SB1 Practical 4

Candidate number: 1015294

May 3, 2019

1. Introduction

The following report consists of two main parts which aim to find bootstrap estimates of variances, quantiles, confidence and prediction intervals. The first part considers data consisting of times between earthquakes from 1970 to 2000. It studies the boostrap estimates of the variance of the maximum likelihood estimator of the rate of these eqrthquakes. Also different quantiles of an exponentially distributed random variable are estimated using parametric bootstrap analysis. Then two types of approximate confidence intervals for the quantiles are considered - normal confidence interval and pivotal confidence interval. In the end of this section an approximate 99% prediction confidence interval for a future observation is deduced.
The second part of the report examines the service time of customers at a college snack bar. The aim is to obtain estimators for the quantiles, find confidence intervals for these quantiles and finally discover a prediction confidence interval. In order to accomplish that, firstly the data is examined so that a parametric or non parametric approach is chosen. Secondly, after giving estimation for the quantiles, 99% pivotal confidence intervals are provided.
The last two sections give a summary of the report and the R code used for all computations.

2. Part A

2.1 Bootstrap estimate of $\mathbf{V}(\hat{\theta}_n)$.

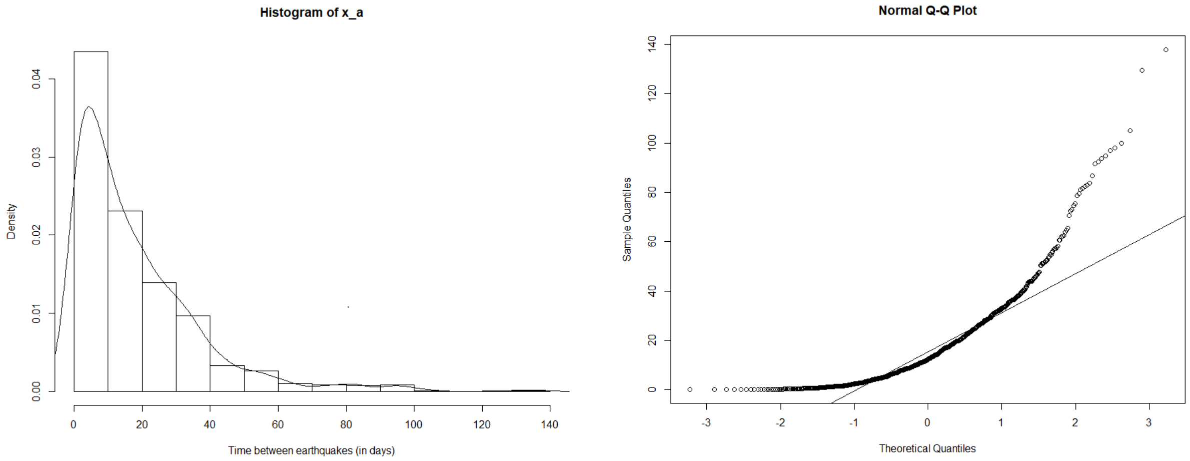
2.1.1 Parametric or non-parametric bootstrap. In the beginning we will examine the given Earthquakes data in order to determine whether to apply parametric or non-parametric approach.As shown below in Figure 1 (a), the density of the data resembles the one of an exponential distribution. Furthermore, considering the graph of the data points in the Normal QQ-plot in figure 1(b), we have more evidence that time between earthquakes follows an exponential distribution. Taking in account that we have $n = 805$ data points, we can safely assume that the cumulative distribution of the given data is of an exponential random variable.

2.1.2 Parametric bootstrap estimate of the variance. Consider a sample $X_1, X_2, ..., X_n$ of independent and identically distributed random variables which follow exponential distribution with cumulative distribution $F = G_\theta$ and probability density function $g_\theta(x) = \theta e^{-\theta x}$. Assume $\hat{\theta}_n$ is the maximum likelihood estimator of θ and $\mathrm{V}(\hat{\theta}_n)$ is its variance. The likelihood of the random sample is:

$$L(\boldsymbol{x}, \theta) = \prod_{i=1}^{n} g_\theta(x_i) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_{i=1}^{n} x_i}$$

Then the log-likelihood is:

$$l(\boldsymbol{x}, \theta) = nlog(\theta) - \theta \sum_{i=1}^{n} x_i \text{ and the derivative is } \frac{dl(\boldsymbol{x}, \theta)}{d\theta} = \frac{n}{\theta} - \sum_{i=1}^{n} x_i$$



