Bootstrapping - Earthquake and Service time datasets

Dominika Bakalarz

1 Introduction

This report presents an analysis of two datasets; the first one contains information about the amount of time (in days) that passed between the consecutive Earthquakes (of Magnitude higher than 6) between 1970 and 2009, while the second dataset provides the service time (in minutes) for 174 customers at a college snack bar.

2 Part A

This part is concerned with the Earthquakes dataset. A histogram summarising the data has been included in the question (and it is also presented in Figure 1). We note that both the shape of the distribution and the real-life meaning of the collected data are good reasons to use the exponential distribution model. Thus, in this part, we assume that the provided data comes from an $\mathrm{Exp}(\theta)$ distribution with probability density function $g_\theta(x) = \theta e^{-\theta x}$.

2.1 Bootstrap estimate of $\mathbb{V}\left(\widehat{\theta}_n\right)$

Since the MLE estimate $\hat{\theta}_n$ of the rate θ of an exponential distribution equals the inverse of the mean of all data samples, we have $\hat{\theta}_n = 0.055$. To calculate the bootstrap estimate $\widehat{\sigma}^2_{n,B}$ of the variance of $\hat{\theta}_n$, I used the following algorithm.

Algorithm 1. Algorithm for variance estimation:

```
• For $j = 1, \ldots, B$
    - Simulate $X_1^{*(j)}, \ldots, X_n^{*(j)}$ using the exponential distribution with rate $\hat{\theta}_n$
    - Evaluate $\hat{\theta}_n^{*(j)} = 1/\mathrm{mean}\left(X_1^{*(j)}, \ldots, X_n^{*(j)}\right)$
• Return the bootstrap variance estimate:
```

$$\widehat{\sigma}^2_{n,B} = \frac{1}{B}\sum_{j=1}^{B}\left(\hat{\theta}_n^{*(j)} - \frac{1}{B}\sum_{j=1}^{B}\hat{\theta}_n^{*(j)}\right)^2$$

I used n = 805 (number of data samples available in the Quakes dataset) and B = 1000, where B is the number of bootstrap iterations, and I obtained variance estimate $\widehat{\sigma}^2_{n,B} = 3.69 \times 10^{-6}$.

2.2 Bootstrap estimates $\hat{q}_n^{\mathrm{P}}(\alpha)$ of $q(\alpha)$

We are interested in estimating $q(\alpha)$ as $\hat{q}_n^{\mathrm{P}}(\alpha) = G^{-1}_{\hat{\theta}_n}(\alpha)$. To do it, I calculated relevant quantiles of an exponential distribution with rate $\hat{\theta}_n = 0.055$; obtained estimates are presented in Table 1.

Table 1: Parametric estimates $\hat{q}_n^{\mathrm{P}}(\alpha)$

value of α	$\hat{q}_n^{\mathrm{P}}(\alpha)$
0.1	1.90
0.25	5.18
0.5	12.50
0.75	25.00
0.9	41.54

Table 2: Confidence intervals for $q(\alpha)$

value of α	99% confidence interval for $q(\alpha)$
0.1	(1.15, 2.37)
0.25	(4.21, 5.99)
0.5	(10.91, 14.00)
0.75	(22.29, 27.77)
0.9	(36.34, 45.83)

2.3 Confidence Intervals for $q(\alpha)$

The 99% bootstrap pivotal confidence interval for $q(\alpha)$ is given by

$$C_n^* = \left[2\hat{q}_n^{\mathrm{P}}(\alpha) - \hat{q}_{0.995}^{q(\alpha)*}, 2\hat{q}_n^{\mathrm{P}}(\alpha) - \hat{q}_{0.005}^{q(\alpha)*}\right]$$

where $\hat{q}_{0.005}^{q(\alpha)*}$ and $\hat{q}_{0.995}^{q(\alpha)*}$ are the 0.005 and 0.995 quantiles of the bootstrap samples $\hat{q}_n^{*(1)}(\alpha), \ldots, \hat{q}_n^{*(B)}(\alpha)$ that were obtained using Algorithm 2.

Algorithm 2. Algorithm for variance estimation of $q(\alpha)$:

- For $j = 1, \ldots, B$
  - Simulate $X_1^{*(j)}, \ldots, X_n^{*(j)}$ using the exponential distribution with rate $\hat{\theta}_n$
  - Evaluate $\hat{q}_n^{*(j)}(\alpha)$ as the α quantile of $X_1^{*(j)}, \ldots, X_n^{*(j)}$ distribution
- Return the bootstrap variance estimate using the same formula as in Algorithm 1 and a list of values $\hat{q}_n^{*(1)}(\alpha), \ldots, \hat{q}_n^{*(B)}(\alpha)$.

We should note that the variance estimate obtained in Algorithm 2 is not required to calculate the pivotal confidence intervals, but it would be necessary to calculate confidence intervals using the normal approximation method, which I also calculated as a safety check; obtained values of can be found in the attached R code. The pivotal confidence intervals are presented in Table 2.

