Hasselt University

Master of Statistics

Computer Intensive Methods

Assignment I

Students:
Christine Jani (1747124)
Joe Gwatsvaira (1747457)
Bache Emmanuel Bache (1747695)
Adama Kazienga (1747603)

Lecturers:
Prof.Ziv Shkedy

January 20, 2019

Contents

1 Introduction

There is an increase use of bootstrap methods, introduced by [Efron and Tibshirani, 1994], in statistical science. In general bootstrap is a resampling procedure for estimating the distributions of the statistic of interest, which acknowledges that there is only one sample. In other words, it provides an indirect method to assess the properties of the distribution underlying the sample and the statistic of interest (summary statistics, confidence interval and statistical inference). Several studies have shown that the bootstrap method is successful and is now being accepted as an alternative to asymptotic methods [Davison et al., 1997] . Parametric and non parametric were the two types of bootstrap methods and their respective generic algorithm were used in most part of the report with some extensions.

2 Methodology

2.1 The non-parametric bootstrap algorithm

This is the case where the random sample (size n) is drawn from an unknown probability distribution F. No distributional assumptions are made. The basic steps are:

1. Construct an empirical distribution F_n from the sample by placing a probability of $\frac{1}{n}$ at each point $x_1, ..., x_n$ of the sample.
2. From F_n draw a random sample of size n with replacement, this is a re-sample (bootstrap sample).
3. Calculate the statistic of interest $\hat{\theta}$ for this sample, called $\hat{\theta}^*$
4. Repeat step (2) and (3)B times in this study, $B = 1000$ was used
5. Construct the relative frequency histogram from the B bootstrap replicates $\hat{\theta}^*$ the distribution obtained is the bootstrap estimate of the sampling distribution of θ. This distribution is used to make inference about the parameter of interest θ

2.2 Parametric bootstrap algorithm

1. Assumption was made that the data comes from a known parametric distribution described by a set of parameters $\mathbf{\Theta}$.
2. Find the estimates $\hat{\mathbf{\Theta}}$ from the sample.
3. Use the Plug-in principle whereby the population parameters are replaced with the parameter estimates from the sample, hence resampling is done from a parametric distribution $F_{\hat{\mathbf{\Theta}}}$
4. The others steps (3),(4),(5) are similar to the algorithm of the non parametric bootstrap.

2.3 Confidence Intervals

A confidence interval (CI) is a type of interval estimate, computed from the statistics of the observed data, that has a higher probability of containing the true value of an unknown population parameter. There are two major types of confidence intervals used in this study the classical and the bootstrap confidence intervals.

2.3.1 Classical confidence Intervals

A symmetric $100(1-\alpha)\%$ confidence interval (CI) has the form: $\hat{\theta} \pm Z_{\frac{\alpha}{2}} * S.e(\hat{\theta})$, it is usually interpreted as through repeated sampling $100(1-\alpha)\%$ calculated confidence intervals are expected to contain the true value. The classical confidence interval assumes that the distribution of the test statistic is symmetric around zero. The major disadvantage of the Classical confidence interval is that it is usually misinterpreted and also in reality we cannot repeat the same experiment unless performing simulations studies, there is a need of bootstrap confidence intervals

2.3.2 Bootstrap Confidence Intervals

Confidence intervals are calculated from the distribution of the bootstrap statistics and this is the interval within which we are reasonably sure that it contains the true value of the population parameter θ of interest. The following are types of confidence intervals used in this study:

1. **Standard Normal CI**: It $\hat{\theta} \pm 1.96 \times S.e_{\hat{\theta}^*}$ where $S.e_{\hat{\theta}^*}$ is the standard deviation of the bootstrap statistic obtained from the bootstrap distribution.

2. **Percentile CI**: estimated from the 2.5% and 97.5% quantiles of the bootstrap distribution

3. **Bootstrap t CI**: estimated from the quantiles t_α and $t_{1-\alpha}$ of the bootstrap distribution of the replicates of the Z statistic where Z formula was $Z^*(b) = \frac{\hat{\theta}^* - \hat{\theta}}{S.e_{\hat{\theta}^*}}$ i.e $[\hat{\theta} - t_{1-\alpha}\hat{s.e}(\hat{\theta}; \hat{\theta} + t_\alpha \hat{s.e}(\hat{\theta})]$

3 Results

Question 1

(a) The classical and the bootstrap confidence interval for the correlation coefficient ρ was estimated. Let $x = Windspeed$ and $y = Temperature$. The sample correlation coefficient is calculated as follows:

$$\hat{\rho} = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} = -0.4579879 \tag{1}$$