(a)

(b)

Figure 1: (a) Histogram: Time between earthquakes vs Density
(b) Normal QQ-plot

The derivative of the log-likelihood is zero at $\theta = \hat{\theta}_n = \dfrac{n}{\sum_{i=1}^{n} x_i} = \dfrac{1}{\bar{x}}$, which is a maximum because the log-likelihood in our case is a concave function. Here with the Earthquakes data we have $\hat{\theta}_n = 0.0554$ *(R code in the Appendix A1).*

Now in order to compute the bootstrap estimate of V$(\hat{\theta}_n)$ we shall simulate $X_1^{*(j)}, X_2^{*(j)}, ..., X_n^{*(j)}, j = 1, .., B$ from a distribution with cumulative distribution function $F_{\hat{\theta}_n}$ of an exponential random variable with rate $\hat{\theta}_n > 0$. Then we can make the first approximation of V$(\hat{\theta}_n)$ by $V_{F_{\hat{\theta}_n}}(\hat{\theta}_n^{*(j)})$, where $\hat{\theta}_n^{*(j)} = \dfrac{1}{\sum_{i=1}^{n} X_i{}^{*(j)}} = \dfrac{1}{\overline{X}^{*(j)}}$ The second approximation we shall make is using the Monte Carlo method: $V_{F_{\hat{\theta}_n}}(\hat{\theta}_n) \approx \dfrac{1}{B}\sum_{j=1}^{B}\Big(\dfrac{1}{\overline{X}^{*(j)}} - \dfrac{1}{B}\sum_{j=1}^{B}\dfrac{1}{\overline{X}^{*(j)}}\Big)^2$, which is the bootstrap estimate of the variance of the MLE of the rate θ. Computing this in R we obtain $V_{F_{\hat{\theta}_n}}(\hat{\theta}_n) \approx 3.74511e - 06$ *(R code in Appendix A2).*

2.2 Parametric estimator for the α quantile.

We are given that the parametric estimator is $\hat{q}_n^P(\alpha) = G_{\hat{\theta}_n}^{-1}(\alpha) = inf\{x : G_{\hat{\theta}_n}(x) > \alpha\}$, where $G_{\hat{\theta}_n}(x) = \hat{\theta}_n e^{-\hat{\theta}x}$. After simple calculation we derive $G_{\hat{\theta}_n}^{-1}(\alpha) = -\frac{1}{\hat{\theta}_n}log(1-\alpha) = -\bar{X}log(1-\alpha)$. Therefore we have the following results: $\hat{q}_n^P(0.1) = 0.1054\bar{X}$ $\hat{q}_n^P(0.25) = 0.2877\bar{X}$ $\hat{q}_n^P(0.5) = 0.6931\bar{X}$ $\hat{q}_n^P(0.75) = 1.3863\bar{X}$ $\hat{q}_n^P(0.9) = 2.3026\bar{X}$. Substituting with $\bar{X} = 18.0388$ we conclude:
$\hat{q}_n^P(0.1) = 1.900574$
$\hat{q}_n^P(0.25) = 5.189430$
$\hat{q}_n^P(0.5) = 12.503521$
$\hat{q}_n^P(0.75) = 25.007043$
$\hat{q}_n^P(0.9) = 41.535799$ *(R code in Appendix A3).*

2.3 Confidence intervals

Firsly, we shall find the normal 99% confidence intervals for $q(\alpha)$. Observe that $E(\hat{q}_n^P(\alpha)) = E(-\bar{X}log(1-\alpha)) = -\frac{1}{\theta}log(1-\alpha) = G_{\theta}^{-1}(\alpha) = q(\alpha)$. Thus, from the Central Limit Theorem $\dfrac{\hat{q}_n^P(\alpha) - q(\alpha)}{\sqrt{V_F(\hat{q}_n^P(\alpha))}}$ converges in distribution to $N(0,1)$. We can estimate $\sqrt{V_F(\hat{q}_n^P(\alpha))}$ by the boostrap estimate of the variance $\sqrt{V_{F_{\hat{\theta}_n}}(\hat{q}_n^P(\alpha))}$ (the calculation of which is very similar to the one in 2.1 but instead of bootstrapping MLE we bootstrap $\alpha-$quantiles). The values of $\sqrt{V_{F_{\hat{\theta}_n}}(\hat{q}_n^P(\alpha))}$ for $\alpha = 0.1, 0.25, 0.5, 0.75, 0.9$ are $0.0675, 0.1838, 0.4351, 0.8869, 1.4697$, re-

