided by the professor (*MyCDFplot2)*, while in the case of the PDF it was obtained using the *histogram* function with parameters that return the PDF.

The mean, median, max, min, std, var and dispersion values were obtained using the Matlab functions with the same name as the ones stated before in this paragraph.

As the survival functions requires a CDF for obtaining the desired data, it was required to firstly obtain the CDF of both bps and IPT. Once the CDF was obtained, the survival function was plotted using the data from the CDF and subtracting it from 1.

For performing a comparison between the exponential distribution for both bps and IPT, against their survival function; it was required to first obtain the exponential distribution of the bps and IPT. It was achieved by using the exponential CDF function where the lambda is the average of the traffic data.

The histograms from both the raw data from the bps and IPT were obtained using the Matlab function *histograms,* as it allowed us to interpret the information faster and easier than using a kernel.

For the mean excess value, a theoretical approach was used; therefore an approximation of the integral was necessary to obtain the sum of the area of rectangles which were under the curve.

Extra:

For simulating the interruption of service it was necessary to use a randomization of data to select different points in time where the data was turned into 0, simulating an interruption. Once the data was modified, the ON/OFF graph was plotted using *stem* as the data was discrete.

## RESULTS AND DISCUSSION

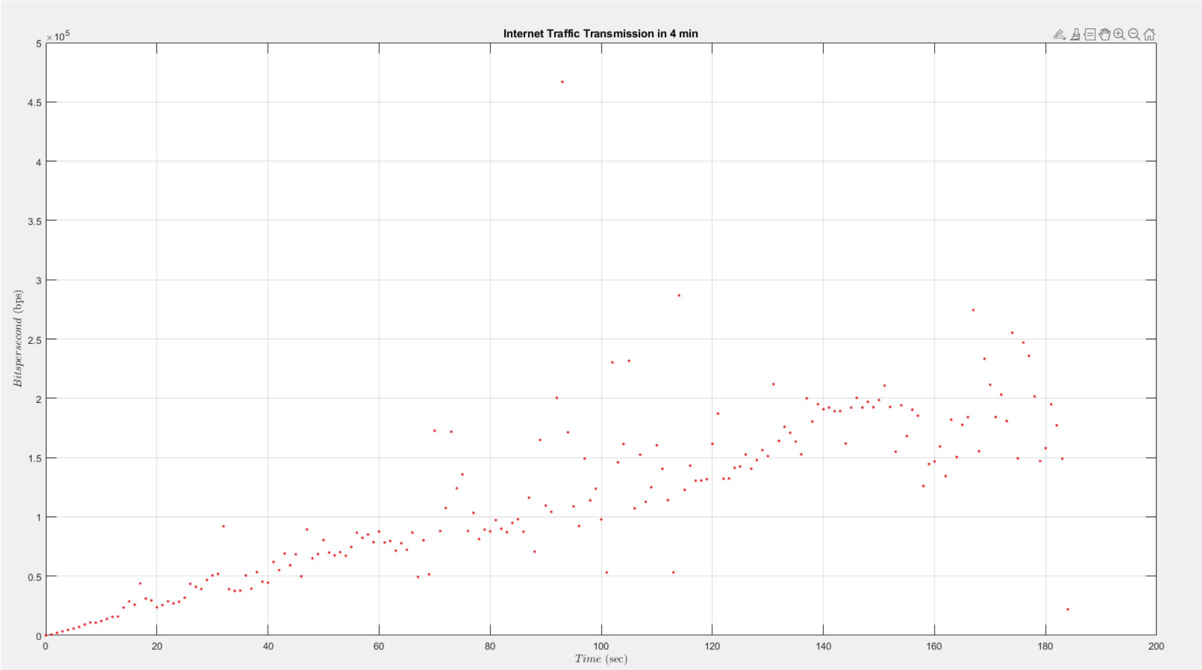


**Figure 1. Internet Traffic Transmission in 4 minutes.**

In figure 1 we can observe the raw data obtained from the analysis of the bytes information gathered by WireShark and then modified to obtain the bits per second (bps) that were required to obtain other information such as the pdf and cdf plots. It is worth noting that even if the data is well scattered throughout the time, the general behavior shown by the plot is increasingly lineal.

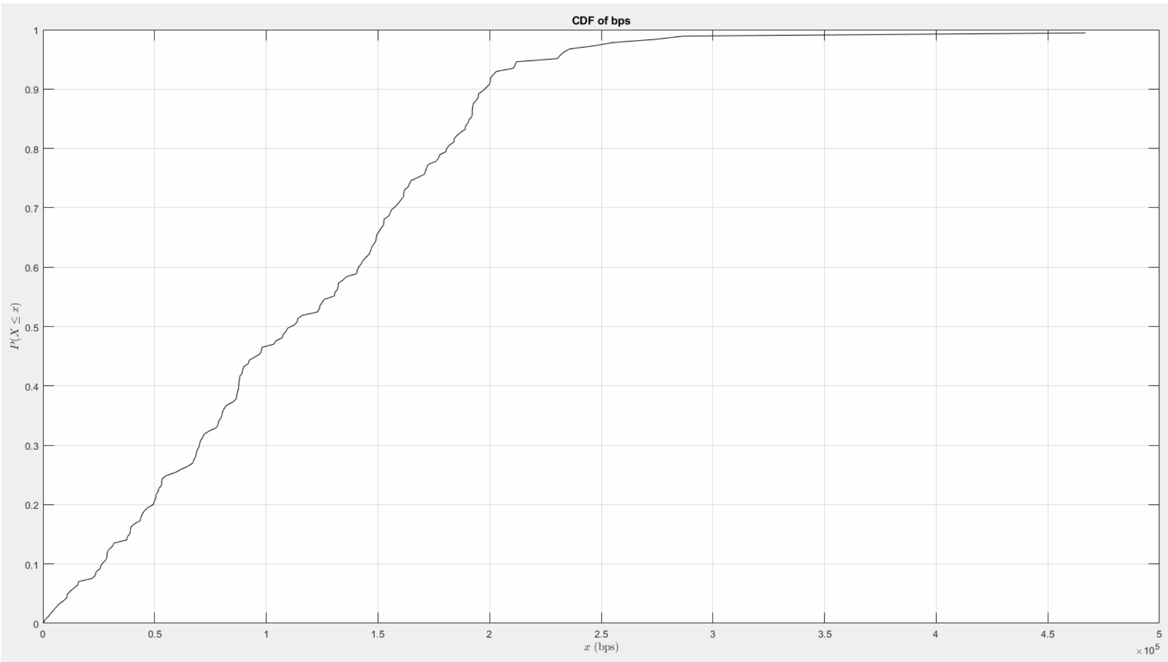


**Figure 2. CDF of bits per second (bps)**

The cumulative distribution function (CDF) calculates the cumulative probability for a given x-value. Use the CDF to determine the probability that a random observation that is taken from the essay will be less than or equal to a certain value.

In this case for characterization purposes Figure 2 and Figure 3 can be associated with theoretical Normal Distribution, where the outcome lies between certain bounds defined by parameters $a$ and $b$ (minimum and maximum values). For this study case the minimum value is close to 0.

Survival function is part of an analysis which contemplates the expected duration of time until one event occurs. In this case it's better to mention it as the reliability function because as the amount of packets increases in time the probability of completing its function drastically reduces, this depends on the modem and Internet Service Providers that are selling you a service and assuring you a certain amount of data to be transmitted without problems over the internet.

