Homework 3

Chapter 6.6.

p. 95-96: ex. 1.

1) Let $X_1, \dots, X_n \sim$ Poisson(λ) and let $\hat{\lambda} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$. Find the bias, SE, MSE.

bias($\hat{\lambda}$) $= E_\lambda(\hat{\lambda}) - \lambda$, if bias $= 0$, then we say unbiased, otherwise biased.

$$\text{bias}(\hat{\lambda}) = E_\lambda(\hat{\lambda}) - \lambda = E_\lambda\left(\frac{1}{n} \cdot \sum_{i=1}^{n} X_i\right) - \lambda =$$

$$= E_\lambda\left(\frac{1}{n}\right) \cdot E\left(\sum_{i=1}^{n} X_i\right) - \lambda = \frac{1}{n} \cdot \sum_{i=1}^{n} E(X_i) - \lambda =$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n} \lambda - \lambda = \frac{\lambda \cdot \cancel{n}}{\cancel{n}} - \lambda = \lambda - \lambda = 0 -$$

- unbiased estimator

$$\widehat{SE}(\hat{\lambda}) = SE\left(\frac{1}{n} \cdot \sum_{i=1}^{n} X_i\right) = \frac{1}{n} SE\left(\sum_{i=1}^{n} X_i\right) =$$

$$= \frac{1}{n} \cdot \sqrt{\sum_{i=1}^{n} SE(X_i)^2} = \frac{1}{n} \cdot \sqrt{\sum_{i=1}^{n} Var(X_i)} =$$

$$= \frac{1}{n} \cdot \sqrt{\sum_{i=1}^{n} \hat{\lambda}} = \frac{1}{n} \cdot \sqrt{\lambda \cdot n} = \frac{\sqrt{n} \cdot \sqrt{\hat{\lambda}}}{\cancel{n}} =$$

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

MSE - Mean Square Error

$$MSE = bias^2(\hat{\lambda}) + Var_n(\hat{\lambda}) =$$
$$= bias^2(\hat{\lambda}) + [SE(\hat{\lambda})]^2 =$$
$$= 0^2 + \left[\sqrt{\frac{\hat{\lambda}}{n}}\right]^2 = 0 + \frac{\lambda}{n} = \frac{\lambda}{n}$$

Chapter 7. 4.

p. 104-105: ex. 2, 3, 9.

2. Let $X_1, \dots, X_n \sim Bernoulli(p)$
   let $Y_1, \dots, Y_m \sim Bernoulli(q)$

1) Find the plug-in estimator and SE_p

In Bernoulli distribution:

mean, $\mu = p$

variance, $Var = p \cdot (1-p)$

Let $\hat{p} = T(F) = \mu = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X_n}$

The standard error is $SE = \sqrt{Var(\overline{X_n})} =$

$$= \sqrt{Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)} = \sqrt{\frac{1}{n^2} \cdot (Var(X_1) + \dots + Var(X_n))} =$$

$$= \sqrt{\frac{n \cdot Var(X_i)}{n^2}} = \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}} \text{ or } \sqrt{\frac{\overline{X_n} \cdot (1-\overline{X_n})}{n}}$$

2) Find an approximate 90% conf. interval for p

Normal-based interval. For 90% conf. interval, $Z_{\frac{1-0.9}{2}} = Z_{\frac{0.1}{2}} = Z_{0.05} = 1.64$

A normal-based 90% conf. interval for p is:

$$\hat{p} \pm Z_{0.05} \cdot \hat{SE}(\hat{p}) = \hat{p} \pm 1.64 \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}$$

or

$$\overline{X}_n \pm 1.64 \cdot \sqrt{\frac{\overline{X}_n \cdot (1-\overline{X}_n)}{n}}$$