spectively. The R code for these results is presented in the *Appendix section A4*. Now we have that $\left(\hat{q}_n^P(\alpha) - z_{\frac{\bar{\alpha}}{2}}\sqrt{V_{F\hat{\theta}_n}(\hat{q}_n^P(\alpha))}, \hat{q}_n^P(\alpha) + z_{\frac{\bar{\alpha}}{2}}\sqrt{V_{F\hat{\theta}_n}(\hat{q}_n^P(\alpha))}\right)$ is an approximate $1-\bar{\alpha}$ normal confidence interval for $q(\alpha)$. Using R we are able to obtain the following results:
99% CI when $\alpha = 0.10 : [1.726782, 2.074366]$

99% CI when $\alpha = 0.25 : [4.715994, 5.662867]$
99% CI when $\alpha = 0.50 : [11.382851, 13.624191]$
99% CI when $\alpha = 0.75 : [22.722589, 27.291496]$
99% CI when $\alpha = 0.90 : [37.750140, 45.321457]$
(R code in Appendix A5)

Secondly, we compute the pivotal 99% confidence intervals for $q(\alpha)$. The 99% bootstrap pivotal confidence interval is given by: $C = [2\hat{q}_n^P(\alpha) - \hat{q}^{q*}_{0.995}, 2\hat{q}_n^P(\alpha) - \hat{q}^{q*}_{0.005}]$, where $\hat{q}^{q*}_{\bar{\alpha}}$ is the $\bar{\alpha}$ quantile of the bootstrap samples $(\hat{q}_n^{P(1)}(\alpha)), \hat{q}_n^{P(2)}(\alpha)), .., \hat{q}_n^{P(B)}(\alpha))$, where $\hat{q}_n^{P(k)}(\alpha) = -\frac{\sum_{i=1}^n X_i^{*(k)}}{n} log(1-\alpha)$, for $k = 1, ..., B$. Now if we plug in $\alpha = 0.1, 0.25, 0.5, 0.75, 0.9$ we obtain the following results:
99% CI when $\alpha = 0.1 : [0.2108\bar{X} - \hat{q}^{q*}_{0.995}, 0.2108\bar{X} - \hat{q}^{q*}_{0.005}]$
99% CI when $\alpha = 0.25 : [0.5754\bar{X} - \hat{q}^{q*}_{0.995}, 0.5754\bar{X} - \hat{q}^{q*}_{0.005}]$
99% CI when $\alpha = 0.5 : [1.3862\bar{X} - \hat{q}^{q*}_{0.995}, 1.3862\bar{X} - \hat{q}^{q*}_{0.005}]$
99% CI when $\alpha = 0.75 : [2.7726\bar{X} - \hat{q}^{q*}_{0.995}, 2.7726\bar{X} - \hat{q}^{q*}_{0.005}]$
99% CI when $\alpha = 0.9 : [4.6052\bar{X} - \hat{q}^{q*}_{0.995}, 4.6052\bar{X} - \hat{q}^{q*}_{0.005}]$
When we substitute the 0.005 and 0.995 quantiles of the bootstaped samples we derive:
99% CI when $\alpha = 0.10 : [1.724940, 2.068059]$
99% CI when $\alpha = 0.25 : [4.706729, 5.633281]$
99% CI when $\alpha = 0.50 : [11.346770, 13.595526]$
99% CI when $\alpha = 0.75 : [22.698440, 27.240694]$
99% CI when $\alpha = 0.90 : [37.686340, 45.147796]$
R code in Appendix A6.
We finish this section by noting that the 99% pivotal and normal confidence interval are almost identical, with the pivotal ones being a little bit tighter.

2.4 Parametric vs Nonparametric approach.

In this case we would prefer to use the parametric approach because we saw that the given data $X_1, X_2, ..., X_n$ seems to come from an exponential distribution, as seen in the beginning of Part A. Hence, we use the fitted cdf $F_{\hat{\theta}_n}$ instead of the empirical cdf in order to obtain better bootstrapped approximations as above.