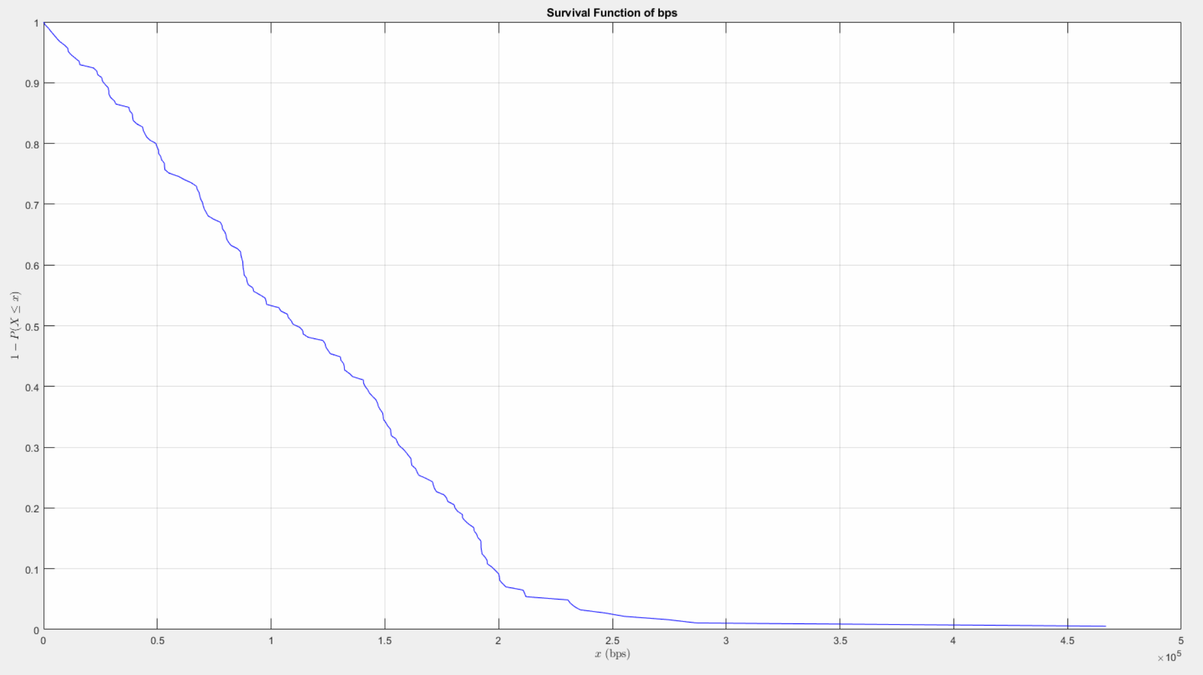


**Figure 3. Survival function of bps**

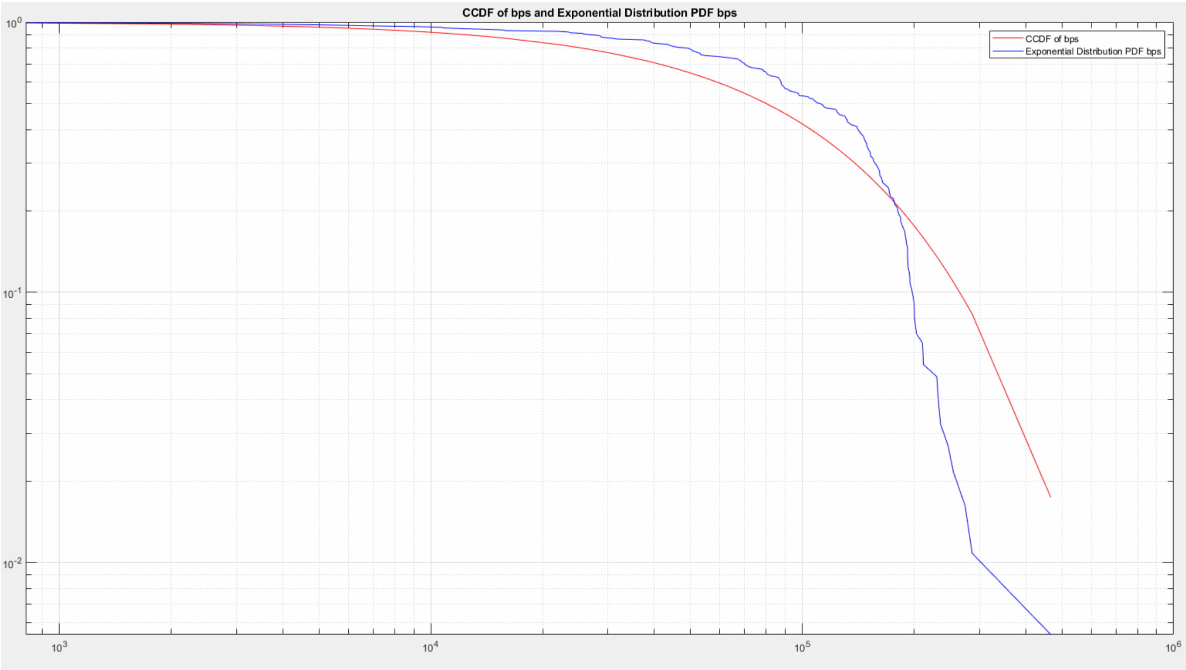


**Figure 4. Survival function comparison with Exponential PDF of bps**

The graphical comparison between both exponential distribution and CCDF of the bps shows that data given can be analyzed in both heavy tail and light tail, as the exponential distribution is in some parts above the CCDF and in other parts it is below the CCDF. However in network traffic analyzing the important information and more problematic analysis comes from the heavy tail distribution, meaning that there is a higher probability of network collapsing since heavy tail means that there is a larger probability of getting very large values (packets transmitted). The exponential distribution is taken as the comparison canon because it has a medium tail so any data that goes below it is considered a light tailed distribution and what goes above is considered as heavy tailed.