3) Find the plug-in estimator and $\hat{SE}$ for $(p-q)$

The plug-in estimator $(p-q) = \int x \, dF_1(x) - \int x \, dF_2(x) = \theta$

$$\hat{\theta} = \int x \, d\hat{F}_1(x) - \int x \, d\hat{F}_2(x) = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{m}\sum_{i=1}^{m} Y_i =$$

$$= \hat{p} - \hat{q} = \overline{X}_n - \overline{Y}_m$$

$$Var(\hat{\theta}) = Var(\hat{p} - \hat{q}) = Var(\hat{p}) + Var(\hat{q}) =$$

$$= \frac{\hat{p}(1-\hat{p})}{n} + \frac{\hat{q}(1-\hat{q})}{m}$$

$$SE(\hat{\theta}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{\hat{q}\cdot(1-\hat{q})}{m}}$$

4) Find an approximate 90% conf. interval for $(p-q)$.

$Z_{\frac{1-\alpha}{2}} = Z_{\frac{1-0.9}{2}} = Z_{\frac{0.1}{2}} = Z_{0.05} = 1.64$, an approximate interval is:

$$(\hat{p}-\hat{q}) \pm 1.64 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{\hat{q}(1-\hat{q})}{m}}$$

or

$$(\overline{X}_n - \overline{Y}_m) \pm 1.64 \cdot \sqrt{\frac{\overline{X}_n \cdot (1-\overline{X}_n)}{n} + \frac{\overline{Y}_m \cdot (1-\overline{Y}_m)}{m}}.$$

3. Simulation.

First, I defined the variable grid, which represents all the x-values from -3 ~~to~~ to +3. More length - more precise and less step will be between values. Next I define F_n -

cummulatative distribution function, which will it for all values between -3 to +3. The graph will be smooth, it is true CDF.

Then I take 100 observation of standard Normal distribution. For plotting - I sort them. Then I create an estimate of CDF $\hat{F}_n(x)$:

$$\hat{F}_n(x) = \frac{\sum_{i=1}^{n} I\{X_i \leq x\}}{n}$$

Then I construct confidence bound

$$\alpha = 1 - 0.95 = 0.05$$

$$\varepsilon_n = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}$$

$$L(x) = \max\{\hat{F}_n(x) - \varepsilon_n, 0\}$$

$$U(x) = \min(\hat{F}_n(x) + \varepsilon_n, 1\}$$

No matter what the true F, then

$$P(L(x) \leq F(x) \leq U(x) \text{ for all } x) \geq 1-\alpha$$

Zhetessov1.R

Source on Save

Run

```
#----------------------------------------------------------------------------------------
#Generate 100 observations from a N(0,1) distribution.
#----------------------------------------------------------------------------------------
par(mfrow=c(2,2))
grid <- seq(-3,3, length=1000)
Fn <- pnorm(grid)
plot(grid, Fn, type="l", xlab="X value",ylab="CDF", sub="Theoretical CDF")


n <- 100
x <- rnorm(n, mean = 0, sd = 1)
x <- sort(x)

Fn_hat <- (1:n)/n #ecdf(x)
plot(x, Fn_hat, type="s", xlab="X Value",ylab="CDF", sub="Empirical CDF", xlim=c(-3,3))

plot(grid, Fn, type="l",xlab="X Value",ylab="CDF")
lines(x,Fn_hat,type="s", lty=3,col=4,lwd=3)
```

```r
#------------------------------------------------------------------------------------------
#Compute a 95 percent confidence band for the CDF F
#------------------------------------------------------------------------------------------
alph <- 1 - 0.95
epsilon <- sqrt(1 / (2 * n) * log(2 / alph)) # I did by the formulas on the page 99 in Wasserman's Book

L <- pmax(Fn_hat - epsilon, 0)
U <- pmin(Fn_hat + epsilon, 1)

plot(grid, Fn, type="l", xlab="x", ylab="cdf", xlim=c(-3,3))
lines(x,L,lty=3,col=2,type="s")
lines(x,U,lty=3,col=2,type="s")