2.5 Parametric bootstrap estimator of $h(\alpha)$.

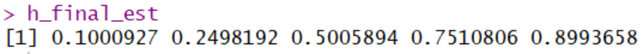
Define $h(\alpha) = P_F(X_{n+1} \leq \hat{q}_n^P(\alpha))$. Assume the future observation X_{n+1} follows the same distribution as the observed data $X_1, X_2, ..., X_n$, i.e X_{n+1} has a cdf of an exponential random variable. Further, we shall simulate $X^{*(1)}_{n+1}, X^{*(2)}_{n+1}, ..., X^{*(B)}_{n+1}$ from the cdf $F_{\hat{\theta}_n}$ of exponential distribution with rate $\hat{\theta}_n = \frac{1}{\bar{X}}$. Then we can define the parametric bootstrap estimator of $h(\alpha)$ to be:

$$\hat{h}_{n,B}(\alpha) = \tfrac{1}{B}\textstyle\sum_{j=1}^{B} P_{F\hat{\theta}_n}(X^{*(j)}_{n+1} \leq \hat{q}_n^P(\alpha))$$

In fact the approximations we have made are illustrated below:

$$\alpha = P_F(X_{n+1} \leq q(\alpha)) \overset{\mathrm{q(\alpha)\approx\hat{q}_n^P(\alpha)}}{\approx} P_F(X_{n+1} \leq \hat{q}_n^P(\alpha)) \overset{\mathrm{F\approx F_{\hat{\theta}_n}}}{\approx}$$
$$P_{F_{\hat{\theta}_n}}(X_{n+1} \leq \hat{q}_n^P(\alpha)) \overset{\mathrm{MonteCarlo}}{\approx} \hat{h}_{n,B}(\alpha)$$

Indeed the values we obtain for $\hat{h}_{n,B}(\alpha)$ are 0.1000927, 0.2498192, 0.5005894, 0.7510806, 0.8993658 for $\alpha = 0.1, 0.25, 0.5, 0.75, 0.9$ respectively. *(R code in Appendix A7)*

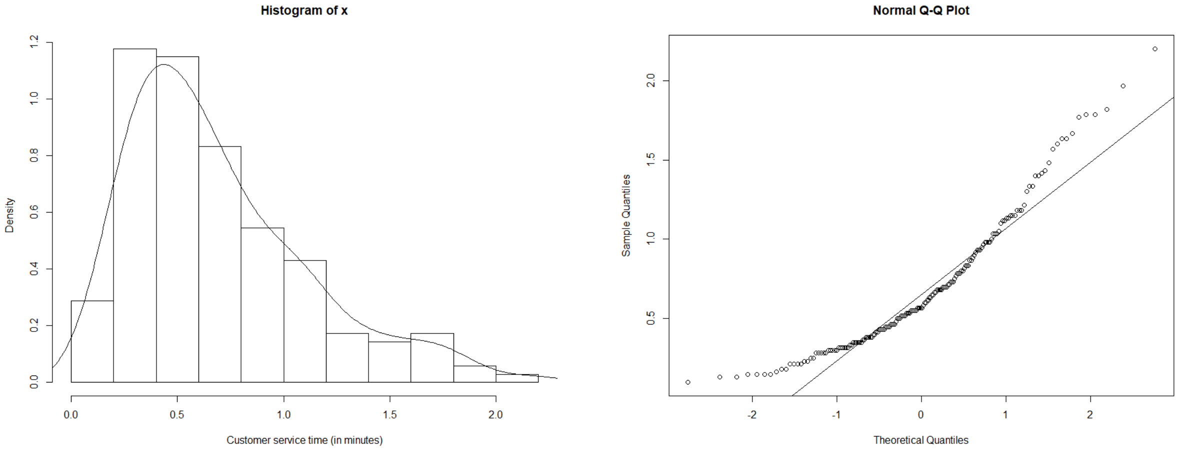


```
> h_final_est
[1] 0.1000927 0.2498192 0.5005894 0.7510806 0.8993658
```

2.6 Approximate 99% prediction confidence interval. By the approximate equations in 2.5 and assuming X_{n+1} comes from the same distribution as the previous observations $X_1, X_2..., X_n$ we can conclude that $(\hat{q}_n^P(0.005), \hat{q}_n^P(0.995)) = 0.0904, 95.5751$ is an approximate 99% confidence interval for the future observation X_{n+1}.