The standard error for $\hat{\rho}$ given by:

$$S.e(\hat{\rho}) = \sqrt{\frac{1-r^2}{n-2}} = 0.07234241$$

Classical confidence interval is given by $\hat{\rho} \pm Z_{\alpha/2} S.e(\hat{\rho})$

$$\begin{aligned} &= && -0.4579879 \pm (1.96 \times 0.07234241) \\ &\implies && \rho \in [-0.599; -0.316] \end{aligned}$$

Interpretation: If we repeat the experiment a 100 times we expect that 95 of the time the correlation coefficient ρ between temperature and wind speed lies between $[-0.6; -0.32]$

(b) **Bootstrap confidence Intervals**

A bivariate normal distribution with the mean vector $\boldsymbol{\mu}$ and the variance covariance matrix $\boldsymbol{\Sigma}$ was assumed for the parametric bootsrap distribution. These were replaced with the plug-in estimates $\bar{\mathbf{X}} = (9.96, 77.88)$ and $\mathbf{S} = \begin{bmatrix} 12.41 & -15.27 \\ -15.27 & 89.59 \end{bmatrix}$ calculated from the observed sample. While for non parametric bootstrap, samples were drawn from the empirical distribution of the observed sample. Pairs were resampled since there is need to preserve the correlation between wind speed and temperature . The bootstrap statistic of interest is $\hat{\rho}^*$ calculated for $B = 1000$ bootstrap samples of size $n = 153$ using formulae (1). The different confidence intervals of ρ are shown in Table 1

Method	Type	Estimate	SE	CI Lower	CI Upper
Classical	-	-0.4579	0.07234241	-0.5998	-0.3162
	Percentile	-	0.0652	-0.561	-0.345
Non Parametric	Percentile	-	0.0643	-0.5743	-0.3199

Table 1: 95% CI for the population correlation coefficient ρ

Distributions of the bootstrap replicates

Figure 1 shows the distribution of the bootstrap replicates for parametric and non-parametric bootstrap. It can be observed that the distributions are symmetric around the observed correlation coefficient $\hat{\rho}$.

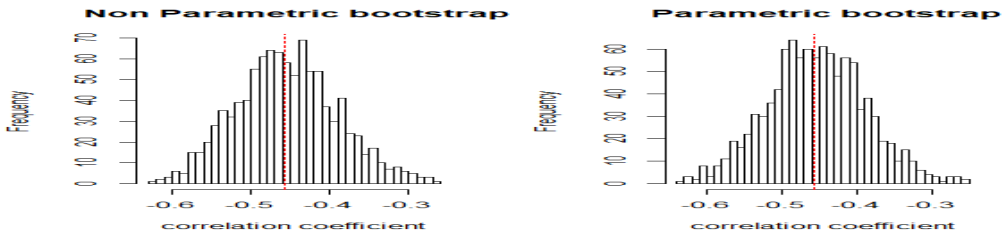

Figure 1: Distribution of bootstrap statistic

(c) **Algorithm and bootstrap replicates distribution of daily minimum temperature with wind speed less than 12.5**

From the airquality data, randomly selected sample of $n = 153$ pairs with replacement of wind speed and temperature was performed. The minimum daily temperature in which the wind speed was less than 12.5mph was kept. This is repeated 10000 times. The 95% confidence interval for the bootstrap replicates and the probability P($\theta < 62.5$) were estimated. The estimated confidence interval for θ using the nonparametric bootstrap ranged from 57 to 61. Moreover, the probability of P($\theta < 62.5$) F was 0.9977