```

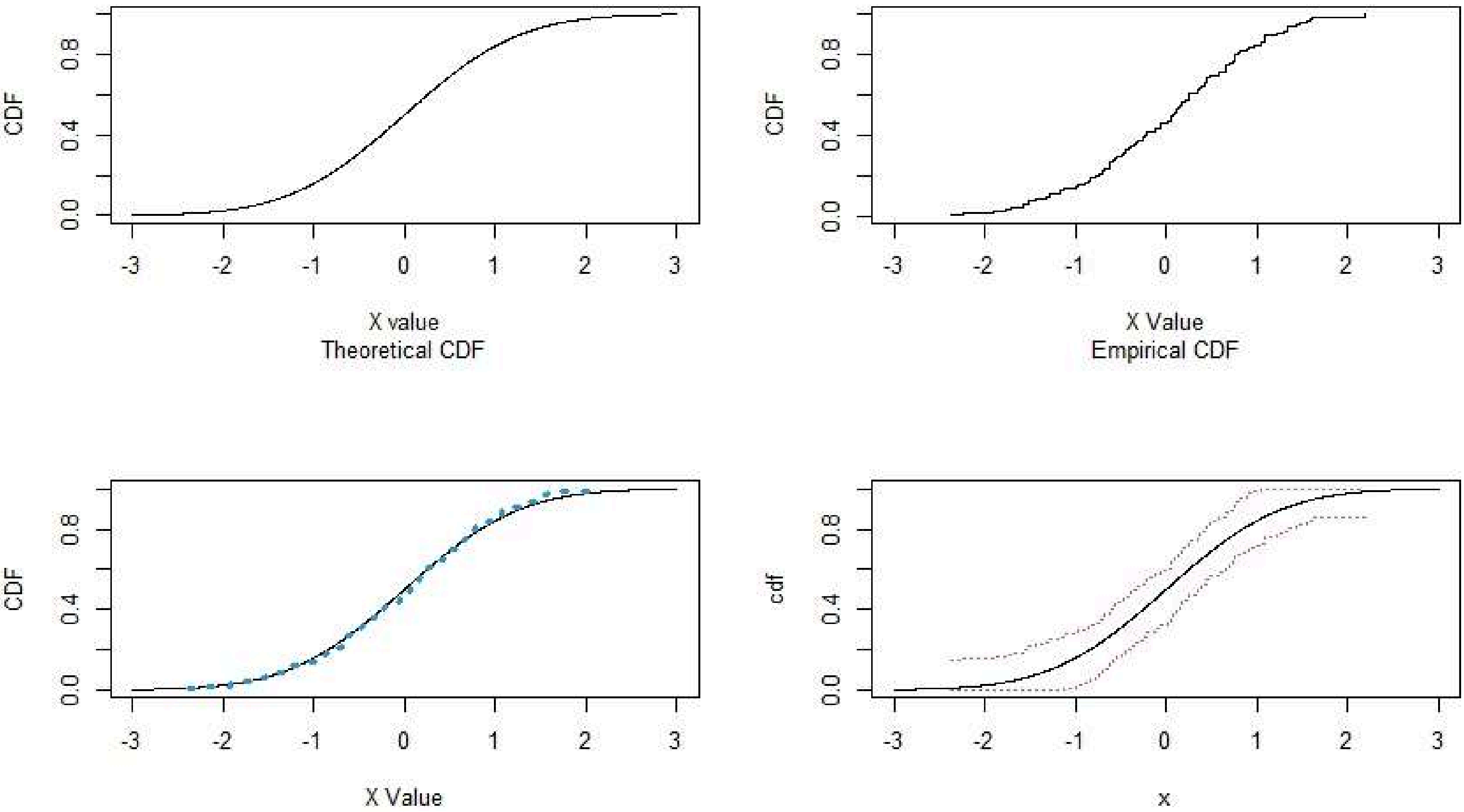
```r
#-------------------------------------------------------------------------------------------
#Repeat this 1000 times and see how often
#the confidence band contains the true distribution function.
#-------------------------------------------------------------------------------------------
ans <- c()

for(i in 1:1000){
  grid <- seq(-3,3, length=1000)
  Fn <- pnorm(grid)

  n <- 100
  x <- rnorm(n, mean = 0, sd = 1)
  x <- sort(x)

  Fn_hat <- (1:n)/n #ecdf(x)


  alph <- 1 - 0.95
  epsilon <- sqrt(1 / (2 * n) * log(2 / alph))

  L <- pmax(Fn_hat - epsilon, 0)
  U <- pmin(Fn_hat + epsilon, 1)

  fraction = c()
  for (i in 1:100)
  {
    logic <- U[i]>=pnorm(x[i]) && L[i]<=pnorm(x[i])
    fraction <- append(fraction, logic)
  }

  ans <- append(ans,all(fraction))
}

mean(ans) # About 0.955 which is >=0.95 (95% confidence interval)