3. Part B

3.1 Parametric vs non parametric approach The given dataset consists of service time in minutes for 174 customers at a college snack bar. In order to decide whether to use parametric or non parametric estimators we shall examine carefully the data and find out if it follows a certain distribution. Considering the histogram in Figure 1 below, it seems that the data might be drawn from a Gamma distribution The density curve on the histogram clearly resembles a probability density function of a Gamma distributed random variable. Furthermore, the QQ-plot in Figure 2 suggests that the service time for the 174 customers indeed follows Gamma distribution. Taking in account these two figures we will proceed with parametric approach and assume that the service times $X_1, X_2, ..., X_n$ are independent and identically distributed from $\Gamma(a, b)$.



(a) (b)

Figure 1: (a) Histogram: Service time vs Density; (b) QQ-plot

3.2 Estimates for the parameters a and b. In order to find estimators for a and b we shall apply the method of moments. We already know that the first and second moment of a gamma distribution satisfy: $E[\Gamma(a, b)] = \frac{a}{b}$ and $V[\Gamma(a, b)] = \frac{a}{b^2}$. Therefore, the moment equations for Gamma distributed $X_1, X_2, ..., X_n$ are:

$$\frac{\sum_{i=1}^n X_i}{n} = \frac{a}{b} \text{ and } \frac{\sum_{i=1}^n X_i^2}{n} = \frac{a(a+1)}{b^2}$$

After solving these two equations with two unknowns we obtain the moment estimators:

$$\hat{a} = \hat{b}\bar{X} \text{ and } \hat{b} = \frac{\bar{X}}{n^{-1}\sum_{i=1}^n X_i^2 - \bar{X}^2}$$

Hence with our data we deduce $\hat{a} = 2.6701$ and $\hat{b} = 3.8424$ *(R code in Appendix B1)*. Alternatively we could estimate $a$and b using there maximum likelihood estimator, which for Gamma distribution we compute numerically to derive $a_{MLE} \approx 2.81$ and $b_{MLE} \approx 4.05$ *(R code in Appendix* B2). However in the rest of the report we shall use the moment estimators for a and b.

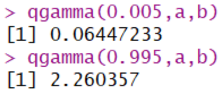
3.3 Parametric estimator for $q(\alpha)$. Define $F_{a,b}$ to be the cumulative distribution function of a gamma distribution with parameters a and b. Also let $F_{\hat{a},\hat{b}}$ be an estimator for $F_{a,b}$, where $\hat{a}$ and $\hat{b}$ are defined as in 3.2. Now we can give a parametric estimator for $q(\alpha)$:

$$\hat{q}^P(\alpha) = F^{-1}_{\hat{a},\hat{b}}(\alpha)$$

Using the quantile function in R **qgamma(**$\alpha, \hat{a}, \hat{b}$ **)** and for $\alpha = 0.1, 0.25, 0.5, 0.75, 0.9$ we have the following results:
$\hat{q}^P(0.1) = 0.2352$
$\hat{q}^P(0.25) = 0.3822$
$\hat{q}^P(0.5) = 0.6104$
$\hat{q}^P(0.75) = 0.9163$
$\hat{q}^P(0.9) = 1.2649$
(R code in Appendix B3)

3.4 Confidence interval for $q(\alpha)$. Simulate datasets $X_1^{*(j)}, X_2^{*(j)}, ..., X_n^{*(j)}$ for each $j = 1...B$ from $\Gamma(\hat{a}, \hat{b})$. For each simulation j obtain a bootstap sample $\hat{q}^{P(j)}(\alpha) = F^{-1}_{\hat{a},\hat{b}}(\alpha)$ (note that $F^{-1}_{\hat{a},\hat{b}}(\alpha)$ depends on $X_1^{*(j)}, X_2^{*(j)}, ..., X_n^{*(j)}$ because $\hat{a}$ and $\hat{b}$ depend on $X_1^{*(j)}, X_2^{*(j)}, ..., X_n^{*(j)}$). Now denote with $\hat{q}^{q_*}_{\bar{\alpha}}$ the $\bar{\alpha}$ quantile of the bootstrap sample $(\hat{q}^{P(1)}(\alpha), \hat{q}^{P(2)}(\alpha), ..., \hat{q}^{P(B)}(\alpha))$. Then the 99% bootstrap pivotal confidence interval is given by: $C = [2\hat{q}^P(\alpha) - \hat{q}^{q_*}_{0.995}, 2\hat{q}^P(\alpha) - \hat{q}^{q_*}_{0.005}]$. Using R we are able to obtain the following results:
99% CI when $\alpha = 0.10 : [0.1724, 0.2260]$
99% CI when $\alpha = 0.25 : [0.3040, 0.3970]$
99% CI when $\alpha = 0.50 : [0.5349, 0.6275]$
99% CI when $\alpha = 0.75 : [0.8045, 0.9151]$
99% CI when $\alpha = 0.90 : [1.0859, 1.2633]$
(R code in Appendix B4)