2.4 Comparison of the parametric and nonparametric approaches

In general, a parametric model is expected to outperform the nonparametric one, provided that it is specified correctly. To investigate how well the exponential model fits the data, I plotted both the histogram of the data and the density function on the same plot (Figure 1). Based on the plot, it is very likely that the data comes from this distribution. Moreover, we know that exponential distribution models the time between events in a process in which events occur continuously and independently at a constant average rate. Domain knowledge about Earthquakes would be useful to confirm if these assumptions are indeed correct in this specific case, but it is very likely that they are, i.e., we can assume that Earthquakes are independent and happen at a constant rate. Another argument in favour of this model comes from the Q-Q plot in Figure 2. We can see that exponential distribution is indeed a good fit for the data. (The point that is second from the right could potentially be classified as an outlier and removed, but since we have 804 data points, it would not affect any of the calculations.) Therefore, since the model is specified correctly, the parametric approach is more suitable for this dataset.

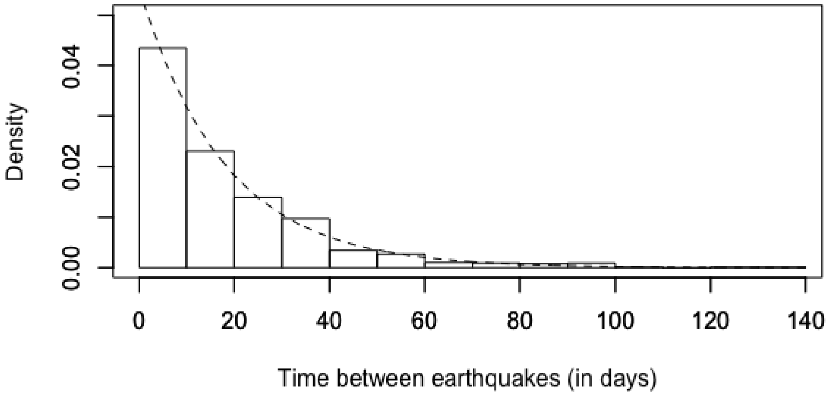

Figure 1: Histogram presents the Quakes data and the dashed line presents the density of Exp($\hat{\theta}_n$) distribution.

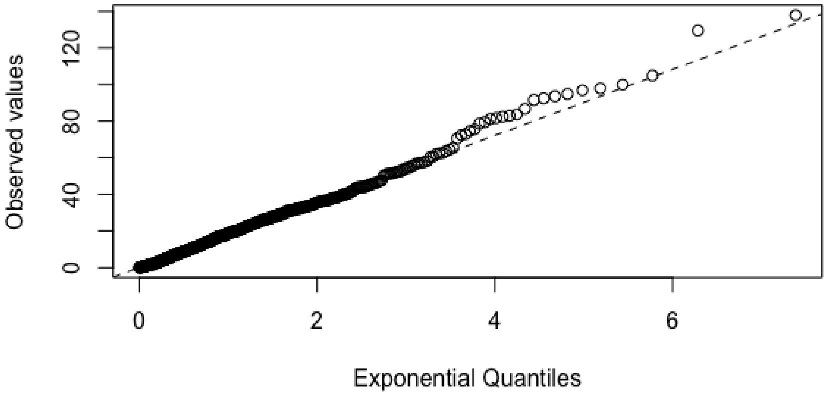

Figure 2: QQ plot for the Quakes data fitted against an exponential distribution. The dashed line has a slope $1/\hat{\theta}_n$ and 0 intercept.

Table 3: Bootstrap estimates of $h(\alpha)$

value of α	$\hat{\text{h}}^*(\alpha)$
0.1	0.118
0.25	0.264
0.5	0.488
0.75	0.768
0.9	0.903
0.005	0.003
0.995	0.992

2.5 Bootstrap estimate of $h(\alpha)$ and confidence interval for a future observation

We do not know the true distribution F, but we have the estimated distribution $F_{\hat{\theta}_n}$, that we used to provide the estimates $\hat{q}_n^{\mathrm{P}}(\alpha)$. Thus, sampling data directly from $F_{\hat{\theta}_n}$ will not help us that much. Instead, we can take all $\hat{\theta}_n^{*(1)}, \ldots, \hat{\theta}_n^{*(B)}$, and draw one sample each of the distributions $F_{\hat{\theta}_n^{*(1)}}, \ldots, F_{\hat{\theta}_n^{*(B)}}$, and then use all these new samples to estimate h(α), while table 3 presents bootstrap estimates of h(α) for a range of values α. This approach resulted in the following algorithm, which I used to calculate bootstrap estimates of h(α).