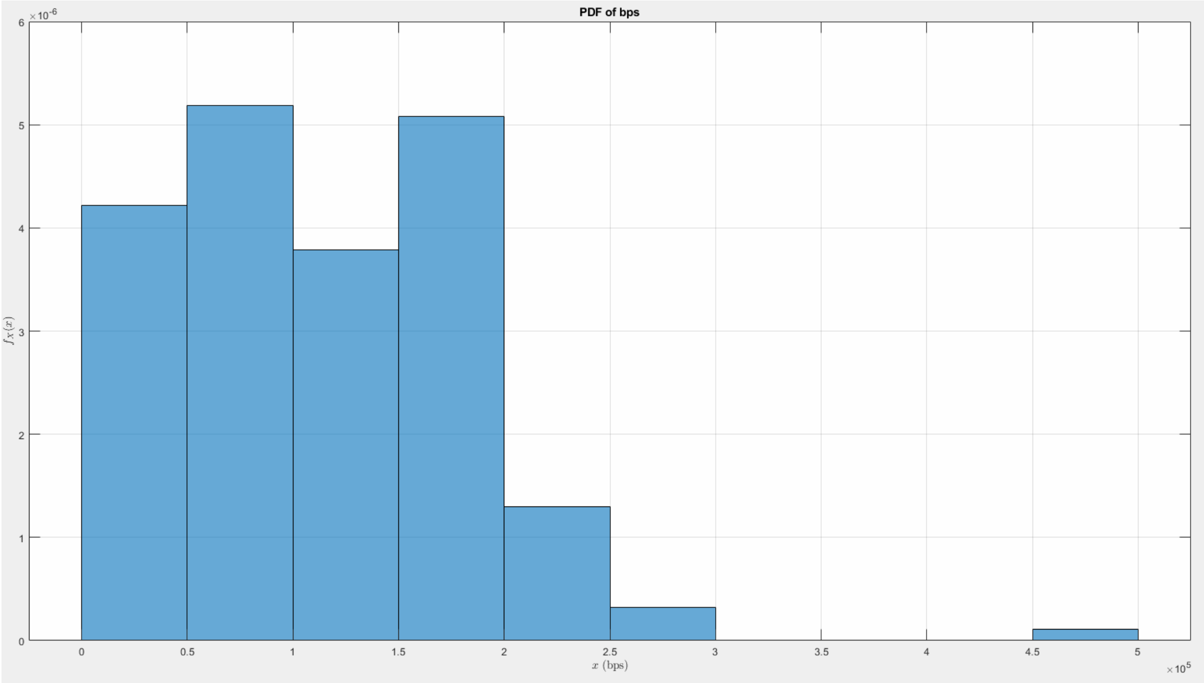


**Figure 5. PDF of bps**

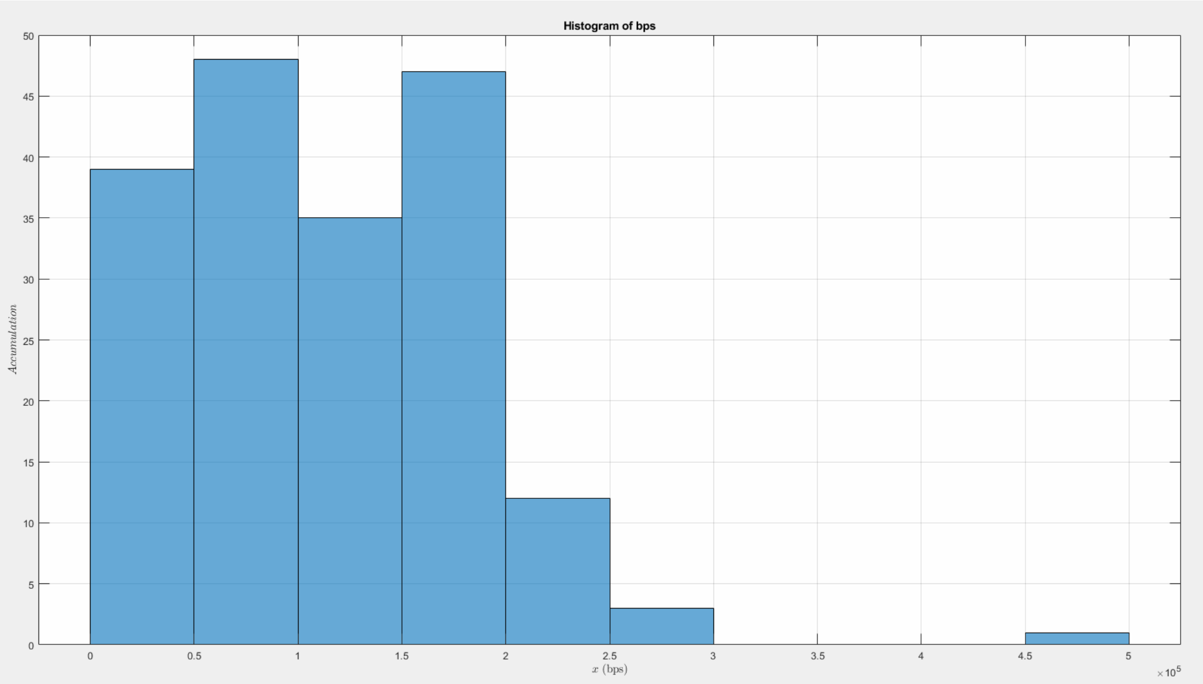


**Figure 6. Histogram of bps**

According to figure 6, after performing the histogram for the bps data, it can be seen that there is a bigger concentration of packets between 0.5 to 1 bit per second and 1.5 to 2 bits per second, meaning that the bps is reliable for browsing through the internet. The range of $y$ axis gives a description of the difference between the greatest and least values in a given data set. How widely dispersed are the frequencies of each bin, where extremely large frequency ranges may indicate a non reliable data. The shape of a histogram can lead to valuable conclusions about the trend of the data. There is remarkable outliner in the right part of the graph from $4.5\ x10^5$ to $5\ x10^5$ $[bps]$ were sent approximately 1 time of the total number of deliveries. It is also important to say that the analysis was conducted on a short time windows, four minutes is not sufficient to characterize a whole network, several test at different time must be conducted to assure a good characterization.

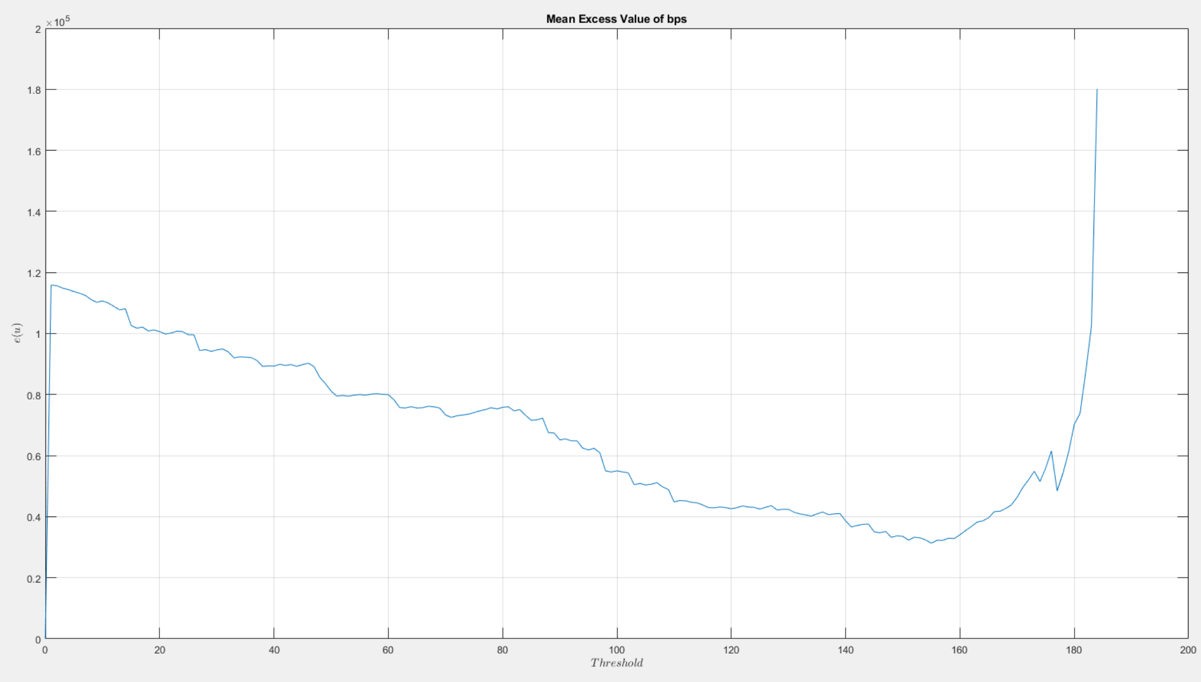


**Figure 7. ME of bps**

By the beginning of the graph the behavior is quite noisy as the samples taken into account started at the second coefficient, this consideration was a consequence of the matricial index. Even so, the rest of the graph works under the expectancy, having an interesting destabilisation when surpassing the 160 value of the threshold approximately. At this point, what seems to get more stable as the samples go through, now gets an exponential quality. This type of behavior was not expected at all, according to the Weibull distribution it must have kept a tendency of becoming more stable as it approached 0 on the vertical axes.

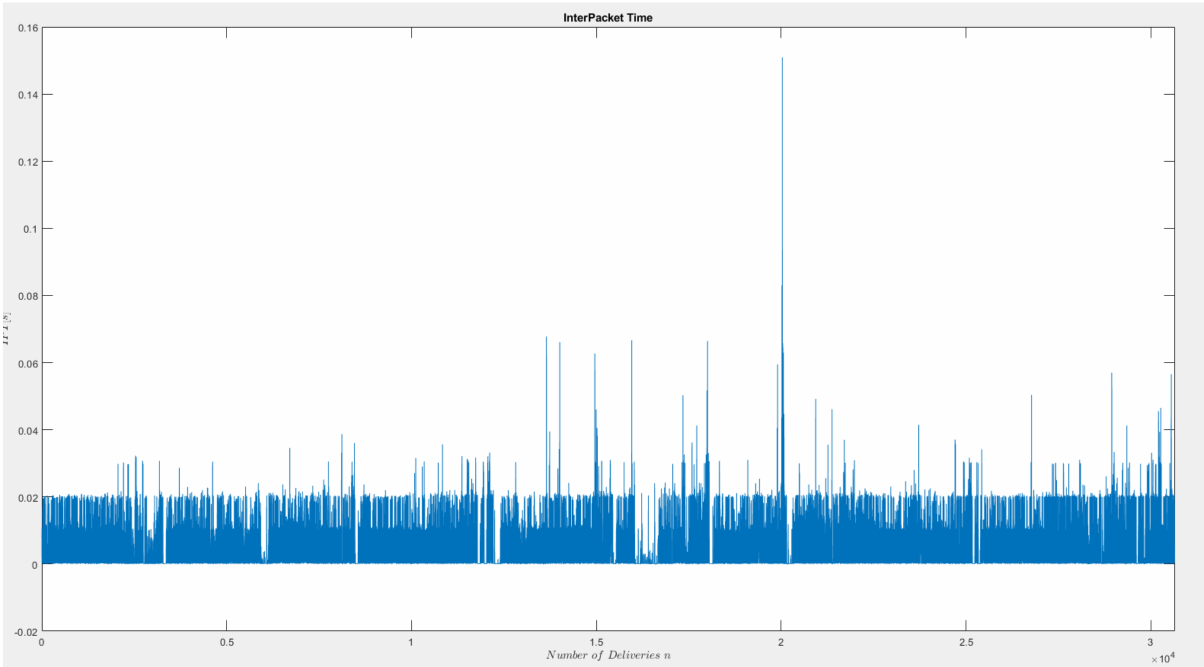


**Figure 8. Inter-packet Time**

Figure 8 shows the graph obtained from plotting the inter packet time or IPT, it is worth mentioning that wireshark does not give time periods with a value of 0. As it can be observed from the graph, each plotted value from the $x$ axis has a corresponding value higher than 0 in the $y$ axis. As shown in this figure, the IPT has a repetitive range between 0 and 0.2, meaning that the majority of deliveries took the same amount of time to be attended, showing a proper work of the internet network.

