F21SA Coursework 1

1. Introduction

In this coursework, we are provided with a csv file which contains the distribution of file sizes sent through an internet network. We have to use the Pareto statistical model and model those sizes by the Maximum Likelihood estimation method.

We have access to $\underline{x}$, which contains sizes in kB of 1000 randomly selected files that were recently sent through the network. We are also provided with the Probability Density Function (PDF) as

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

2. Data Summarization

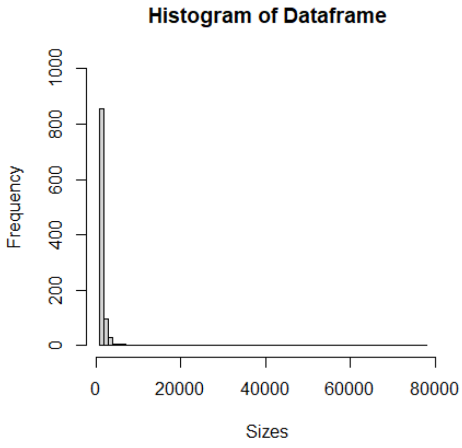

Figure 1: Histogram of File size

```
summary(file_size)
         x

Min.      :  1000
1st Qu.:  1098
Median  :  1285
Mean      :  1622
3rd Qu.:  1637
Max.      :  77538
S.D.      :  2552.119
Var.      :  6513309
IQR       :  539
```

Listing 1: Numerical Summary

Figure 1 and Listing 1 displays the Histogram and the Numerical Summary of the file size data respectively. The file size data has a mean of 1622kB, median of 1285kB, standard deviation of 2552.119kB, Q1 of 1098kB, Q3 of 1637kB, and an IQR of 539kB. This explains that the average file size is about 1622, and usually not over 1637kB or under 1098kB.

We can observe from the histogram that the distribution is positively skewed. This is also correct because the file size of any transferred file cannot be negative.

3. Maximum Likelihood Estimation (MLE)

The likelihood function is,

$$\mathcal{L}(x,\alpha,x_m) = \prod_{i=1}^{n} \frac{\alpha x_i^\alpha}{x_i^{\alpha+1}} = \alpha^n x_i^{n\alpha} \prod_{i=1}^{n} \frac{1}{x_i^{\alpha+1}} \tag{2}$$

The log-likelihood function is,

$$= ((\alpha x_m^\alpha)^m) - \left(\sum_{i=1}^{m} \ln x_i^{\alpha+1}\right) = m\ln(\alpha x_m^\alpha) - \left(\sum_{i=1}^{m} \ln x_i^{\alpha+1}\right) \tag{3}$$

$$= m \ln \alpha + m\alpha \ln (x_m) - \ln (x_m) - \sum_{i=1}^{m} \ln x_i^{\alpha+1} \tag{4}$$

To obtain the MLE for α, $\hat{\alpha}$, we solve for 0 after calculating the first derivative.

$$= \frac{\partial}{\partial \alpha}(m \ln \alpha) + \frac{\partial}{\partial \alpha}(m\alpha \ln x_m) - \frac{\partial}{\partial \alpha}\left((\alpha+1)\sum_{i=1}^{m} \ln x_i\right) \tag{5}$$

$$\Rightarrow \frac{m}{\hat{a}} + m \ln (x_m) - \sum_{i=1}^{m} \ln x_i = 0 \tag{6}$$

$$\Rightarrow \frac{m}{\hat{a}} = \sum_{i=1}^{m} \ln x_i - m \ln x_m \tag{7}$$

$$\Rightarrow \boxed{\hat{\alpha} = \frac{m}{\sum_{i=1}^{m} \ln x_i - m \ln x_m}} \tag{8}$$

4. Fisher information for α to approximate the distribution of $\hat{\alpha}$

The Fisher information is found as follows:-

$$I(\alpha) = -\left[E\left[\frac{\partial^2 \ell}{\partial \alpha^2}\right]\right] \tag{9}$$

$$I(\alpha) = -\frac{\partial^2}{\partial \alpha^2}\left[\frac{m}{\alpha} + m \ln x_m - \sum_{i=1}^{m} \ln x_i\right] \tag{10}$$

$$\begin{aligned} &= -\left[\frac{-m}{\alpha^2} + 0\right] \\ &\Rightarrow \boxed{I(\alpha) = \frac{m}{\alpha^2}} \end{aligned} \tag{11}$$

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

5. 95% Confidence Interval (CI) for $\hat{\alpha}$

$$[\alpha_L(\underline{x}), \alpha_U(\underline{x})] = \hat{\alpha} \pm (Z_{\frac{\alpha}{2}} * ese(\hat{\alpha})) \tag{12}$$

From eq. (8), we have,

$$\hat{\alpha} = \frac{m}{\sum_{i=1}^{m} \ln x_i - m \ln x_m} \tag{13}$$

The value of $\hat{\alpha}$ is 2.793079. [Refer Appendix for R code]
From eq. (11), we have,

$$ese(\hat{\alpha}) = \sqrt{\frac{1}{I(\hat{\alpha})}} = \sqrt{\frac{\hat{\alpha}^2}{m}} \tag{14}$$

The value of $ese(\hat{\alpha})$ is 0.08832491. [Refer Appendix for R code]

Substituting the values of $\hat{\alpha}$ and $ese(\hat{\alpha})$ in eq.(12) and with $Z_{\frac{\alpha}{2}} = 1.96$ taken from NCST Table 5 [1], the confidence interval is

$$I_{0.95} = [\alpha_L(\underline{x}), \alpha_U(\underline{x})] \approx [2.619965, 2.966192] \tag{15}$$

6. Estimation of Y'

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 size, R is used to simulate the predicted file sizes and find the distribution of Y'.