Algorithm 3. Algorithm for estimating $h(\alpha)$:

```
• For $j = 1, \ldots, B$
    - Sample $X_j^*$ from the exponential distribution with rate $\hat{\theta}_n^{*(j)}$
• Evaluate $\hat{h}^*(\alpha) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i^* < p(\alpha)}$
```

To estimate the 99% confidence interval for a future observation, I used Algorithm 2 to estimate $q(0.005)$ and $q(0.995)$ of the true underlying distribution F as quantiles of the estimated distribution $F_{\hat{\theta}_n}$. I obtained $\hat{q}(0.005) = 0.11$, and $\hat{q}(0.995) = 93.80$, what means that (0.11, 93.80) is the 99% confidence interval for a future observation.

3 Part B

This part is concerned with the Service dataset, which provides the service time (in minutes) for 174 customers at a college snack bar. A histogram summarising the data has been included in the question (and it is also presented in Figure 3).

3.1 Exploratory Data Analysis and proposed approach

As mentioned in Part A, in general, parametric models outperform the nonparametric ones, provided that the specified model corresponds well to the collected data. Thus, in order to decide between parametric and nonparametric approach, I tried to see how well we can fit a specific distribution to our data. Based on the shape of the histogram I decided to try the Gamma distribution. Obtained MLEs of shape and scale parameters are following: shape $\hat{\alpha} = 2.813$ and scale $\hat{\beta} = 0.247$. Figure 3 presents the histogram and the density of Gamma($\hat{\alpha}, \hat{\beta}$); we can see from it that assuming that our data comes from this distribution is suitable. Moreover, the gamma distribution is a generalization

Table 4: Parametric estimates $\hat{q}_n^{\mathrm{P}}(\alpha)$

value of α	$\hat{q}_n^{\mathrm{P}}(\alpha)$
0.1	0.24
0.25	0.39
0.5	0.61
0.75	0.91
0.9	1.25

of the exponential distribution that can be used to model the amount of time between events in an otherwise Poisson process in which the event rate is not necessarily constant, and our data (waiting time in the queue) could be model in this way. Therefore, I decided to use the parametric approach.

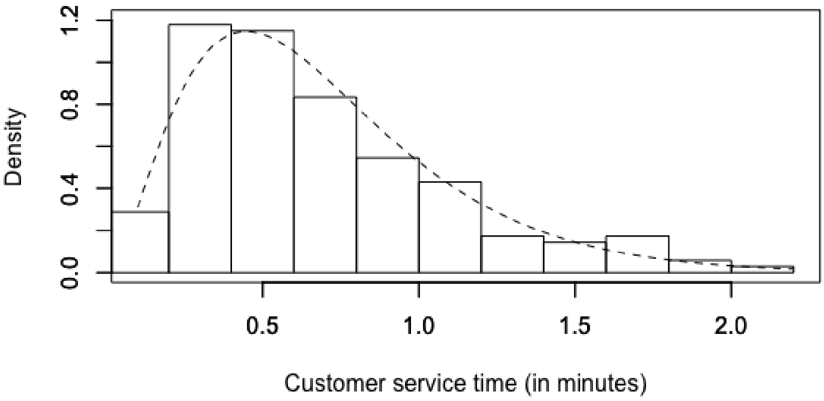

Figure 3: Histogram presents the Service data and the dashed line presents the density of Gamma$(\hat{\alpha}, \hat{\beta})$ distribution.

3.2 The estimates $\hat{q}_n^{\mathrm{P}}(\alpha)$

I estimated the quantiles using Gamma$(\hat{\alpha}, \hat{\beta})$ distribution; obtained estimates are presented in Table 4.

3.3 Confidence Intervals for $q(\alpha)$

To obtained the confidence intervals, I followed the same method as in Part A, i.e., I used a modified version of Algorithm 2, where instead of sampling $X_1^{*(j)}, \ldots, X_n^{*(j)}$ from the exponential distribution, I sampled the values from the Gamma$(\hat{\alpha}, \hat{\beta})$ distribution. Then, I used the same formula for the 99% bootstrap pivotal confidence intervals; obtained values are presented in Table 5.

Table 5: Confidence intervals for $q(\alpha)$

value of α	99% confidence interval for $q(\alpha)$
0.1	(0.18 0.30)
0.25	(0.32, 0.46)
0.5	(0.51, 0.69)
0.75	(0.82, 1.06)
0.9	(1.05, 1.40)

3.4 99% confidence interval for a future observation

To estimate the 99% confidence interval for a future observation, I used the modified version of Algorithm 2 to estimate $q(0.005)$ and $q(0.995)$ of the true underlying distribution Gamma(α, β) as quantiles of the estimated distribution Gamma$(\hat{\alpha}, \hat{\beta})$. I obtained $\hat{q}(0.005) = 0.074$, and $\hat{q}(0.995) = 2.171$, what means that (0.074, 2.171) is the 99% confidence interval for a future observation.

4 Attachments

R code is attached on the next page.