```

33:90 (Untitled)

Console Terminal Jobs

R 4.1.2 · ~/

```
+   ans <- append(ans,all(fraction))
+ }
>
> mean(ans)
[1] 0.955
>
```


CDF
0.0
0.4
0.8
-3
-2
-1
0
1
2
3
X value
Theoretical CDF
CDF
0.0
0.4
0.8
-3
-2
-1
0
1
2
3
X Value
Empirical CDF
CDF
0.0
0.4
0.8
-3
-2
-1
0
1
2
3
X Value
cdf
0.0
0.4
0.8
-3
-2
-1
0
1
2
3
x

```r
#-----------------------------------------------------------------------------------------------
#Repeat using data from a Cauchy distribution.
#-----------------------------------------------------------------------------------------------
ans <- c()

for(i in 1:1000){
  grid <- seq(-3,3, length=1000)
  Fn <- pnorm(grid)

  n <- 100

  x <- rcauchy(n) #I only changed the function for random sample
  x <- sort(x)

  Fn_hat <- (1:n)/n #ecdf(x)


  alph <- 1 - 0.95
  epsilon <- sqrt(1 / (2 * n) * log(2 / alph))

  L <- pmax(Fn_hat - epsilon, 0)
  U <- pmin(Fn_hat + epsilon, 1)

  fraction = c()
  for (i in 1:100)
  {
    logic <- U[i]>=pnorm(x[i]) && L[i]<=pnorm(x[i])
    fraction <- append(fraction, logic)
  }

  ans <- append(ans,all(fraction))
}

mean(ans) # About 0.19 - 0.20 which is (< 0.95) lower than 95% confidence interval

#Zhetessov Nur M.

```

In Cauchy distribution, calculating the mean will provide no useful information, because mean is underfined, as the Var(X). Therefore, I can't calculate confidence interval properly for Cauchy distribution. In experiment it shows only 80% match.

9.

$n = 100$ people

$m = 100$ people

$k_1 = 90$ recovered people

$k_2 = 85$ recovered people

p_1 - standard treatment

p_2 = new treatment

Estimate $\theta = p_1 - p_2$, $SE(\theta)$

80% conf. interval.

95% conf. interval.

$$p_1 = \frac{\# \text{ recovered by standard treat.}}{\# \text{ total}}$$

$$p_2 = \frac{\# \text{ recovered by new treat.}}{\# \text{ total}}$$

$p_1 = \frac{90}{100} = 0.9$ - probability of recovery by standard. treat

$p_2 = \frac{85}{100} = 0.85$ - prob. recover. by new. treat.

An estimate is

$$\theta = p_1 - p_2 = 0.9 - 0.85 = 0.05$$

Standard error is

I showed it in exercise 2.

$$[SE(\theta)]^2 = Var(\theta) = Var(p_1 - p_2) =$$

$$= Var(p_1) + Var(p_2) = \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}$$

$$SE(\theta) = \sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}$$

$$SE(\theta) = \sqrt{\frac{0.9 \cdot 0.1}{100} + \frac{0.85 \cdot 0.15}{100}} = 0.0466$$

Z-score for 80% Conf. interval is

$Z_{\frac{1-0.8}{2}} = Z_{\frac{0.2}{2}} = Z_{0.1} = 1.282$

Upper bound:

$(p_1 - p_2) + Z_{0.1} * SE(\theta)$

$0.05 + 1.282 * 0.0466$

0.110

Lower bound:

$(p_1 - p_2) - Z_{0.1} * SE(\theta)$

$0.05 - 1.282 * 0.0466$

-0.010

Conf. interval 80% is $(-0.010; 0.110)$

Z-score for 95% is $Z_{0.025} = 1.96$

Lower bound: $(p_1 - p_2) - 1.96 * SE(\theta)$

$0.05 - 1.96 * 0.0466$

-0.041

Upper bound: $0.05 + 1.96 \cdot 0.0466$

0.1413 Interval is $(-0.041; 0.1413)$

Chapter 8.6

p. 116-118: ex. 6, 7a

6. In this our statistic is a function

$g(\mu) = e^{\mu}$

I have create a random normal sample ~ $N(5,1)$ of 100 observations.

Then I found $\hat{\theta}$ by $g(\bar{X}_{\hat{\theta}})$

Using bootstrap method, I sample randomly our original sample and found its statistic by $g(x)$. All $g(\theta^*)$ were contained in a vector.

To find conf. interval, I need SE. We find SE from vector with all θ^*, taking their mean and $\sqrt{Var}$. With it I can construct conf. interval:

$$\hat{\theta} \pm z * \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\theta_i^* - \frac{1}{n}\sum_{j=1}^{n}\theta_j^*\right)^2}$$

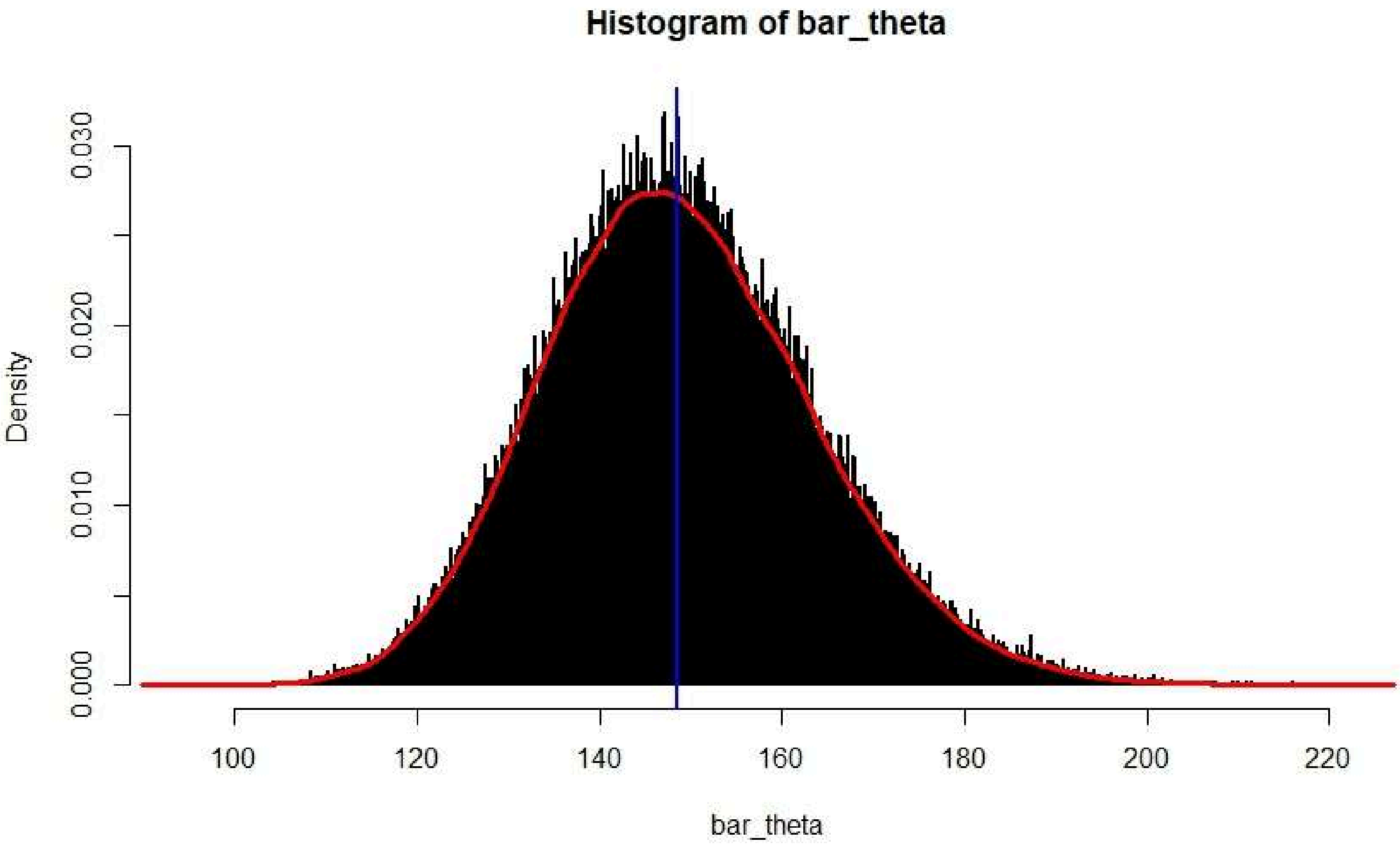
Then I plot all θ^* versus its frequency and