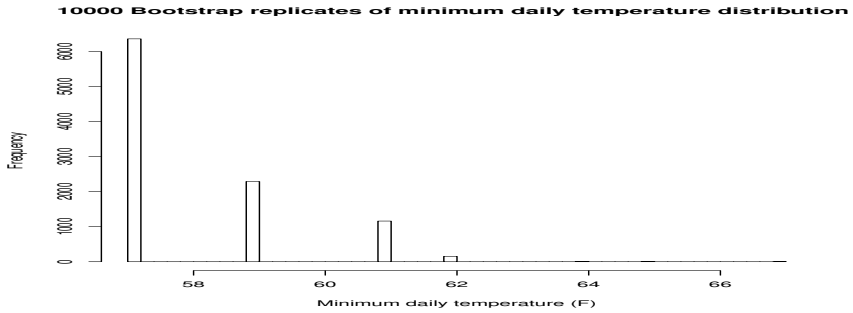

Figure 2: Histogram bootstrap replicates

Question 2

(a) **Classical method**- An odds ratio (OR) is a measure of association between an exposure (Gender) and an outcome (vote). The OR represents the odds that an outcome (vote) will occur given a particular exposure (gender), compared to the odds of the outcome (vote) occurring in the other exposure (gender) . The general formulae to calculate the odds ratio statistic is given by $OR = \frac{n_{11} \times n_{22}}{n_{12} \times n_{21}}$. The data used is displayed in the contingency Table 2. The study aimed at finding the classic and bootstrap confidence interval for the odds ratio.

	Gender	
Vote	Female	Male
Hillary Clinton	54	45
Trump	35	44

Table 2: Dataset

$$OR = \frac{n_{11} \times n_{22}}{n_{12} \times n_{21}} = \frac{54 \times 44}{45 \times 35} = 1.509$$

Confidence interval for $\log(OR)$ is constructed and then exponentiated to get the CI for OR $CI = \exp\left(\log(1.509) \pm 1.96 \times \sqrt{(\frac{1}{54} + \frac{1}{45}) + (\frac{1}{35} + \frac{1}{44})}\right)$
The estimated odds ratio using the classical methods is $1.508 (SE = 0.330)$, while the 95% confidence intervals for the odds ratio ranged from 0.83 to 2.73.

Interpretation : Through repeated experiment, 95 out of 100 confidence interval, would be expected to contain the true value of the odds ratio. Moreover, Since the confidence Interval contains one it implies that there is no association between vote and gender at a 5% level of significance.

(b) **Bootstrap Distribution for the Odds Ratio statistic**

From the data represented in tabular format transformation was made to the structure of the dataset into zero and one for each vote of female and male respectively for both candidates. Female was coded as one whereas it was zero for a male. So under Clinton, there were 54 'ones' and 45 'zeros' while for Trump it was 35 'ones' and 44 'zeros'. Random samples were drawn with replacement from each group (size 99 for Clinton and 79 for Trump), calculating the OR for each bootstrap sample and repeat this 1000 times (non parametric bootstrap). A percentile confidence interval was obtained by taking the 2.5% and the 97.5% of the bootstrap distribution It can be seen from Figure 3 that the bootstrap distribution was not symmetric around the sample OR, and have a long right tail. The bootstrap percentile confidence interval is given by $OR \in [0.829; 2.813]$ and the standard error was 0.507 .

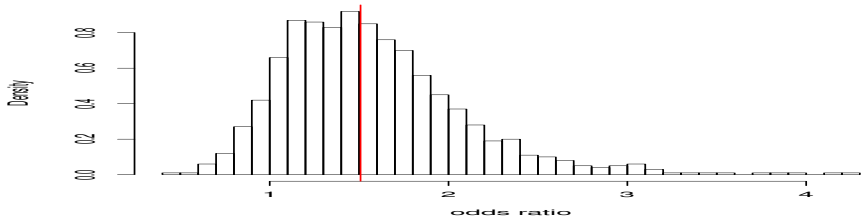

Figure 3: Histogram bootstrap replicates

(c) **Bootstrapping for Proportion of female voted for Hilary Clinton**

Of interest here is the fraction of the Females who voted for Hilary Clinton and denoted by θ. θ was approximated by the sample proportion

$$\hat{\theta} = \frac{X}{n} = \frac{\text{number of females who voted Clinton}}{\text{Total females in the sample}} \tag{2}$$

We denoted 54 females who voted for Hilary Clinton as 'ones' and 35 who voted Donald Trump as 'zeros'. We have a total $n = 89$ females in the study. B=1000 bootstrap replicates are taken with replacement from the empirical distribution F_n and at each bootstrap $\hat{\theta}^*$ is calculated. Histogram of bootstrap estimates Figure 4 and confidence intervals (percentile and bootstrap t) are shown in Table 3.

For the parametric bootstrap we resample from $\hat{F} = \text{Bin}(n, \hat{\theta})$ where $\hat{\theta} = 0.607$ and $n = 89$ is the plug in estimate of θ and . The distribution of the bootstrap replicate $\hat{\theta}^*$ for the two cases (parametric and non parametric) are shown in Figure 4.