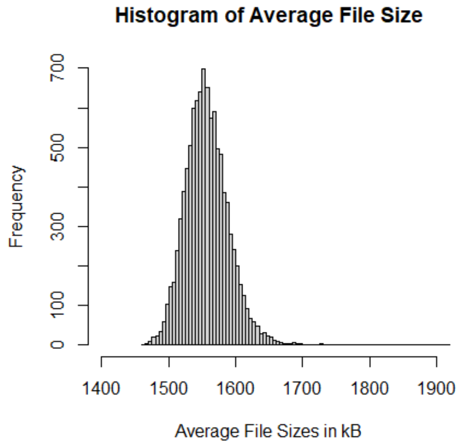

Figure 2: Histogram of Average File size

```
summary(average_file_size)
          x

Min.    :  1463
1st Qu.:  1535
Median  :  1555
Mean    :  1557
3rd Qu.:  1577
Max.    :  1972
S.D.    :  32.5771
Var.    :  1061.267
IQR     :  42
```

Listing 2: Numerical Summary

It seems from simulation in R that the predicted mean file sizes has an approximate Normal distribution. The simulation is set at 10,000 times and by using the rPareto(n,t,α) function, where n is length(file size), t is x_m and α is $\hat{\alpha}$.

From the results, we can see that the average file size is in the range of 1463 kB and 1972 kB with interquartile range of 42 kB.

7. Maximum Possible Limit

We are supposed to set an upper limit on the file size to discard the files that are above the limit. We need to find the max possible limit to accept atleast 99% of the incoming files.

To implement this, I've made use of the qPareto(p,t,α) function where p,t and α are 0.99, 1000 (file size data length) and 2.793079 (from eq. (13)) respectively.

After calculation, the function returned the value as 5200.627 kB. [Refer Appendix for R code]

From this calculation, we understand that 99% of the values in the dataset are below 5200 kB. So the maximum possible limit such that 99% of the files will be accepted would be 5200 kB.

8. Conclusion

We have successfully analyzed the dataset which contains the distribution of file sizes that are being sent over the network using Pareto Statistical Model. Also, we found that the mean file sizes has a normal distribution after running multiple simulations. We also found the limit on the file size, so that the files of size that are higher than the limit will be rejected but also 99% of the files would still be accepted by the network.

References

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

[2] https://www.statisticshowto.com/fisher-information/

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

Appendix

CW1.r

```
# Importing the dataframe file as csv

df = read.csv(”filesize.csv”, header = TRUE)

# Installing Pareto library

library(Pareto)

# Fetching the column from the dataframe

file_size = df$x

# Calculating the numerical summaries

mean = mean(file_size)
summary = summary(file_size)
standard_deviation = sd(file_size)
variance = var(file_size)

# Printing the summaries

print(summary)
cat(”Standard␣Deviation:␣”, standard_deviation)
cat(”Variance:␣”, variance)

# Creating histograms

hist(file_size, main = ”Histogram␣of␣Dataframe”,
     xlab=’Sizes’, xlim=c(1000,80000), ylim=c(0,1000),breaks = 100)

# MLE

x_m=1000 # Given
alpha_hat = length (file_size) / (sum(log(file_size))
                                  - length(file_size)*(log(x_m)))
alpha_hat

# Estimated standard error

ese_alpha = alpha_hat / sqrt(length(file_size))
ese_alpha

# Confindence Interval

z_025 <- qnorm(p = 0.025, lower.tail = FALSE)
alpha_CI_L = alpha_hat - (z_025*ese_alpha)
```

```
alpha_CI_U = alpha_hat + (z_025*ese_alpha)

cat("Confidence Interval: [", alpha_CI_L, ",", alpha_CI_U, "] ")

# Simulations

y_prime=0
for(i in 1:10000)
{ y_prime[i]=mean(Pareto::rPareto(length(file_size),x_m,alpha_hat))}
hist(y_prime,main = "Histogram of Average File Size",
     xlab='Average File Sizes in kB',xlim=c(1400,1900),ylim=c(0,700),breaks = 100)

# Calculating the summaries for Y'

mean_avg = mean(y_prime)
summary_avg = summary(y_prime)
standard_deviation_avg = sd(y_prime)
variance_avg = var(y_prime)
print(summary_avg)
cat("Standard Deviation: ", standard_deviation_avg)
cat("Variance: ", variance_avg)

# Calculating the maximal possible limit

max_possible_limit = qPareto(0.99, length(file_size),alpha_hat)
print(max_possible_limit)
```