6Ex_Zhetessov.R

```r
#-------------------------------------------------------------------------------------------
#Create a data set mu=5 consisting of n=100 observations.
#-------------------------------------------------------------------------------------------

n <- 100
mu <- 5
x <- rnorm(n, mean=mu, sd=1)

#-------------------------------------------------------------------------------------------
# Use the bootstrap method
#-------------------------------------------------------------------------------------------
est_hat_theta <- exp(mean(x))

bar_theta <- c()
for(i in 1:100000){
  rand_sampling <- sample(x, size = n, replace = TRUE)
  star_theta = exp(mean(rand_sampling))

  bar_theta <- append(bar_theta, star_theta)
}


#-------------------------------------------------------------------------------------------
#(a) Get the SE and 95 percent confidence interval for Theta
#-------------------------------------------------------------------------------------------
SE = sqrt(var(bar_theta))

alpha = 1 - (95/100)
z_score = abs(qnorm(alpha/2))

Upp_bound = est_hat_theta + z_score * SE
Low_bound = est_hat_theta - z_score * SE

c(Low_bound,Upp_bound)

#-------------------------------------------------------------------------------------------
#(b) Plot a histogram of the bootstrap replications. Compare this to the true sampling distribution
#-------------------------------------------------------------------------------------------
hist(bar_theta, breaks = 1000, col = "darkmagenta",freq = FALSE)
lines(x = density(x = bar_theta), col = "red", lwd = 3)

true_theta = exp(mu)
abline(v=true_theta,col="blue",lwd=2)

#Zhetessov Nur M.
```

Console

```
> Low_bound = est_hat_theta - z_score * SE
>
> c(Low_bound,Upp_bound)
[1] 119.2606 176.9549
>
> #-------------------------------------------------------------------------------------------
> #(b) Plot a histogram of the bootstrap replications.
```


Histogram of bar_theta
Density
0.000
0.010
0.020
0.030
100
120
140
160
180
200
220
bar_theta

draw a line - blue line - the true θ which by statistic is $e^{\mu} = e^5$. I see that ≥ 0.95 of times out confidence interval contain the true parameter.

