HERIOT-WATT UNIVERSITY DUBAI
SCHOOL OF MATHEMATICAL AND COMPUTER SCIENCES

F21SA - Statistical Modelling and Analysis - 2023-24

Coursework

Student ID Number	Student Name	Programme
H00441124	Ilia	MS Data Science

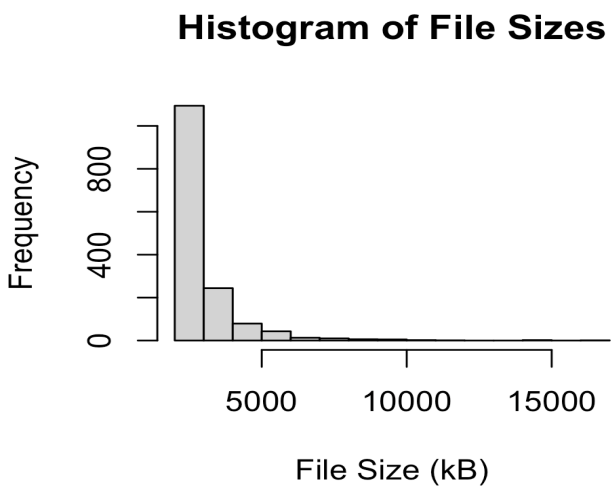
Introduction

The aim of this coursework is to practice parameter estimation techniques such as the Maximum Likelihood Estimation (MLE) and Method of Moments Estimation (MME) and to understand how to model real-world data with theoretical distributions, in this case, the Pareto distribution. Also, I investigated the provided dataset of file sizes using R by calculating major statistical characteristics and plotting the data with a histogram of distribution.

Workflow

I began by loading the dataset containing file sizes and plotted a histogram to visualize the distribution. The histogram showed that the data is skewed with a long tail, which is characteristic of the Pareto distribution.

Statistical metrics	Actual values
Min value	2000
1st quantile	2205
Median	2510
3rd quantile	3103
Max value	16728
Mean	2903
Standard deviation	1263



The histogram of the file sizes showed a highly skewed distribution, confirming the appropriateness of the Pareto distribution for modeling the data. Then I proceed with step by step calculation of estimators:

1. **Calculating the Alpha MLE Value (α):** Firstly, derive the likelihood function, then get a log-likelihood function and finally get the alpha value by taking the derivative of step 2 and equating it to zero. Resulting formula of alpha_mle and its value:

$$\hat{\alpha}_{\text{MLE}} = \frac{n}{\sum_{i=1}^{n} \ln(x_i) - n \ln(x_m)} = 3.16$$

2. **Calculating Fisher Information for Alpha:** Fisher information was calculated to understand the amount of information that our data sample holds about the alpha parameter. Resulting formula of Fisher information and its value:

$$I(\alpha) = \frac{n}{\alpha^2} = 149.22$$

3. **Calculating the Alpha Value using MME:** I also estimated the MME alpha parameter to compare with the MLE results. Resulting formula of alpha_mme and its value:

$$\alpha_{\text{MME}} = \frac{\bar{x}}{\bar{x} - x_m} = 3.21$$

In examining the alpha values obtained through both the Maximum Likelihood Estimation (MLE) and the Method of Moments Estimation (MME), I find that they are quite close — 3.16 for MLE and 3.21 for MME. These values suggest that the distribution of file sizes in our dataset is skewed, with smaller sizes than larger ones, which is typical for data that follows a Pareto distribution. The slight difference between the MLE and MME could be indicative of the different mathematical approaches each method takes. The MLE's slightly lower value suggests a slightly heavier tail, meaning there's a slightly higher chance of encountering larger-than-average file sizes.

However, there is no straight answer which estimator is the best one. MLE is generally preferred for its efficiency — especially with larger datasets — as it tends to give more 'weight' to values that are more likely. On the other hand, MME can be easier to compute and sometimes more robust against data anomalies. Given that our dataset isn't massive but isn't tiny either, both estimators provide valuable insight.

