F21SA Assessed Project

1 Introduction

The file sizes $\underline{x}$, is used as a frequency to analyze the internet network data with the help of the Pareto statistical model. It's used for weighing the pros and cons of solving situations in conflict which enables you to concentrate on the solutions that will bring you the greatest benefits.

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

In order to estimate future file sizes, the model is fitted to the data using maximum likelihood estimation of the distribution's scale parameter σ. The likelihood that file size variance will increase in the upcoming months is calculated, and the robustness of the finding is examined.

2 Data Summary

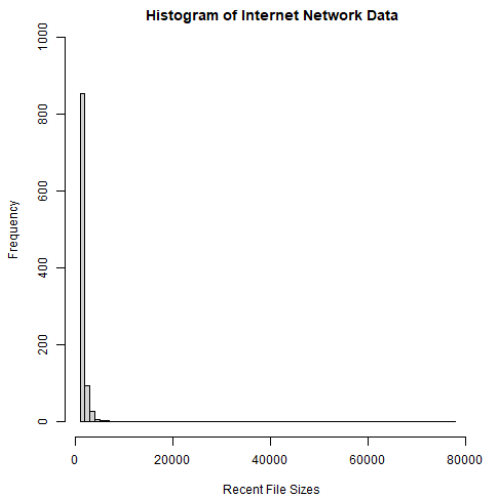

Figure 1: Histogram of File sizes

```
> cat("Network Data\n")
Network Data

> cat("Mean:", mean(myData), "kB\n")
Mean: 1621.621 kB

> cat("Standard deviation:", sd(myData), "kB\n")
Standard deviation: 2552.119 kB

> cat("Median:", median(myData), "kB\n")
Median: 1285.402 kB

> cat("Quantiles:", quantile(myData), "(all kB)\n\n")
Quantiles: 1000.16 1098.374 1285.402
           1637.462 77538.43 (all kB)

> cat("Skewness:", skewness(myData))
Skewness: 26.53065
```

Listing 1: Summary & Standard Deviation

The histogram and numerical summary, including the sample mean, standard deviation, median, and other quartiles, have been obtained from Figure 1 and Listing 1. Quartiles such as Q1 (1098.374) and Q3 (1637.462) have been found as well.

From the above skewness, we can conclude that the above data is positivly skewed which shows that the mean is greater than the median. This indicates that majority of the file sizes sent over were on the lower size(kb) limit while fewer outliers with higher file sizes(kb) are lower in frequency.

3 Maximum Likelihood Estimation (MLE) for $\hat{\alpha}$

A method for estimating the parameters of a particular distribution using some observed data is the likelihood function. By determining specific values for the mean and variance that make the observation the most likely outcome, MLE can be used to estimate the sample variance using a small sample of the distribution. The

likelihood function is,

$$\begin{aligned}\mathcal{L}(x,\alpha,x_m) &= \prod_{i=1}^{n}\frac{\alpha x_i^{\alpha}}{x_i^{\alpha+1}} = \alpha^n x_i^{n\alpha}\prod_{i=1}^{n}\frac{1}{x_i^{\alpha+1}} \\ &= \ln\left((\alpha x_m^{\alpha})^m\right) - \left(\sum_{i=1}^{m}\ln x_i^{\alpha+1}\right) = m\ln\left(\alpha x_m^{\alpha}\right) - \left(\sum_{i=1}^{m}\ln x_i^{\alpha+1}\right) \\ &= m\ln\alpha + m\alpha\ln\left(x_m\right) - \ln\left(x_m\right) - \sum_{i=1}^{m}\ln x_i^{\alpha+1}\end{aligned}$$

For α, we set the partial derivative of ℓ with respect to α equal to 0:

$$\begin{aligned}&\frac{\partial}{\partial\alpha}(m\ln\alpha) + \frac{\partial}{\partial\alpha}\left(m\alpha\ln x_m\right) - \frac{\partial}{\partial\alpha}\left((\alpha+1)\sum_{i=1}^{m}\ln x_i\right) \\ &= \frac{m}{\hat{a}} + m\ln\left(x_m\right) - \sum_{i=1}^{m}\ln x_i = 0;\ = \frac{m}{\hat{a}} = \sum_{i=1}^{m}\ln x_i - m\ln x_m \\ Therefore, \rightarrow \hat{\alpha} &= \frac{m}{\sum_{i=1}^{m}\ln x_i - m\ln x_m}\end{aligned}$$