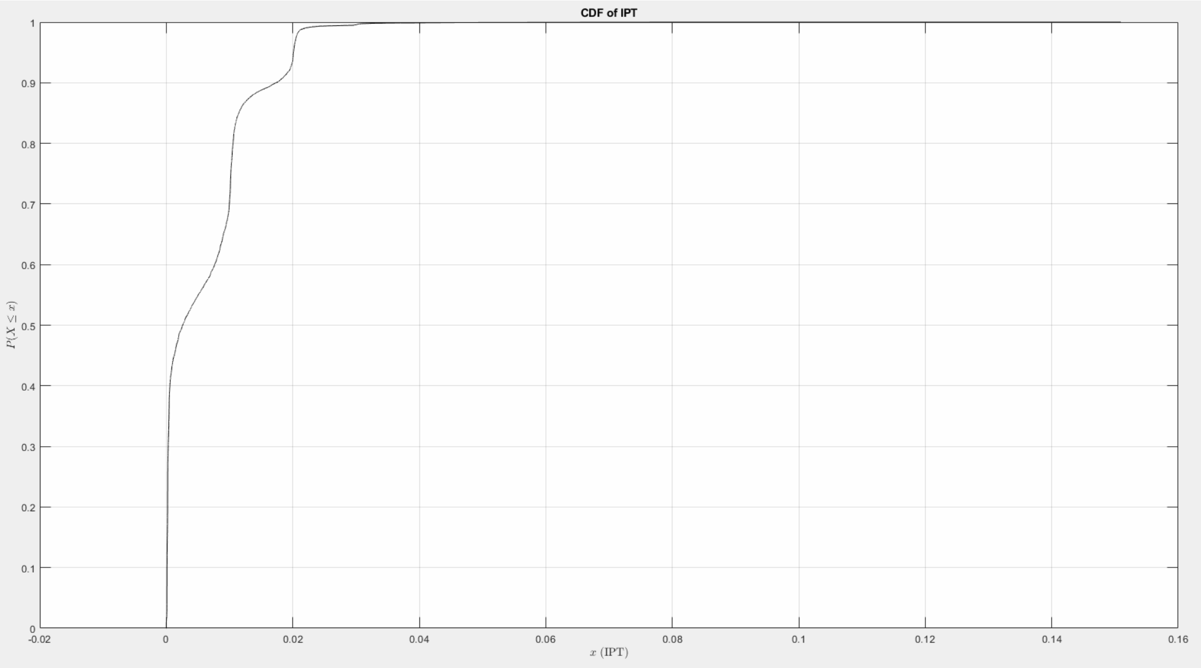


**Figure 9. CDF of IPT**

To formally characterize the cumulative distribution function of IPT a larger experiment is required but as shown in figure 9 the distribution that most approaches would be the Wiebull distribution. It is difficult because it does not follow a defined pattern.

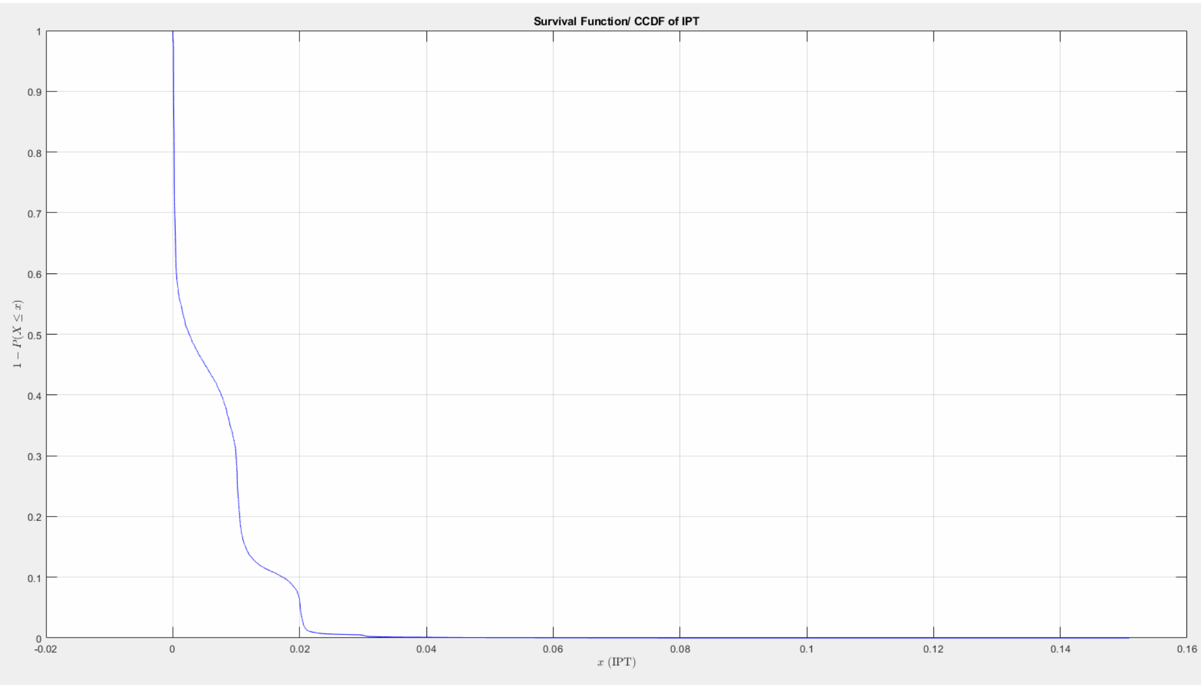


**Figure 10. Survival function of IPT**

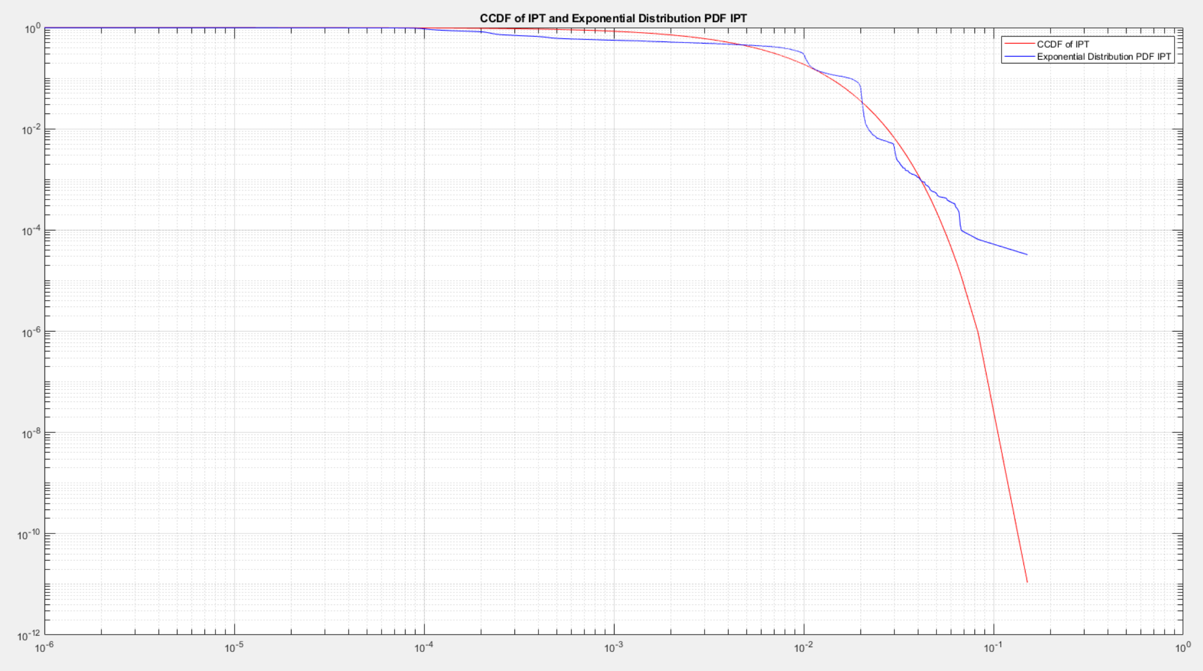


**Figure 11. Survival function comparison with Exponential PDF of IPT**

After graphically comparing the CCDF and the exponential distribution from the IPT through their plots, it is clear that an analysis for both heavy tail and light tail can be performed, as the exponential distribution plot is in some parts above and in others below the CCDF plot. It is important to mention that the relevant analysis and information comes directly from the heavy tail distribution, since it shows a higher probability of obtaining larger values or packets that might cause a network failure.