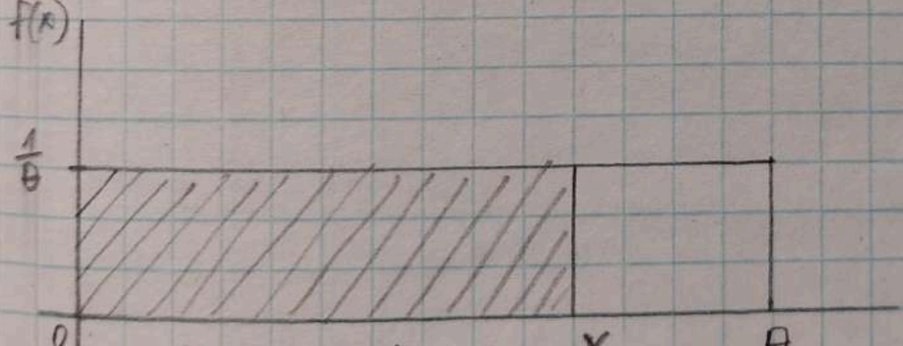
7. a) $X_1, \dots, X_n \sim Uniform(0, \theta)$.

$\hat{\theta} = X_{max} = \max\{X_1, \dots, X_n\}$.

Generate data set of 50 with $\theta = 1$

In Uniform distribution the probability density function (PDF) is $f(x) = \frac{1}{b-a}$ $(a \leq x \leq b)$

$f(X) = \frac{1}{\theta - 0} = \frac{1}{\theta}$



$P(k < X)$

Assume all X_i are independent. The maximum $X_i < \theta$, if and only if all $X_i < \theta$

$P(\hat{\theta} < X) = P(\max\{X_1 \dots X_n\} < X) = P(X_1 < X, X_2 < X, \dots, X_n < X) =$

$= P(X_1 < x) \cdot P(X_2 < x) \cdot \ldots \cdot P(X_n < x) =$

$= (x-0) \cdot \left(\frac{1}{\theta - 0}\right) \cdot (x-0) \cdot \left(\frac{1}{\theta - 0}\right) \ldots (x-0) \cdot \left(\frac{1}{\theta - 0}\right) =$

$= \left(x \cdot \frac{1}{\theta}\right)^n = \left(\frac{x}{\theta}\right)^n$ since $\theta = 1$, then $(x)^n$

where the distribution is:

$$F(x) = \begin{cases} 0, & x < 0 \\ \left(\frac{x}{\theta}\right)^n, & 0 < x < \theta \quad \text{for} \quad 0 \leq \theta \leq 1 \\ 1, & x > \theta \end{cases}$$

The confidence interval in computer simulation showed good coverage 0.95 - lower bound to 1,02 upper bound, which contains the true $\theta = 1$