4. **Constructing a Confidence Interval for Alpha:** A 90% confidence interval was constructed around the MLE estimate of alpha to gauge the precision of our estimation. A 90% CI is given by:

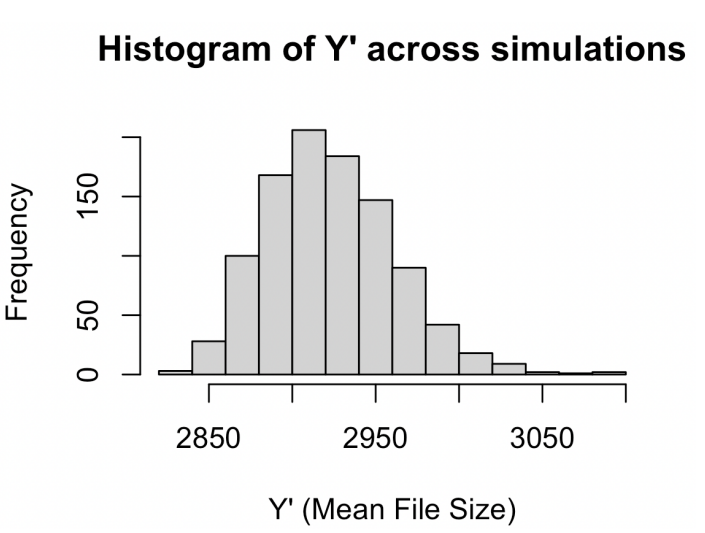
$$\left(\hat{\alpha}_{\text{MLE}} \pm z_{5\%} \cdot \frac{\hat{\alpha}_{\text{MLE}}}{\sqrt{n}} \right)$$

$$(3.04,\ 3.31)$$

The values of a 90% confidence interval confirm that alpha values from both estimators are correct.

5. **Simulating the Distribution of Predicted Mean File Size (Y'):** To understand the behavior of the mean file size, I simulated 1000 samples, each containing 1500 file sizes, and calculated the mean for each sample. Then plotted a histogram of these means to observe the distribution.

Statistical metrics	Actual values
Min value	2828
1st quantile	2895
Median	2920
3rd quantile	2948
Max value	3094
Mean	2923
Standard deviation	38.79



The histogram of the Y' values across simulations appeared to be normally distributed, as expected due to the support of the Central Limit Theorem stating that the distribution of sample means will tend to be normal regardless of the original distribution, as the number of samples increases.

Conclusion

The statistical analysis of file sizes using the Pareto distribution provided valuable insights into the nature of the data. The confidence interval for the alpha parameter indicated a reasonable level of precision in our estimates, and the simulation exercise illustrated the theoretical principles of statistical distributions in practice. My findings indicate that both MLE and MME can be used to estimate parameters of the Pareto distribution, but they may give different results. The MLE method provided a lower estimate for alpha, which could suggest a heavier tail in the distribution.

References

1. Dalgaard, P. (2008). Introductory Statistics with R. Springer
2. F20SA / F21SA Statistical Modelling and Analysis
3. RStudio Education. (2021). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.

Appendix:

- R code used for analysis
- Hand calculations

R script:

```
# Loading the data
data =
read.csv("/Users/ilya/Desktop/GitHub_Repositories/HW_University/Statistic/CW/filesizes.csv
")

# Plotting a histogram
hist(data$x, main="Histogram of File Sizes", xlab="File Size (kB)", ylab="Frequency")

# Calculating numerical summaries
summary(data$x)
mean_value = mean(data$x)
sd_value = sd(data$x)
print(mean_value)
print(sd_value)


# Exercise 2 (calculating alpha value from MLE)
n = dim(data)[1]
x_m = 2000
sum_of_ln_xi = sum(log(data$x))
alpha_MLE = n / (sum_of_ln_xi - (n * log(x_m)))


# Exercise 3 (calculating fisher information for alpha)
fisher_information = n / alpha_MLE**2

# Calculating distribution of alpha_hat
print(sprintf("Alpha_hat has approximately N(%.3f, %.3f)", alpha_MLE, 1 /
fisher_information))


# Exercise 4 (calculating alpha value from MME)
alpha_MME = mean_value / (mean_value - x_m)


# Exercise 5 (calculating a 90% CI for alpha)
lower_bound = alpha_MLE - 1.645 * (alpha_MLE / n**0.5)
upper_bound = alpha_MLE + 1.645 * (alpha_MLE / n**0.5)

cat(sprintf("A 90%% CI for alpha is (%f, %f)", round(lower_bound, 3), round(upper_bound,
3)))
```

```
# Exercise 6 (compering alpha value from MLE and MME)
print(alpha_MLE)
print(alpha_MME)


# Exercise 7 (Mean simulation)
library(VGAM)

# Number of simulations
num_simulations = 1000

# Array to store Y' for each simulation
Y_primes = numeric(num_simulations)

for (i in 1:num_simulations) {
  # Simulate file sizes for each sample
  file_sizes <- rpareto(1500, x_m, alpha_MLE)

  # Calculate Y' for this sample and store it
  Y_primes[i] = mean(file_sizes)
}

# Plot a histogram of the Y' values
hist(Y_primes, main="Histogram of Y' across simulations", xlab="Y' (Mean File Size)",
ylab="Frequency")

# Output numerical summaries of Y'
summary(Y_primes)
mean(Y_primes)
sd(Y_primes)
```