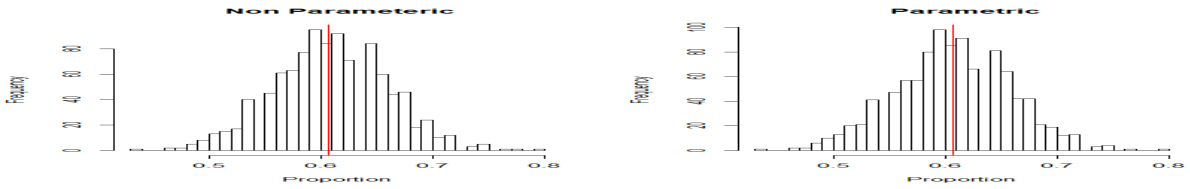

Figure 4: The Distribution of the bootstrap replicate for proportion

Bootstrap	Std.Error	Percentile	T-interval
Non-Parametric	0.051	[0.517:0.697]	[0.507:0.715]
Parametric	0.051	[0.517;0.685]	[0.507;0.715]

Table 3: Confidence Intervals for the bootstrap proportion

The 95% Bootstrap confidence intervals were almost similar starting from approximately 0.5 to 0.72. The true proportion of females who voted Hilary Clinton is between [0.5;0.72].

Question 3

(a) **Estimate the ratio ($\hat{R}$) in the data**
Ratio is a mathematical quantity that tells us how many times the ozone level in May is greater than ozone level in September. Let us denote the ozone level in May and September for a particular year by y_i and x_i respectively then $\hat{R} = \frac{y_i}{x_i}$ and the average ratio is denoted $\bar{\hat{R}}$ From the sample data $\bar{\hat{R}} = 1.138$ the ozone level in May is 1.138 times larger than the ozone level in September.

(b) **Use parametric and non parametric bootstrap to estimate the standard error and to construct 95% confidence interval for the $\hat{R}$**
A new data frame which consist of the relevant columns was created. Non parametric bootstrap was done by sampling from the Empirical distribution $\hat{F}_n$. To keep the structure of the data and maintain the relationship between ozone level for May and September for a particular year we take 1000 bootstrap replicates (n=30) of pairs with replacement. At each bootstrap replicate the ratio $\bar{\hat{R}}^*$ was estimated. For The parametric bootstrap we assume that the data comes from a log-normal(μ, σ^2) distribution since ozone level data is always positive. The plug-in estimates of the parameters are $\boldsymbol{\mu} = (23.08, 31.45)'$ and variance, $\mathbf{s^2} = (506.74, 654.61)$ for May and September respectively. Bootstrap replicates were drawn form multivariate lognormal. The approximation of the sampling distribution was found by plotting the histogram of bootstrap replicates Figure 5. The standard error was 0.274 using the non parametric bootstrap whereas it was 0.294 with the parametric bootstrap. Furthermore, the 95% percentile bootstrap confidence interval for parametric and non

parametric method were respectively $[0.730; 1.872]$ and $[0.663; 1.706]$. The bootstrap replicates distribution for classical, parametric and non parametric bootstrap are displayed in the figure 5. It can therefore, be observed that the ratio using all these methods is always positive, right skewed.

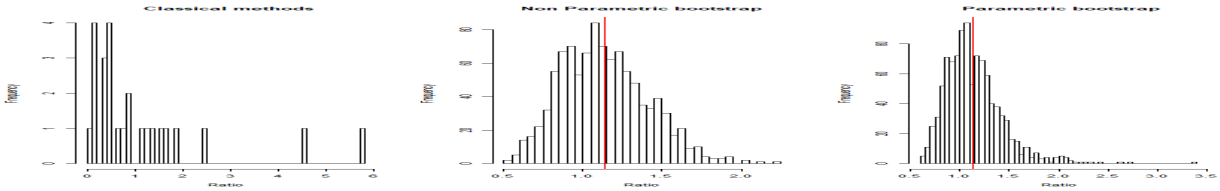

Figure 5: Distributions of ratios.

(c) Estimate the probability $\hat{R} < 0.7$

$$P = \frac{\sum_{i=1}^{B} R < 0.7}{B} = \frac{37}{1000} = 0.037 \text{ using non parametric bootstrap}$$
$$= \frac{16}{1000} = 0.016 \text{ using parametric bootstrap}$$

Question 4

In this question we considered a random sample of size $10(n = 10)$ from $N(\mu = 1, \sigma^2 = 1)$.