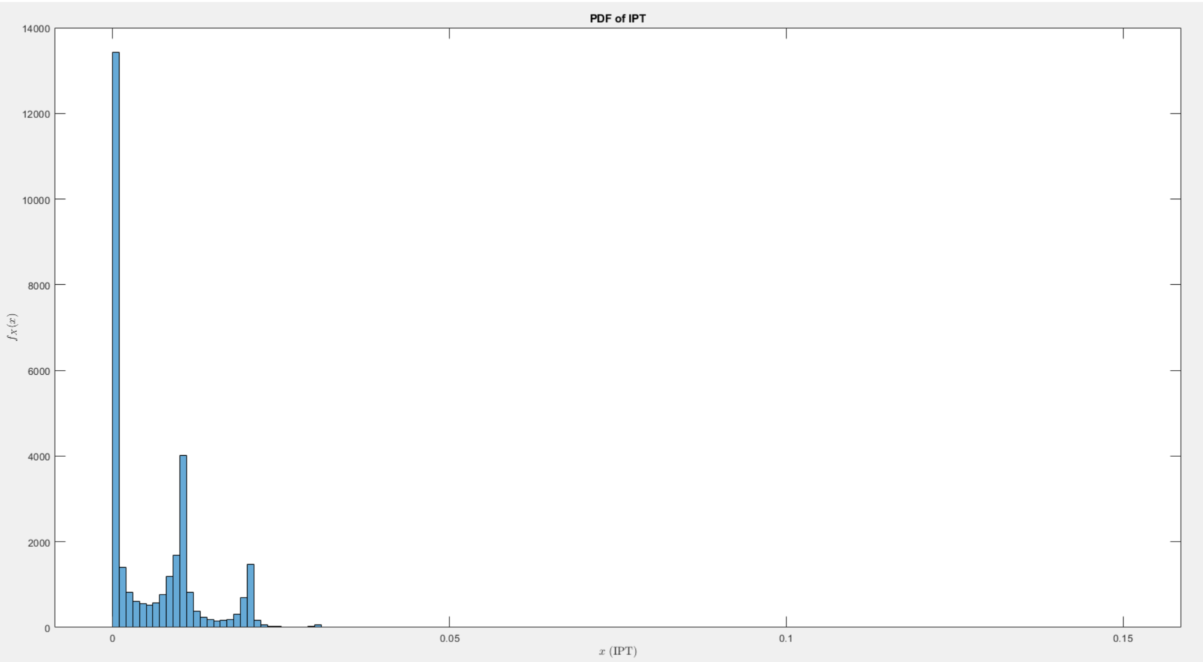


**Figure 12. PDF of IPT**

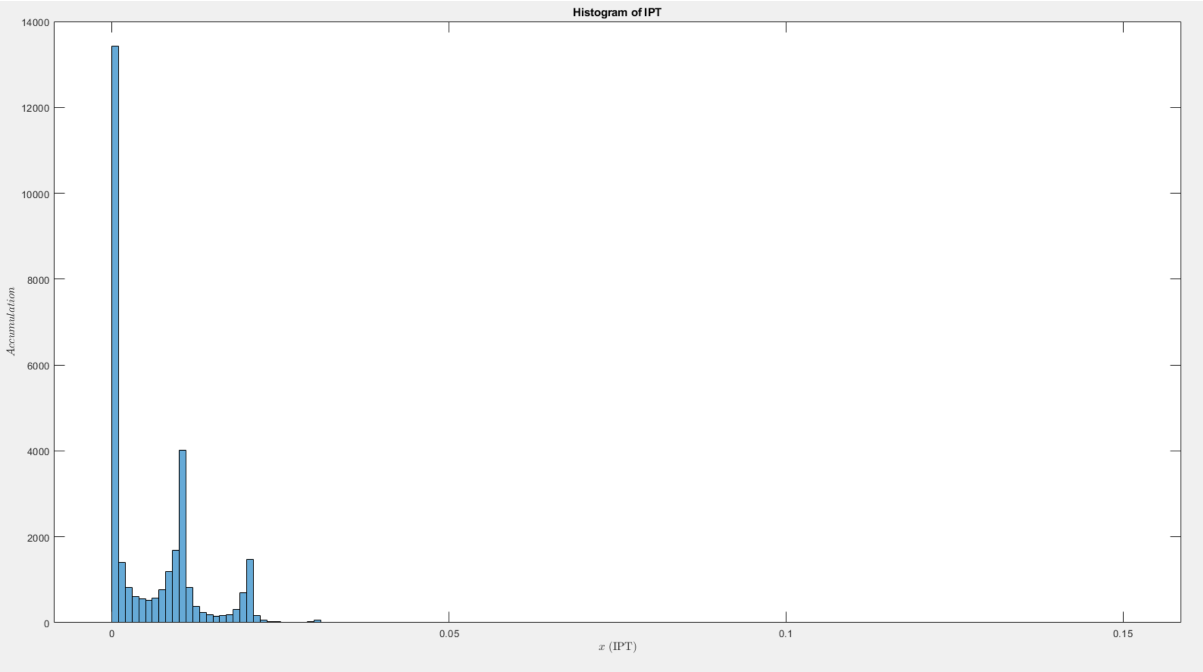


**Figure 13. Histogram of IPT**

As shown in Figure 8, the range most repeated IPT relies in the 0-0.8 interval so it is logical to find an irrefutable higher accumulation in that interval. A future work would be to list the parameters for a network to be considered as properly working and analyse these intervals of time to assure quality of service in highly demanding environments.

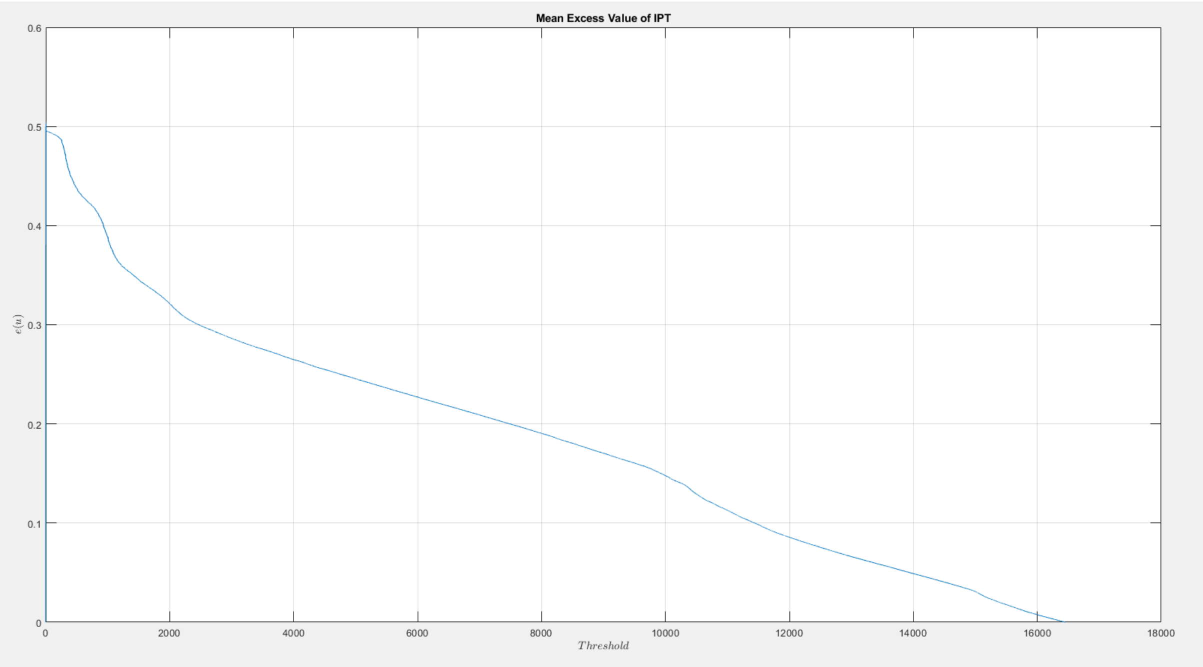


**Figure 14. ME of IPT**

As in the previous Mean excess value distributions, the initial values were particularly noisy at the beginning, with an abrupt jump from zero to a real initial value at the second sample; this as the matricial index forces the program to have an initial value of 2 instead of 1. For this case, the graph does have an evident Weibull behaviour from start to end, approaching softly but purposefully the zero.

## CONCLUSION

In conclusion, the analysis of internet traffic through different tools is a fundamental ability IT engineers should have, as well as having knowledge of the basic behavior that packets and networks have. Therefore it is really important to be able to use tools such as matlab and wireshark to perform different analyses of network traffic. Particularly for this case, the objectives were achieved, as the BPS were modelled in different ways. Even the coding of the models was done properly, the samples taken in 5 minutes are not enough to fully characterize a network. Further analysis must be performed in order to have a reliable enough characterization of a network. This time, a particular aspect to improve is the performance at the Mean Excess Function for the BPM, where unexpected values were presented at the tail. On the other hand, for these measurements the network had a phenomenal connectivity where all the bits arrived in almost the same time and with no errors, this type of behavior is not the same for everyday, giving range to further analysis on the conditions that allowed this performance for this time. Finally, it was learned that random processes are all around us and even if the modeling of all random phenomena is sometimes hard, it is also very important for its understanding and control, pushing the scientific community to a constant evolution in the sake of gaining more understanding of such important topics.