3.5 Approximate 99% prediction confidence interval. It is reasonable to assume that the future observation X_{n+1} has $\Gamma(\hat{a}, \hat{b})$ distribution. Then $(\hat{q}^{P(j)}(0.005), \hat{q}^{P(j)}(0.995)) = (0.0645, 2.2604)$ is an approximate 99% confidence interval for X_{n+1} *(R code in Appendix B6)*.

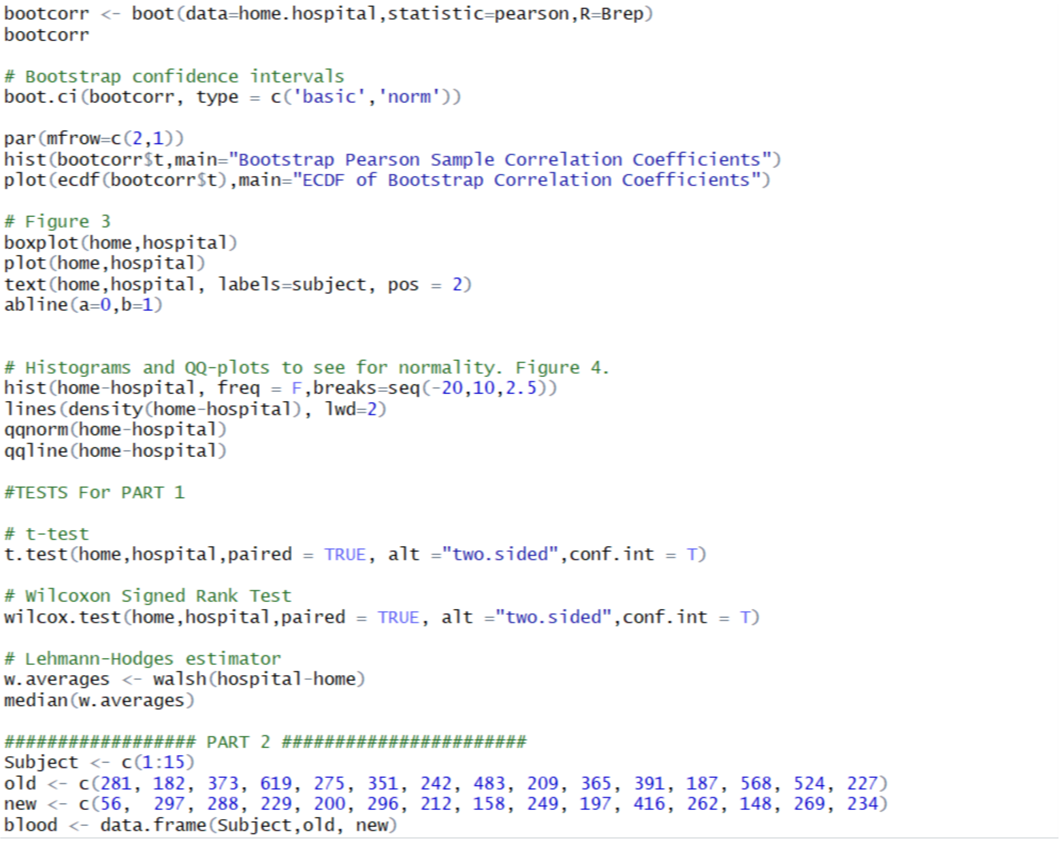


```
> qgamma(0.005,a,b)
[1] 0.06447233
> qgamma(0.995,a,b)
[1] 2.260357
```