4 Fisher's information for I(α) with the distribution of $\hat{\alpha}$

$$\begin{aligned}I(\alpha) &= -\left[E\left[\frac{\partial^2\ell}{\partial\alpha^2}\right]\right] \\ I(\alpha) &= -\frac{\partial^2}{\partial\alpha^2}\left[\frac{m}{\alpha} + m\ln x_m - \sum_{i=1}^{m}\ln x_i\right] \\ I(\alpha) &= -\left[\frac{-m}{\alpha^2} + 0\right] \\ I(\alpha) &= \frac{m}{\alpha^2}\end{aligned}$$

For large m, $\hat{\alpha}$ is approximately distributed as $N(\alpha, \frac{1}{I(\alpha)})$. In this case that is $N(\alpha, \frac{\alpha^2}{m})$

5 Using the results of 3 and 4 , finding $\hat{\alpha} = \hat{\alpha}(\underline{x})$ to approximate 95% equal-tailed confidence interval $I = [\alpha_L(\underline{x}), \alpha_U(\underline{x})]$ for α

$$[\alpha_L(\underline{x}), \alpha_U(\underline{x})] = \hat{\alpha} \pm (Z_{\frac{\alpha}{2}} * ese(\hat{\alpha}))$$

Fromg the above MLE equation, we can derive the following;

$$\hat{\alpha} = \frac{m}{\sum_{i=1}^{m}\ln x_i - m\ln x_m}$$

The $\hat{\alpha}$ value derived from the equation is 2.793079. [Refer Appendix:Code]

$$ese(\hat{\alpha}) = \frac{1}{\sqrt{I(\hat{\alpha})}}$$

The $ese(\hat{\alpha})$ value derived from the equation is 0.08832491. [Refer Appendix:Code]
Substituting the values of $\hat{\alpha}$ and $ese(\hat{\alpha})$ in and with $Z_{\frac{\alpha}{2}} = 1.96$ taken from NCST Table 5, the confidence interval is

$$I_{0.95} = [\alpha_L(\underline{x}), \alpha_U(\underline{x})] \approx [2.619965, 2.966192]$$

6 Estimation of the distrubtion Y'

This section aims to Calculate, Predict and Estimate Y', the mean file sizes for the next 1000 files.

By letting $X_i' \sim \text{Pareto}(\hat{\alpha}, x_m)$, and $Y' = \frac{1}{1000}\sum_{i=1}^{1000} X_i'$ be the predicted mean file sizesfor the next 1000 files, R is used to simulate the predicted file sizes and estimate the distribution of Y'.

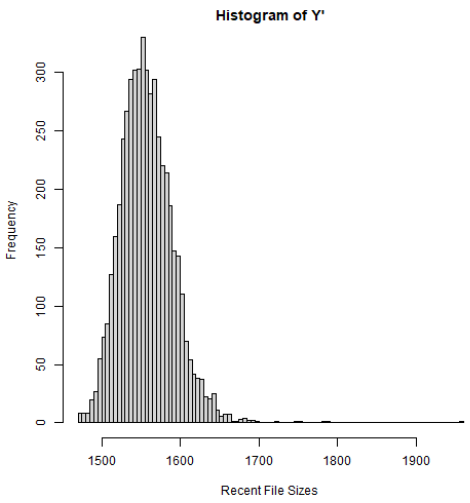

Figure 2: Histogram of File sizes

```
> cat("Network Data\n")
Network Data

> cat("Mean:", mean(y_dash), "kB\n")
Mean: 1557.879 kB

> cat("Standard deviation:", sd(y_dash), "kB\n")
Standard deviation: 33.22118 kB

> cat("Median:", median(y_dash), "kB\n")
Median: 1555.585 kB

> cat("Quantiles:", quantile(y_dash), "(all kB)\n")
Quantiles: 1462.642 1535.643 1555.585
           1576.872 1944.324 (all kB)

> cat("Skewness:", skewness(y_dash) )
Skewness: 1.179502
```

Listing 2: Summary & Standard Deviation

The histogram and numerical summaries, including the sample mean, standard deviation, median, and other quartiles, have been calculated from Figure 2 and Listing 2. The histogram and numerical summaries, including the sample mean, standard deviation, median, and other quartiles, have been calculated from Figure 2 and Listing 2. Quartiles such as Q1 (1535.643) and Q3 (1576.872) have been found as well.