(a) **Show that that coverage probability of the classical 95% confidence interval for the mean was approximately 0.95.**
Random samples were drawn from N(1,1). The lower and upper bound were estimated $(\bar{X} \pm Z_{0,025} \times \frac{\sigma}{\sqrt{n}})$ and 100 CIs were constructed. The estimated coverage probability (P) is:

$$P = \frac{\text{Number of intervals containing } \mu}{R} = 0.94 \tag{3}$$

Figure 6 shows the plot of the confidence intervals. In red are the confidence intervals that does not include the true population parameter μ, and black are the intervals that include μ.

(b) **Show that coverage probability of the percentile bootstrap 95% confidence interval for the mean was approximately 0.95 using non parametric bootstrap**
Random sample of size 10 was drawn from N(1,1). From the empirical distribution of this sample, 1000 bootstrap samples each of size 10 were drawn with replacement and a bootstrap replicate $mean = \bar{X}$ computed for each. A 95% bootstrap percentile confidence interval was estimated. This iterative procedure is then repeated

$R = 100$ times to estimate the coverage probability of the percentile bootstrap confidence interval using the formula in Equation 3. Figure 6 shows that all the confidence intervals contained the true parameter (coverage probability = 100%) which may be due to small variability within the samples.

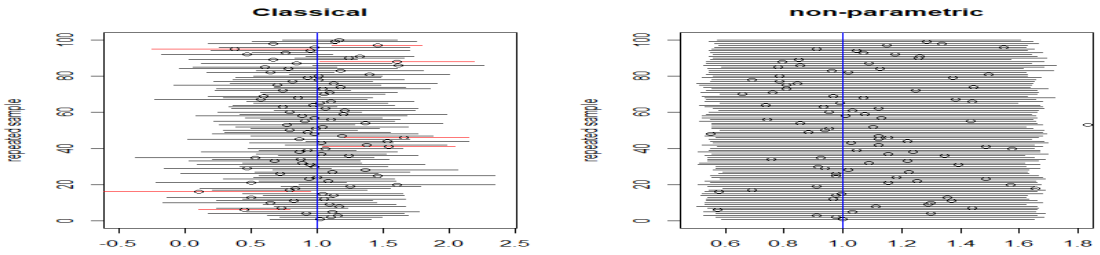

Figure 6: Bootstraps classical & non parametric (bootstrap percentile) CIs

(c) **Repeat on (a) and (b) when the sample was drown from Poisson($\lambda = 2$).** Algorithm applied is similar to the on above. Since our sample size is small we use the exact confidence intervals $\left[\frac{1}{2n}\chi^2_{(\lambda n, \frac{\alpha}{2})}; \frac{1}{2n}\chi^2_{(\lambda n, \frac{1-\alpha}{2}}\right]$. To calculate the classical coverage probability R=100 simulations of the confidence intervals are done and Equation 3 applied. Figure 7(a) for classical method shows a 99% coverage probability. For the non parametric bootstrap we draw one sample from a Poisson($\lambda = 2$); this is the original sample. B=1000 bootstrap samples are drawn with replacement from the empirical distribution of the original sample. $\hat{\lambda}$ is calculated at each bootstrap, then a 95% percentile confidence interval for the bootstrap replicates of λ determined. This is repeated R=100 times and count the number of confidence intervals that include $\lambda = 2$. Figure 7(b) shows a 100 percent coverage probability.

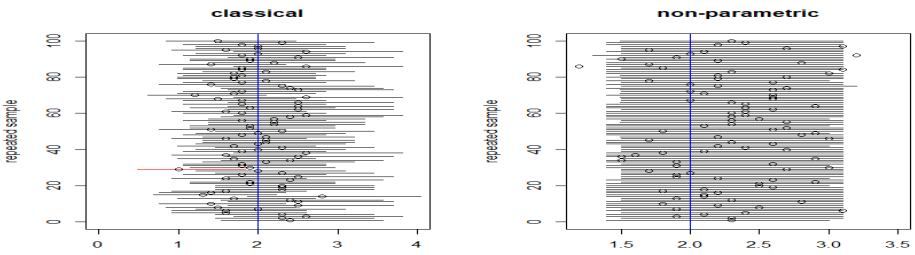

Figure 7: Classical & non parametric (bootstrap percentile) CIs for Poisson distribution

References

[Davison et al., 1997] Davison, A. C., Hinkley, D. V., et al. (1997). *Bootstrap methods and their application*, volume 1. Cambridge university press.

[Efron and Tibshirani, 1994] Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.