## EXTRA - ON/OFF Process

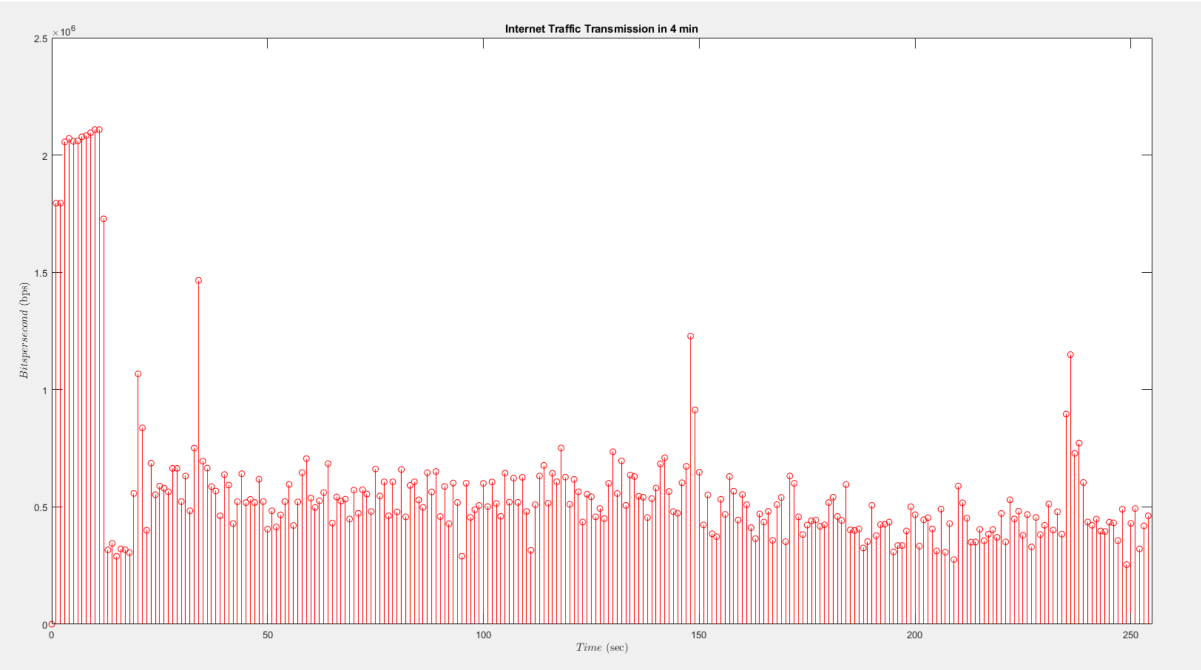


**Figure 15. Internet Traffic.**

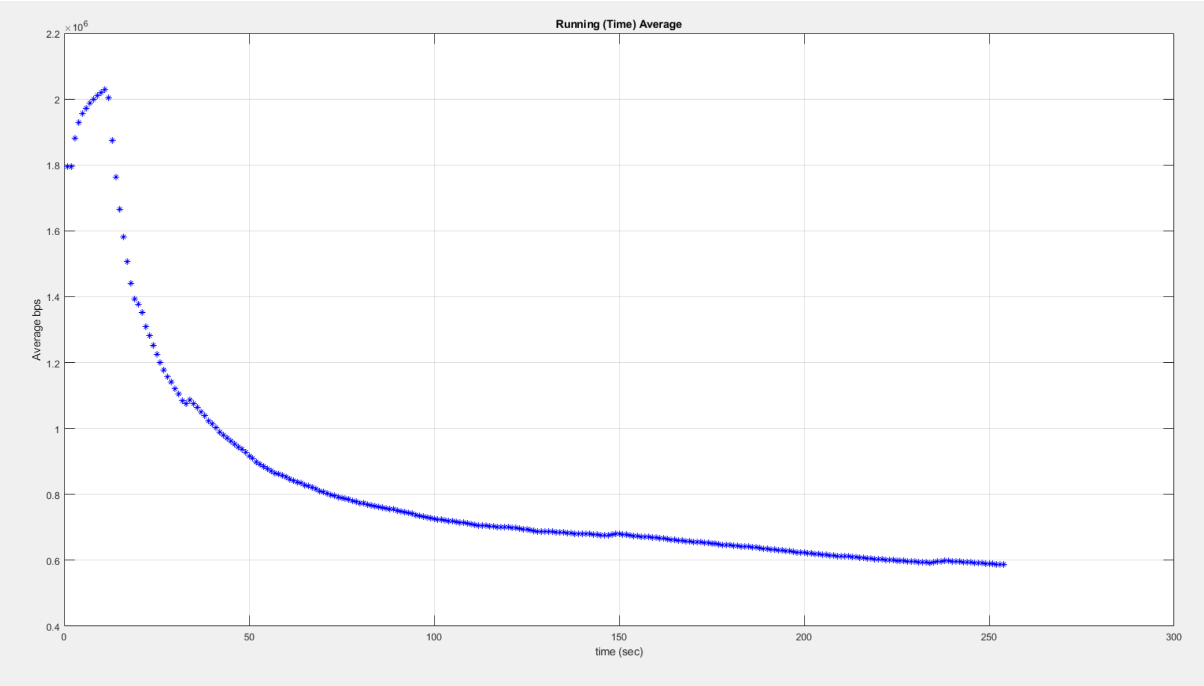


**Figure 16. Running average  of Internet Traffic**

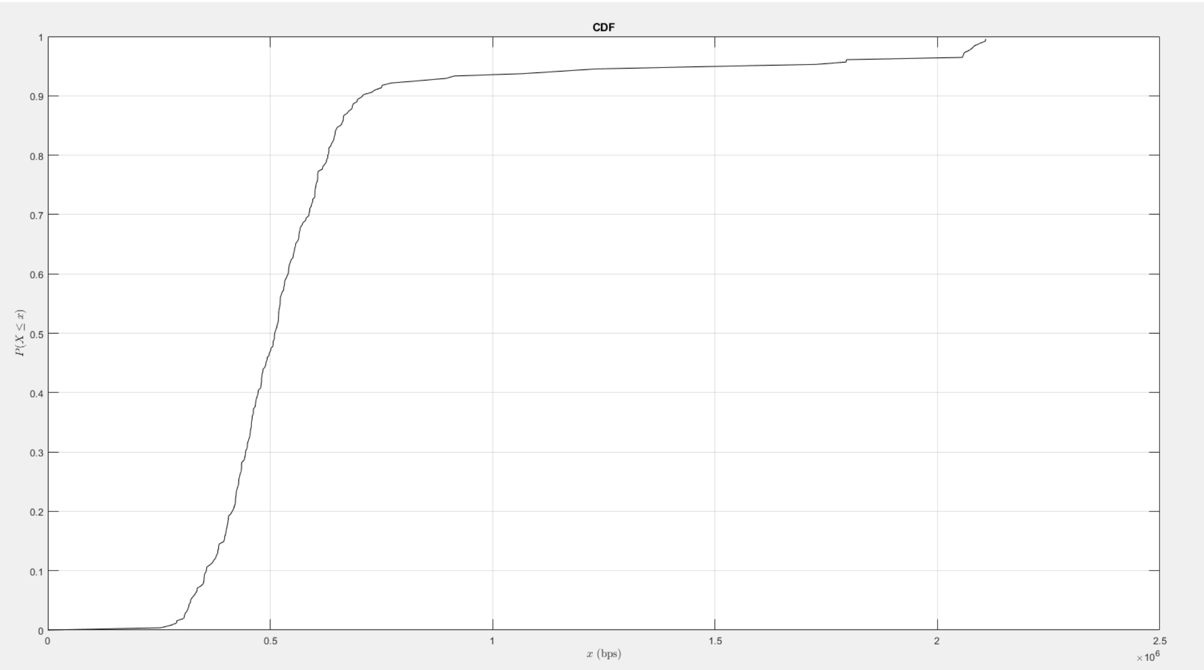


**Figure 17. CDF  of Internet Traffic**

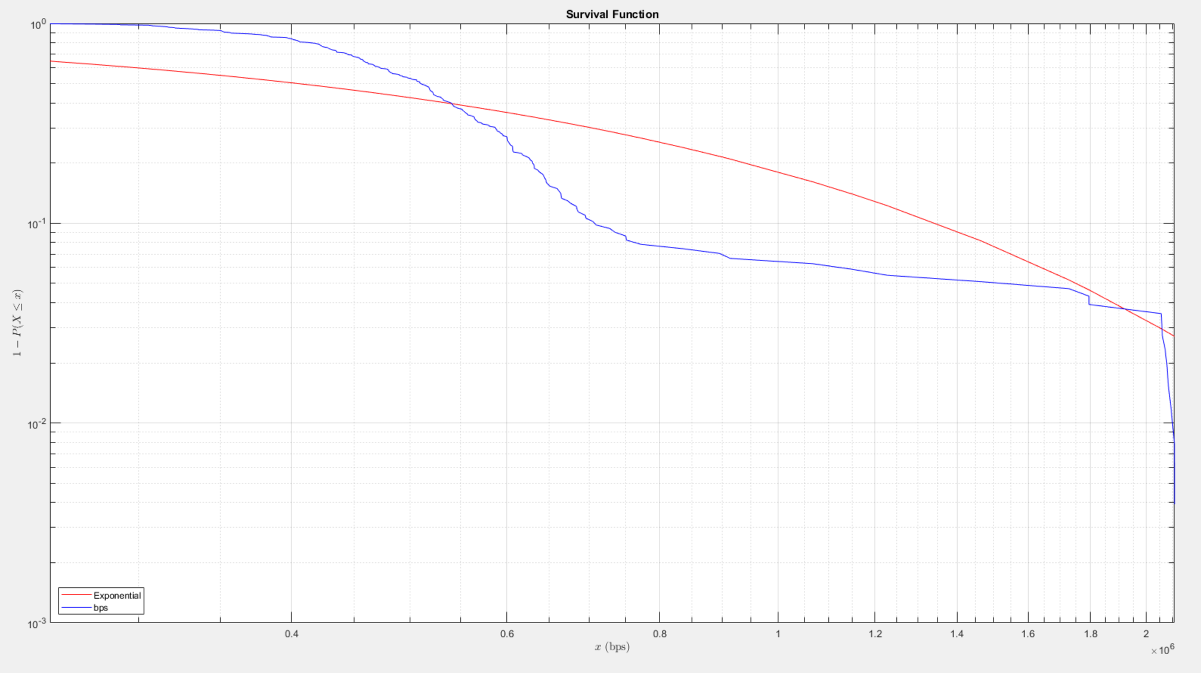


**Figure 18. Survival function of Internet Traffic**

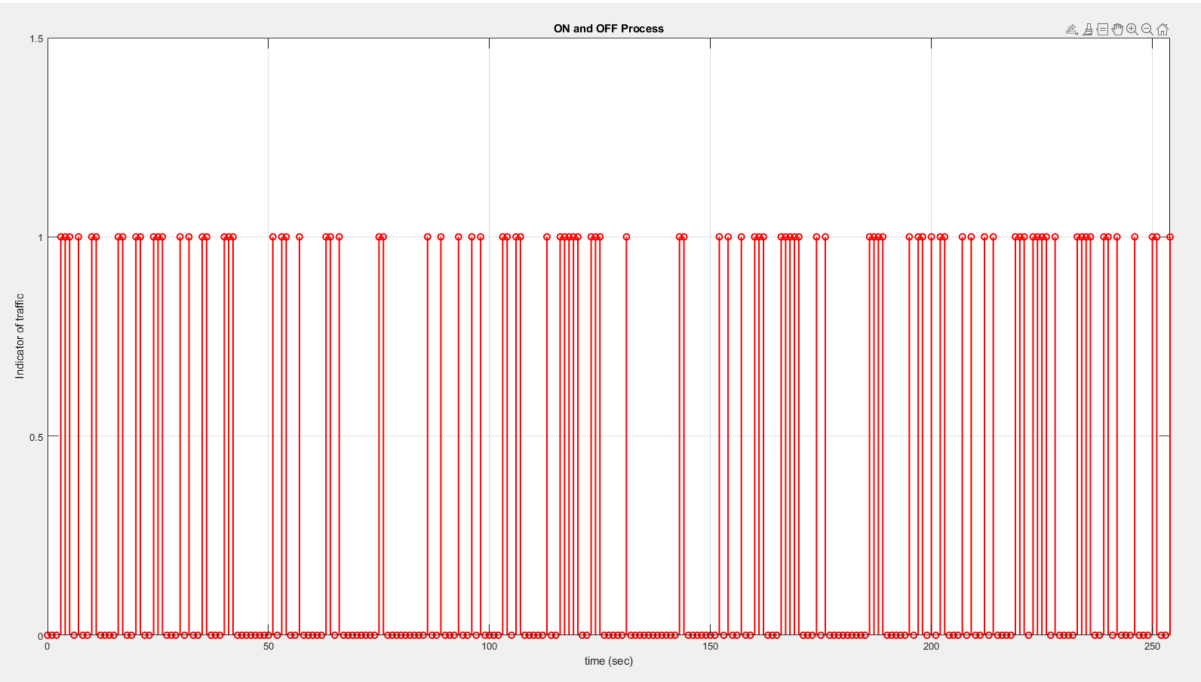


**Figure 19. ON/OFF process of Internet Traffic**

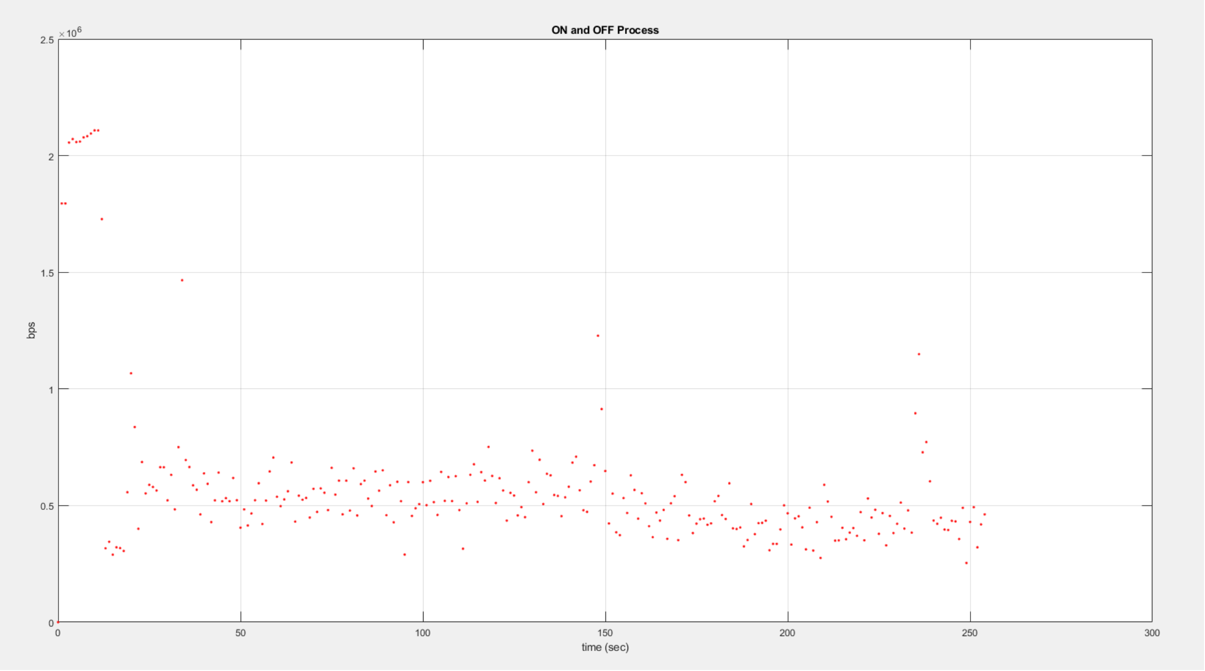


**Figure 20. On and OFF Process**

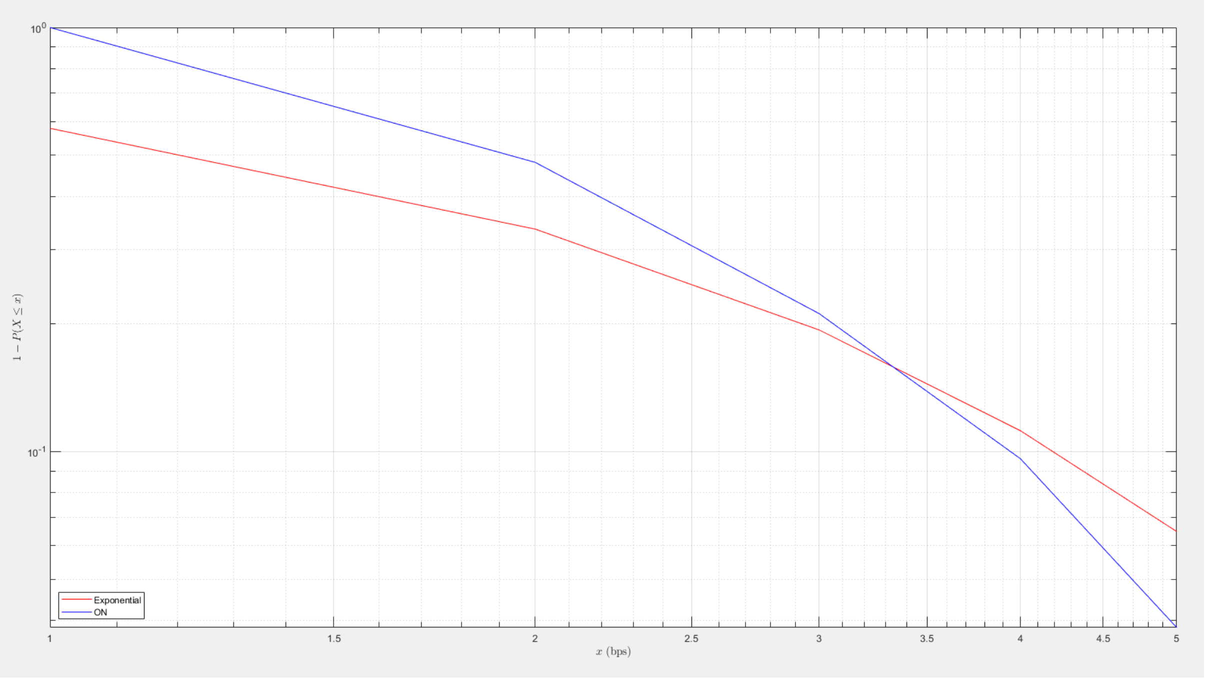


**Figure 21. Exponential and ON**

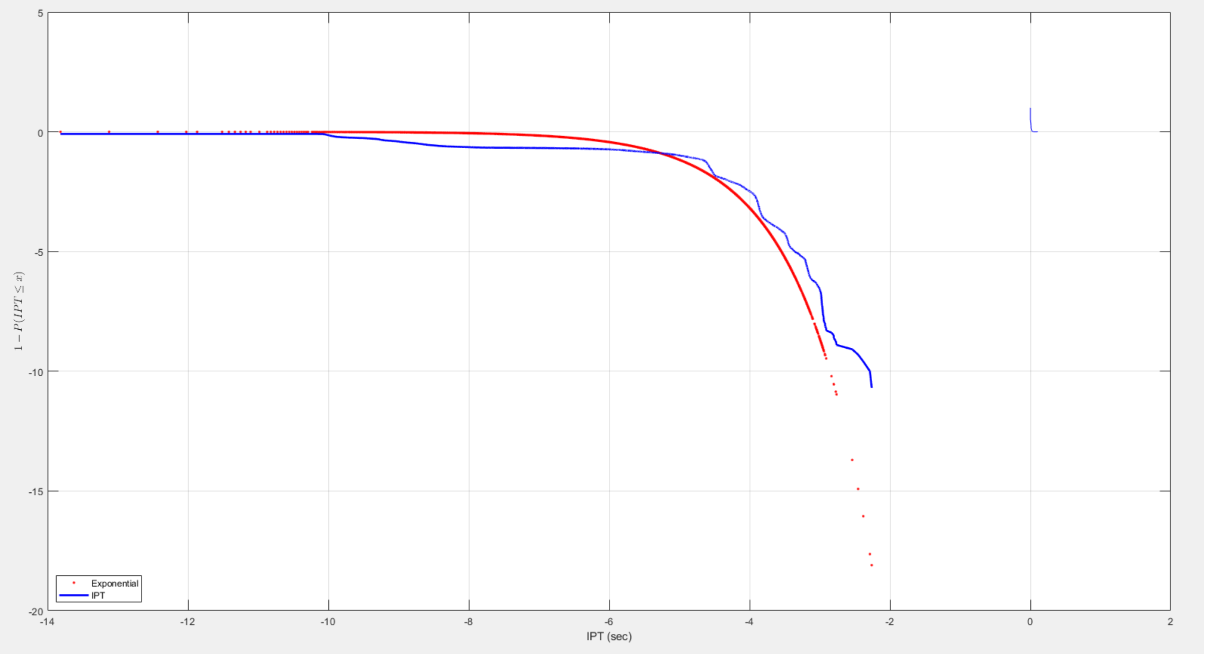