From the above skewness, we can conclude that the above data is normally skewed distribution which shows that the mean in this case is almost equal to the median with minute differences. This tells us that the file sizes are roughly around [1550kB].

7 Maximum Possible Limit

The maximum possible limit is set to the 99% so that all the outliers are removed,in such a way that 99% of files would be accepted by the network. The qPareto function is utilized [qPareto(p(0.99),t(1000),α)(2.793079)] from Maximum Likelihood Estimation(3).

From this observation we can undertand that the MPL(Maximum Possible Limit) with the above calculation is 5200.627, which states that all the values under the value of the result(5200.627kB) will be accepted by the network removing all the outliers that are above 5200.627. [Refer Appendix:Code]

8 Conclusion

With the given data set from above we were able to understand the statistical summary analysis of the file sizes using the Pareto function. In this dataset the mean is more than the median which proves the distribution is positively skewed. When attempted to calculate the mean with more number of simulation

with the appropriate function, it is seen that the distribution becomes normally skewed. From the histogram, it is evident that the mean and median are roughly the same.

By calculating the 99% limit of the total data, we are able to remove any of the outliers above the limit, allowing the network to accept the the files under the limit only.

References

[1] https://cran.r-project.org/web/packages/Pareto/Pareto.pdf

[2] https://online.stat.psu.edu/stat415/lesson/1/1.2

[3] Lindley, D. and Scott, W., 1995. New Cambridge Statistical Tables. 2nd ed. New York: Cambridge University Press.

[4] https://www.scribbr.com/statistics/skewness/

9 Appendix

9.1 Appendices: Code

```
# F21SA Statistical Modelling and Analysis
# H00410394 Sasha Fathima Suhel

# R language Version
R.version.string

# Reading and storing csv
# Internet network data is stored in an csv file, x column contains file sizes
myData = read.csv("filesize.csv")[["x"]]

library(moments)
library(Pareto)

# Histogram to summarize data
png(filename = "Histogram.png",
    width = 480, height = 480, units = "px", pointsize = 12,
    bg = "white")
par(mar=c(4,4,1,1)+0.1)
hist(myData, main = "Histogram of Internet Network Data",
     xlab="Recent File Sizes",ylim = c(0,1000),xlim = c(1000,80000) ,breaks = 100)
suppress <- dev.off()
# Summary statistics
cat("Network Data\n")
cat("Mean:", mean(myData), "kB\n")
cat("Standard deviation:", sd(myData), "kB\n")
cat("Median:", median(myData), "kB\n")
cat("Quantiles:", quantile(myData), "(all kB)\n")
cat("Skewness:", skewness(myData) )

# MLE
alphahat = length(myData)/(sum(log(myData)) - length(myData)*(log(1000))) ; alphahat

# Estimated Standard Error
ese = alphahat / sqrt(length(myData)) ; ese
# Confidence Interval
z_025 <- qnorm(p = 0.025, lower.tail = FALSE) # From NCST Table 5
CI_L = alphahat - (z_025*ese)
CI_U = alphahat + (z_025*ese)
cat("Confidence Interval: [",CI_L,",",CI_U,"] ")

# Estimating the derivation for Y'
y_dash=numeric(0)
for(i in 1:5000){
  y_dash[i]=mean(Pareto::rPareto(length(myData),1000,alphahat))
}
# Plotting histogram for the sample means from Y'
png(filename = "Y_dash.png",
```

```
    width = 480, height = 480, units = "px", pointsize = 12,
    bg = "white")
par(mar=c(4,4,1,1)+0.1)
hist(y_dash, main = "Histogram of Y'",
     xlab="Recent File Sizes", breaks = 100)
suppress <- dev.off()

# Summary of the new data
cat("Network Data\n")
cat("Mean:", mean(y_dash), "kB\n")
cat("Standard deviation:", sd(y_dash), "kB\n")
cat("Median:", median(y_dash), "kB\n")
cat("Quantiles:", quantile(y_dash), "(all kB)\n")
cat("Skewness:", skewness(y_dash) )

# To obtain Maximum Possible Limit
m_p_l = qPareto(0.99,length(myData),alphahat) ;m_p_l
```