7a_Ex.R × 6Ex_Zhetessov.R ×

Source on Save

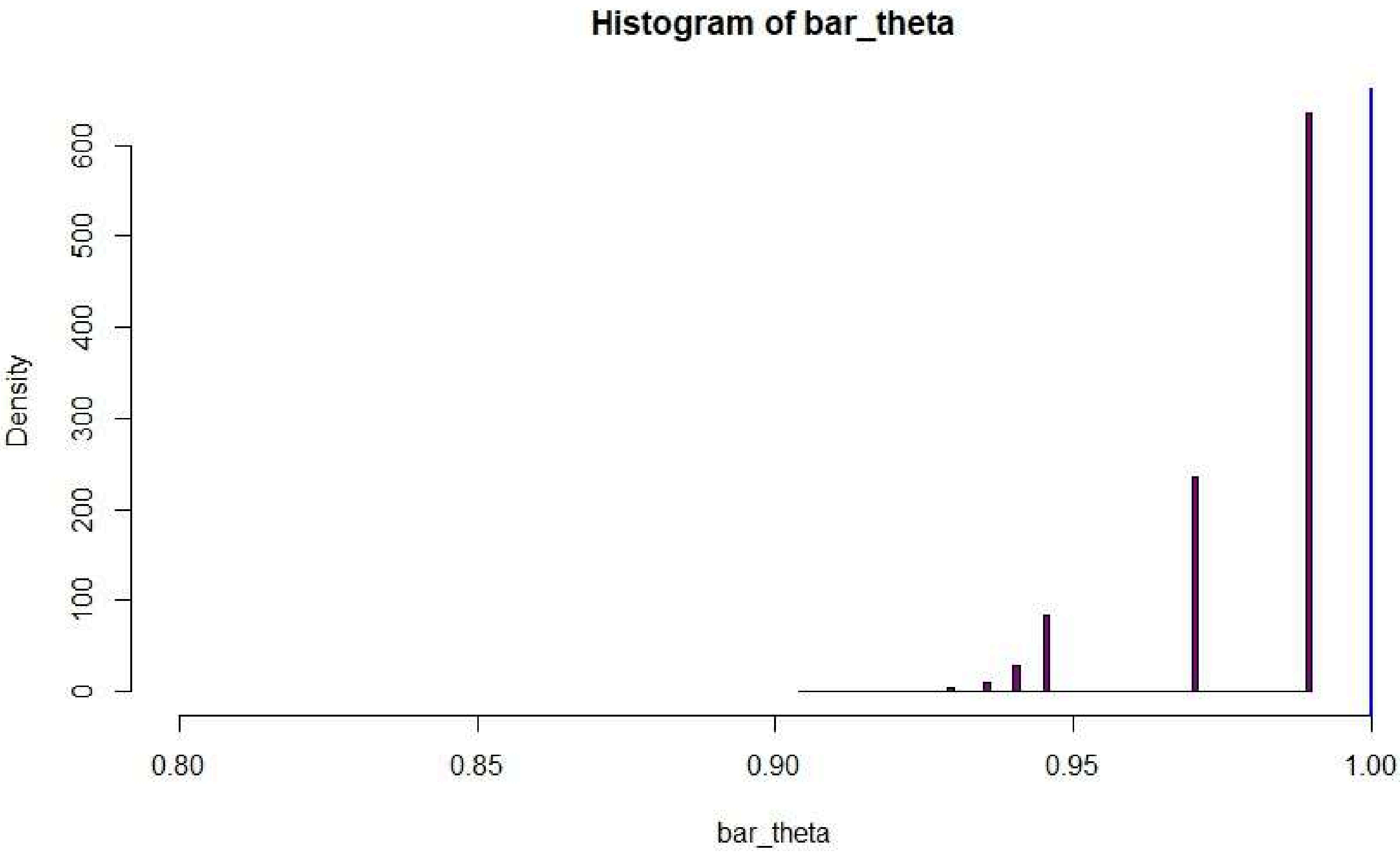
```r
n<-50
x <- runif(n, 0, 1)

hat_theta = max(x)

bar_theta <- c()
for(i in 1:100000){
  rand_sampling <- sample(x, size = n, replace = TRUE)
  star_theta = max(rand_sampling)

  bar_theta <- append(bar_theta, star_theta)
}

SE = sqrt(var(bar_theta))

alpha = 1 - 0.95
z = abs(qnorm(alpha/2))

Lower_bound = hat_theta - z * SE
Upper_bound = hat_theta + z * SE

Lower_bound
Upper_bound

hist(bar_theta, breaks = 100, col = "darkmagenta",freq = FALSE, xlim=c(0.8, 1))
#lines(x = density(x = bar_theta), col = "red", lwd = 3)

true_theta = max(1)
abline(v=true_theta,col="blue",lwd=2)

#Zhetessov Nur M.

```

31:18 (Top Level)

Console Terminal × Jobs ×

R 4.1.2 · C:/Users/Nur/Desktop/дз/Стат/HW3/

```
> z = abs(qnorm(alpha/2))
>
> Lower_bound = hat_theta - z * SE
> Upper_bound = hat_theta + z * SE
>
> Lower_bound
[1] 0.958237
> Upper_bound
[1] 1.021634
>
> hist(bar_theta, breaks = 100, col = "darkmagenta",freq = FALSE, xlim=c(0.8, 1))
> #lines(x = density(x = bar_theta), col = "red", lwd = 3)
>
> true_theta = max(1)
> abline(v=true_theta,col="blue",lwd=2)
>
> #Zhetessov Nur M.
>
```


Histogram of bar_theta
Density
0
100
200
300
400
500
600
0.80
0.85
0.90
0.95
1.00
bar_theta