Hand calculations

Ex. 2 MLE α ; $f(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}}$; $n = 1500$

(S_1) likelihood function

$$L(\alpha, x) = f(x_1) f(x_2) \dots f(x_n) = \left(\frac{\alpha x_m^\alpha}{x_1^{\alpha+1}}\right)\left(\frac{\alpha x_m^\alpha}{x_2^{\alpha+1}}\right)\dots$$

$$\dots \left(\frac{\alpha x_m^\alpha}{x_n^{\alpha+1}}\right) = \frac{\alpha^n x_m^{\alpha n}}{(x_1 x_2 \dots x_n)^{\alpha+1}}$$

(S_2) log-likelihood function

$$L(\alpha) = \ln(L[\alpha]) = n \ln(\alpha) + n\alpha \ln(x_m) -$$

$$-(\alpha+1) \sum \ln(x_i)$$

(S_3) The MLE is the solution of:

$$\frac{dL}{d\alpha}(\hat{\alpha}) = 0 \iff \frac{n}{\hat{\alpha}} + n \ln(x_m) - \sum \ln(x_i) = 0$$

$$\frac{n}{\hat{\alpha}} = \sum \ln(x_i) - n \ln(x_m)$$

$$n = \hat{\alpha}\left(\sum \ln(x_i) - n \ln(x_m)\right)$$

$$\hat{\alpha} = \frac{n}{\sum \ln(x_i) - n \ln(x_m)}$$

Ex 3.

$$I(\alpha) = -E\left[\frac{d^2 \ln f(x)}{d\alpha^2}\right]$$ — second derivative

p.d.f of Pareto distribution; $f(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}}$

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

Distribution of $\tilde{\alpha}$

$$\tilde{\alpha} \sim N\left(\alpha, \frac{1}{I_n(\alpha)}\right) \Leftrightarrow \tilde{\alpha} \sim N\left(\alpha, \frac{\alpha^2}{n}\right)$$

This means that the MLE $\tilde{\alpha}$ of α is asymptotically normally distributed with mean(α) and variance $\left(\frac{\alpha^2}{n}\right)$

Ex 4.

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

The MME is the solution of $E(x) = \bar{x}$

$$\frac{\alpha x_m}{\alpha - 1} = \frac{\bar{x}}{1}$$

$$\alpha x_m = \bar{x}\alpha - \bar{x}$$

$$\bar{x} = \alpha(\bar{x} - x_m) \Leftrightarrow \alpha = \frac{\bar{x}}{\bar{x} - x_m}$$

Ex 5

from Ex 2. $\tilde{\alpha} = \frac{n}{\sum \ln(x_i) - n \ln(x_m)}$

A 90% CI is given by:

$$\left(\tilde{\alpha} \pm z_{5\%} \sqrt{\frac{1}{I(\tilde{\alpha})}}\right)$$

$$I(\alpha) = \frac{n}{\alpha^2}$$

$$\hat{\alpha} \sim N\left(\alpha, \frac{1}{I(\alpha)}\right) \Leftrightarrow \hat{\alpha} \sim N\left(\alpha, \frac{\hat{\alpha}^2}{n}\right) \rightarrow \text{ese}(\hat{\alpha}) = \sqrt{\frac{\hat{\alpha}^2}{n}} =$$

$$= \frac{\hat{\alpha}}{\sqrt{n}}$$

A 90% CI for λ is given by:

$$\left(\hat{\lambda} \pm 1.645 \cdot \frac{\hat{\lambda}}{\sqrt{n}}\right)$$