# Pareto Distribution Modeling for Network File Sizes: A Statistical Approach

Lakshmi Suresh Thandayaan

# Contents

# 1. Introduction

The purpose of this document is to analyse the distribution of file sizes that are send through an internet network. For analysing the distribution, we will initially take a random sample containing the size (in kB) of 1000 files and analyse the data from the sample. We will be using the programming language R to analyse the data from our distribution and make inferences.

# 2. Analysing the Random Sample

The figure below shows a histogram of the file sizes:

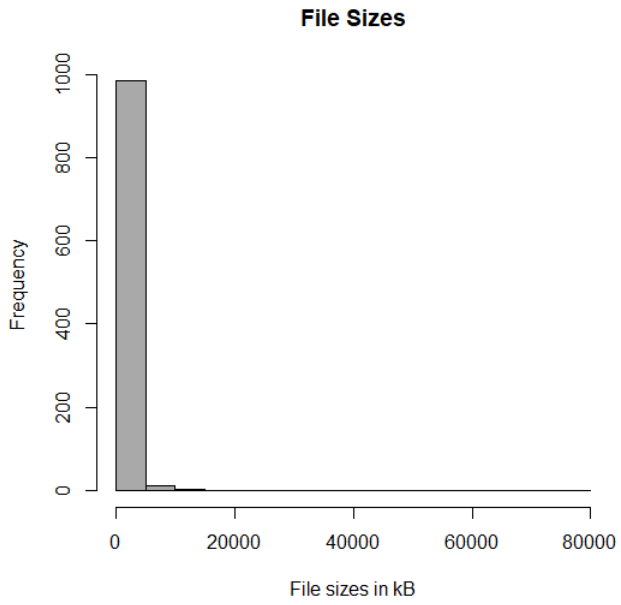


Figure 1: Histogram of a sample of the distribution

The histogram shows us that the distribution is positively skewed with a heavy tail. We obtain the mean, standard deviation, median and the quantiles of the data as follows:

mean = 1621.621, standard deviation = 2552.119, median = 1285.402
quantiles:

| 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|
| 1000.160 | 1098.374 | 1285.402 | 1637.462 | 77538.426 |

## 3. Pareto Model

From our analysis we understand that our distribution is positively skewed, with a heavy tail, which is characteristic of a Pareto distribution. Thus, we can conclude that a Pareto model would be suitable for our distribution of file sizes. So, the probability density function for our distribution will be:

$$f(x, \alpha, x_m) = \begin{cases} \dfrac{\alpha x_m^{\alpha+1}}{x^{\alpha+1}}, & x > x_m \\ \\ 0, & x < x_m \end{cases}$$

where $x_m$ is the minimum file size from the sample. So we assign $x_m$ to be 1000. $\alpha$ is currently an unknown parameter of the distribution.

## 4. Estimation of $\alpha$

For estimating the value of $\alpha$, we consider the likelihood function of $\alpha$ and our random sample $X = (x_1, x_2, ..., x_{1000})$. The likelihood function is given by:

$$L(\alpha, X) = \left(\frac{\alpha x_m^{\alpha}}{x_1^{\alpha+1}}\right) * \left(\frac{\alpha x_m^{\alpha}}{x_2^{\alpha+1}}\right) * \cdots * \left(\frac{\alpha x_m^{\alpha}}{x_{1000}^{\alpha+1}}\right)$$

$$= \frac{\alpha^{1000} x_m^{1000\alpha}}{(\prod_{i=1}^{1000} x_i)^{\alpha+1}}$$

Using the likelihood function ( here log is taken to the base $e$) we get:

$$l(\alpha) = \log L(\alpha, X)$$

$$= \log \frac{\alpha^{1000} x_m^{1000\alpha}}{(\prod_{i=1}^{1000} x_i)^{\alpha+1}}$$

$$= 1000 \log \alpha + 1000\alpha \log x_m - (\alpha + 1) \sum_{i=1}^{1000} \log x_i$$

And using the score function we get:

$$U(\alpha) = \frac{\partial l}{\partial \alpha} = \frac{1000}{\alpha} + 1000 \log x_m - \sum_{i=1}^{1000} \log x_i$$

Equating $U(\alpha)$ to 0, we get:

$$\alpha = \frac{1000}{\sum_{i=1}^{1000} \log x_i - 1000 \log x_m}$$

With $x_m = 1000$, we will get the value of $\alpha$ as 2.793.