**Figure 22. Figure 21. Exponential and IPT**

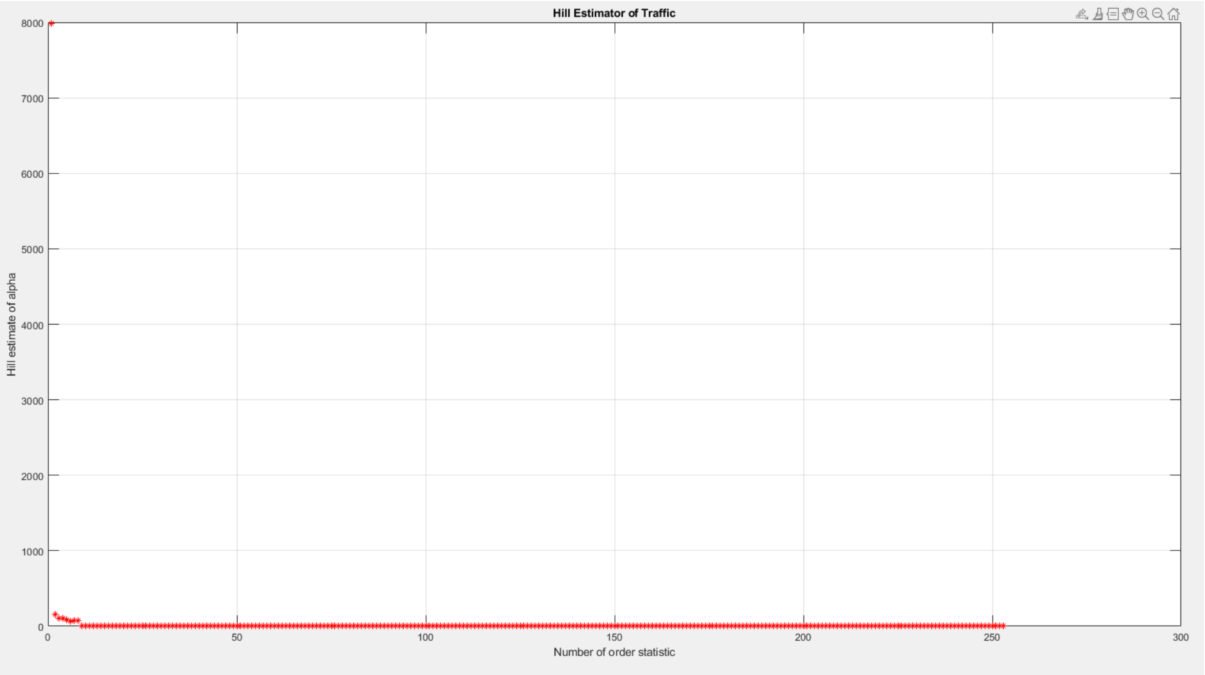


**Figure 23. Hill Estimator of Traffic.**

## REFERENCES:

[1]Ghost,S. & Resnick, S. (2010). A discussion on mean excess plots. ScienceDirect. Retrieved from: https://arxiv.org/pdf/0907.5236.pdf

[2]Ghost,S. & Resnick, S. (2010). When does the mean excess plot look linear? Retrieved from: https://arxiv.org/pdf/1011.6447.pdf

[3]Das, B. &Ghost, S. (2015). Detecting tail behavior: mean excess plots with confidence bounds. Retrieved from: https://arxiv.org/pdf/1501.00518.pdf

[4] (N.A). (2019). Retrieved 2 June 2021, from https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Poisson_Regression.pdf

[5] Roback, P., & Legler, J. (2021). *Beyond multiple linear regression: Applied generalized linear models and multilevel models in R*. CRC Press.

[6] Date, S. (2019). An Illustrated Guide to the Poisson Regression Model. Retrieved 2 June 2021, from https://towardsdatascience.com/an-illustrated-guide-to-the-poisson-regression-model-50cccba15958
[7] Páez, G. K. S., & Amaya, J. C. G. (2011). Algoritmos de gestión de tráfico: Leaky Bucket, Token Bucket y Vir tual. *Tecnura: Tecnología y Cultura Afirmando el Conocimiento, 15*(29), 76-89.