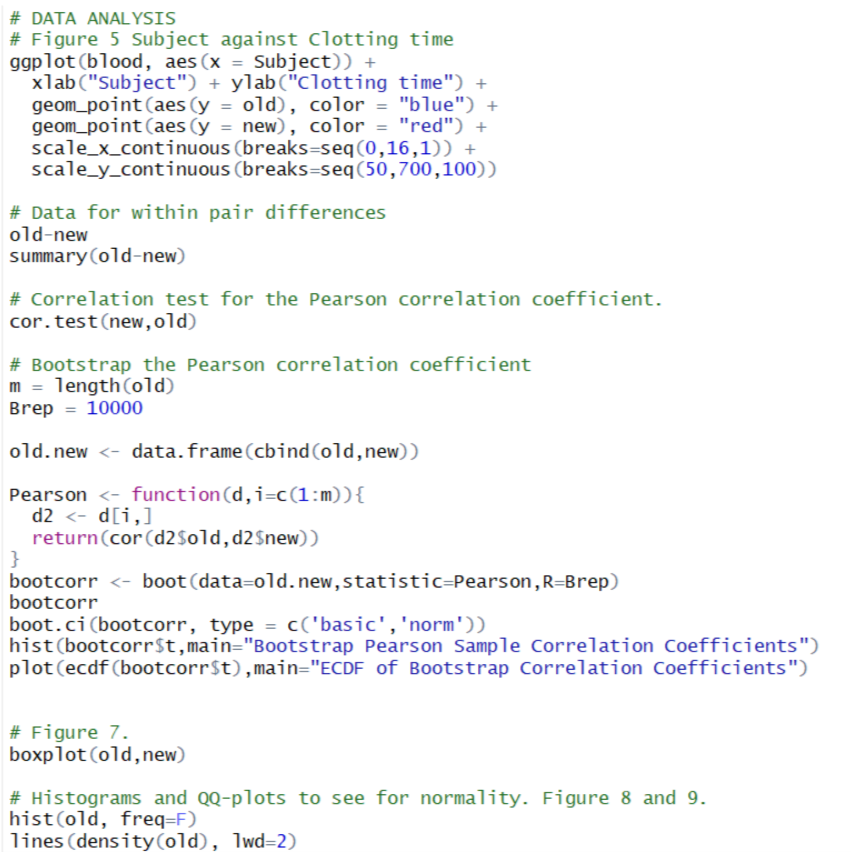
4. Conclusion

In the beginning of part B we decided to use a parametric approach because the histogram and QQ-plot of the given data suggested that the data follows a Gamma distribution$\Gamma(a, b)$. Further, we obtained estimators for the parameters $aandb$ of the Gamma distribution by applying the method of moments estimation - $\hat{a} = 2.6701$ and $\hat{b} = 3.8424$. Using these moment estimators, a parametric estimator for the quantile $q(\alpha)$ was found. Then by using the Monte Carlo method a bootstrap pivotal confidence interval was obtained. Finally, using the estimated quantiles, a prediction confidence interval for future observation X_{n+1} was given - $C = (0.0645, 2.2604)$.