## 5. Fischer Information of $\alpha$

The Fischer Information for $\alpha$, $I(\alpha)$ is given by:

$$I(\alpha) = -E\left[\frac{\partial^2 l}{\partial \alpha^2}\right]$$

We have:

$$\frac{\partial l}{\partial \alpha} = \frac{1000}{\alpha} + 1000 \log x_m - \sum_{i=1}^{1000} \log x_i$$

Differentiating the equation we get,

$$\frac{\partial^2 l}{\partial \alpha^2} = \frac{-1000}{\alpha^2}$$

So we get,

$$I(\alpha) = -E\left[\frac{\partial^2 l}{\partial \alpha^2}\right] = \frac{1000}{\alpha^2}$$

Thus we get the value of $I(\alpha)$ as 128.18. Since our sample chosen is very large, by maximum likelihood theorem we have:

$$\hat{\alpha} \sim N\left(\alpha, \frac{1}{I(\alpha)}\right)$$

$$\sim N(2.793, 0.01)$$

## 6. Confidence Interval of $\alpha$

The 95% equal-tailed confidence interval for $\hat{\alpha}$ can given by:

$$\hat{\alpha} \pm (1.96 \times se(\hat{\alpha}))$$

Since, $se(\hat{\alpha}) = \sqrt{1/I(\alpha)}$, we get that:

$$95\% \text{ CI of } \hat{\alpha} : \hat{\alpha} \pm 1.96 \times 0.088$$

Thus we can say that $\alpha$ is in the interval: [2.620,2.966] 95% of the time.

# 7. Estimating the Distribution of Sample Mean

In this section of the document, we will be discussing the results obtained by using the Pareto package in R to simulate 1000 random sample $X_i$ of a Pareto distribution with $x_m = 1000$ and $\alpha = 2.793$ and using that to estimate and analyze the distribution of the sample mean Y. The figure given below gives the histogram for the distribution we have obtained for Y:

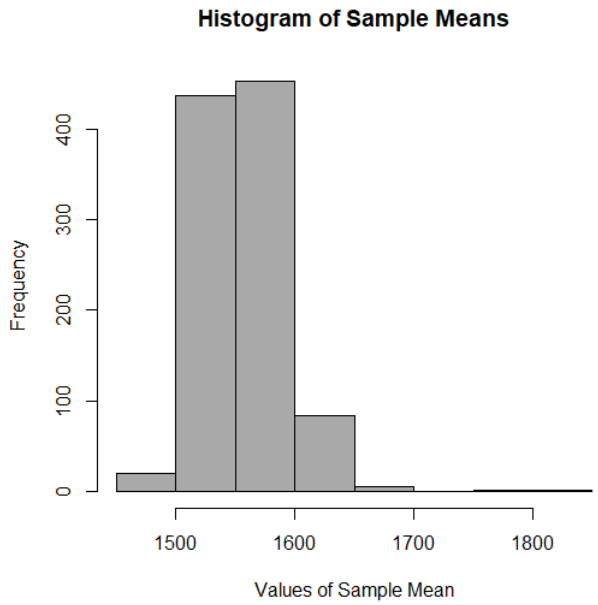


Figure 2: Histogram of the sample mean

We obtain mean, standard deviation, median and quantiles of the $Y$ as follows:

mean: 1557.396, standard deviation: 33.539, median: 1554.31
quantiles:

| 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|
| 1476.129 | 1533.956 | 1554.310 | 1577.386 | 1844.284 |

For large random samples, the sample mean would converge to normal distribution. This can be confirmed by the histogram of $Y$.

## 8. Upper Limit for the File Sizes

By analysing Pareto distributions in R we can estimate that 99% of the files have size less that **5201 kB**. This can be used by a network administrator to set an upper limit for that files sizes, so that 99% of the files will be accepted by the network.

## 9. Conclusion

Based on the analysis conducted, we can conclude that the file sizes transmitted through the network follow a Pareto distribution. For our data, we identified the minimum possible file size $x_m$ as 1000 kB, given that all sampled files are at least this size. We estimated the shape parameter $\alpha$ of the Pareto distribution to be 2.793.

Our analysis also involved calculating the Fisher information for $\alpha$, which was found to be 128.18. Consequently, the estimator $\hat{\alpha}$ is approximated to follow a normal distribution $N(2.793, 0.01)$. We determined that with 95% confidence, the true value of $\alpha$ lies within the interval [2.620, 2.966].

Additionally, we established that to ensure 99% of file sizes are accepted by the network, an upper limit of 5201 kB should be set. This limit will effectively accommodate the majority of the files based on the Pareto distribution parameters estimated from the sample.