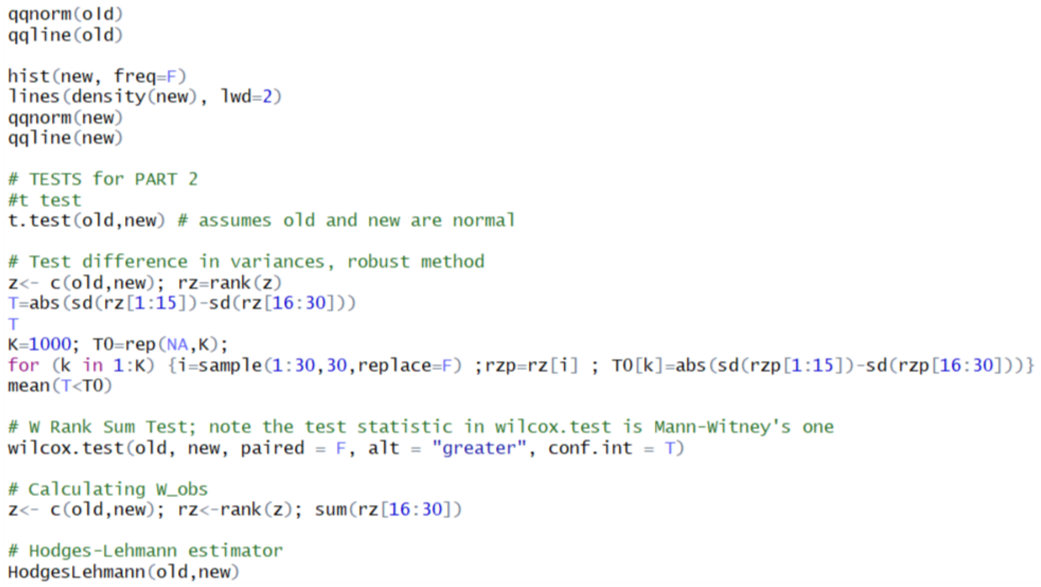
5. Appendix



```
bootcorr <- boot(data=home.hospital,statistic=pearson,R=Brep)
bootcorr

# Bootstrap confidence intervals
boot.ci(bootcorr, type = c('basic','norm'))

par(mfrow=c(2,1))
hist(bootcorr$t,main="Bootstrap Pearson Sample Correlation Coefficients")
plot(ecdf(bootcorr$t),main="ECDF of Bootstrap Correlation Coefficients")

# Figure 3
boxplot(home,hospital)
plot(home,hospital)
text(home,hospital, labels=subject, pos = 2)
abline(a=0,b=1)


# Histograms and QQ-plots to see for normality. Figure 4.
hist(home-hospital, freq = F,breaks=seq(-20,10,2.5))
lines(density(home-hospital), lwd=2)
qqnorm(home-hospital)
qqline(home-hospital)

#TESTS For PART 1

# t-test
t.test(home,hospital,paired = TRUE, alt ="two.sided",conf.int = T)

# Wilcoxon Signed Rank Test
wilcox.test(home,hospital,paired = TRUE, alt ="two.sided",conf.int = T)

# Lehmann-Hodges estimator
w.averages <- walsh(hospital-home)
median(w.averages)

################## PART 2 #######################
Subject <- c(1:15)
old <- c(281, 182, 373, 619, 275, 351, 242, 483, 209, 365, 391, 187, 568, 524, 227)
new <- c(56,  297, 288, 229, 200, 296, 212, 158, 249, 197, 416, 262, 148, 269, 234)
blood <- data.frame(Subject,old, new)
```



```
# DATA ANALYSIS
# Figure 5 Subject against Clotting time
ggplot(blood, aes(x = Subject)) +
  xlab("Subject") + ylab("Clotting time") +
  geom_point(aes(y = old), color = "blue") +
  geom_point(aes(y = new), color = "red") +
  scale_x_continuous(breaks=seq(0,16,1)) +
  scale_y_continuous(breaks=seq(50,700,100))

# Data for within pair differences
old-new
summary(old-new)

# Correlation test for the Pearson correlation coefficient.
cor.test(new,old)

# Bootstrap the Pearson correlation coefficient
m = length(old)
Brep = 10000

old.new <- data.frame(cbind(old,new))

Pearson <- function(d,i=c(1:m)){
  d2 <- d[i,]
  return(cor(d2$old,d2$new))
}
bootcorr <- boot(data=old.new,statistic=Pearson,R=Brep)
bootcorr
boot.ci(bootcorr, type = c('basic','norm'))
hist(bootcorr$t,main="Bootstrap Pearson Sample Correlation Coefficients")
plot(ecdf(bootcorr$t),main="ECDF of Bootstrap Correlation Coefficients")


# Figure 7.
boxplot(old,new)

# Histograms and QQ-plots to see for normality. Figure 8 and 9.
hist(old, freq=F)
lines(density(old), lwd=2)
```



```
qqnorm(old)
qqline(old)

hist(new, freq=F)
lines(density(new), lwd=2)
qqnorm(new)
qqline(new)

# TESTS for PART 2
#t test
t.test(old,new) # assumes old and new are normal

# Test difference in variances, robust method
z<- c(old,new); rz=rank(z)
T=abs(sd(rz[1:15])-sd(rz[16:30]))
T
K=1000; T0=rep(NA,K);
for (k in 1:K) {i=sample(1:30,30,replace=F) ;rzp=rz[i] ; T0[k]=abs(sd(rzp[1:15])-sd(rzp[16:30]))}
mean(T<T0)

# W Rank Sum Test; note the test statistic in wilcox.test is Mann-Witney's one
wilcox.test(old, new, paired = F, alt = "greater", conf.int = T)

# Calculating W_obs
z<- c(old,new); rz<-rank(z); sum(rz[16:30])

# Hodges-Lehmann estimator
HodgesLehmann(old,new)
```