GR5234 - HW2

Mathieu Sauterey - UNI: mjs2364

September 19, 2018

Problem 1

Let N = 6 and n = 3. For purposes of studying sampling distributions, assume that all population values are known.

$y_1 = 98$ $y_2 = 102$ $y_3 = 133$ $y_4 = 154$ $y_5 = 175$ $y_6 = 190$

We are interested in $\bar{y}_U$, the population mean. Two sampling plans are proposed.

Plan 1. Eight possible samples may be chosen with equal probability 1/8

Plan 2. Three possible samples may be chosen.

```
pop = c(98,102,133,154,175,190)


# Plan 1

sample1 = matrix(NA, ncol = 3, nrow = 8)

sample1[1,] = pop[c(1,3,5)]

sample1[2,] = pop[c(1,3,6)]

sample1[3,] = pop[c(1,4,5)]

sample1[4,] = pop[c(1,4,6)]

sample1[5,] = pop[c(2,3,5)]

sample1[6,] = pop[c(2,3,6)]

sample1[7,] = pop[c(2,4,5)]

sample1[8,] = pop[c(2,4,6)]


Prob1 = 1/8


# Plan 2

sample2 = matrix(NA, ncol = 3, nrow = 3)

sample2[1,] = pop[c(1,3,5)]

sample2[2,] = pop[c(2,3,6)]
```

```
sample2[3,] = pop[c(1,4,6)]


Prob2 = c(1/4,1/2,1/4)
```

(a) Let $\bar{y}$ be the mean of the sample values. For each sampling plan, find:

(i) E[y_bar]

(ii) V[y_bar]

(iii) Bias(y_bar)

(iv) MSE(y_bar)

```
# Sampling plan 1

mean1 = rowMeans(sample1)

E_Mean1 = Prob1*sum(mean1)
print(E_Mean1)
```

```
## [1] 142
```

```
Var_Mean1 = sum(Prob1*(mean1-E_Mean1)^2)
print(Var_Mean1)
```

```
## [1] 18.94444
```

```
Bias1 = E_Mean1 - mean(pop)
print(Bias1)
```

```
## [1] 0
```

```
MSE1 = Bias1^2 + Var_Mean1
print(MSE1)
```

```
## [1] 18.94444
```

For the sampling plan 1, we have:

$\mathrm{E}[\bar{y}] = 142$

$\mathrm{V}[\bar{y}] = 18.9444$

$\mathrm{Bias}(\bar{y}) = 0$

$\mathrm{MSE}(\bar{y}) = 18.9444$

```
# Sampling plan 2

mean2 = rowMeans(sample2)

E_Mean2 = sum(Prob2*mean2)
print(E_Mean2)
```

```
## [1] 141.5
```

```
Var_Mean2 = sum(Prob2*(mean2-E_Mean2)^2)
print(Var_Mean2)
```

```
## [1] 18.02778
```

```
Bias2 = E_Mean2 - mean(pop)
print(Bias2)
```

```
## [1] -0.5
```

```
MSE2 = Bias2^2 + Var_Mean2
print(MSE2)
```

```
## [1] 18.27778
```

For the sampling plan 2, we have:

$\mathrm{E}[\bar{y}] = 141.5$

$\mathrm{V}[\bar{y}] = 18.02778$

$\mathrm{Bias}(\bar{y})$ = -0.5

$\mathrm{MSE}(\bar{y}) = 18.2777$

(b) Which sampling plan do you think is better? Why?

Although sample 1 is unbiased, I think that sampling plan 2 is better because it has lower MSE.

Problem 2

Consider an artificial situation in which we know the value of y_i for each of the N = 8 units in the whole population. The index set for the population is:

$U = \{1,2,3,4,5,6,7,8\}$

and the values of y_i are:

$y = \{1, 2, 4, 4, 7, 7, 7, 8\}$

Consider the following sampling plan.

(a) Find the probability of selection Pi_i for each unit i.

```
pop = c(1, 2, 4, 4, 7, 7, 7, 8)

sample = matrix(NA, ncol = 4, nrow = 5)

sample[1,] = c(1,3,5,6)

sample[2,] = c(2,3,7,8)

sample[3,] = c(1,4,6,7)

sample[4,] = c(2,4,7,8)
```

```
sample[5,] = c(4,5,6,8)

Prob = c(1/8, 1/4, 1/8, 3/8, 1/8)

Pi = c()
for (i in 1:length(pop)){
  Pi[i] = sum(rowSums(sample == i)*Prob)
}

print(Pi)
```

```
## [1] 0.250 0.625 0.375 0.625 0.250 0.375 0.750 0.750
```

Based on the above we get:

$\pi_1 = 0.25$

$\pi_2 = 0.625$

$\pi_3 = 0.375$

$\pi_4 = 0.625$

$\pi_5 = 0.25$

$\pi_6 = 0.375$

$\pi_7 = 0.75$

$\pi_8 = 0.75$

(b) What is the sampling distribution of t_hat = 8y_bar?

```
sample = matrix(NA, ncol = 4, nrow = 5)

sample[1,] = pop[c(1,3,5,6)]

sample[2,] = pop[c(2,3,7,8)]

sample[3,] = pop[c(1,4,6,7)]

sample[4,] = pop[c(2,4,7,8)]

sample[5,] = pop[c(4,5,6,8)]

mean = rowMeans(sample)

E_Mean = sum(Prob*mean)

t_hat = 8*E_Mean

print(t_hat)
```

```
## [1] 42.25
```

```
Var_t_hat = sum(Prob*(8*mean-t_hat)^2)
print(Var_t_hat)
```

```
## [1] 16.4375
```

We find the following sampling distribution for $\hat{t}$

$\mathrm{E}[\hat{t}] = 42.25$

$\mathrm{V}[\hat{t}] = 16.4375$

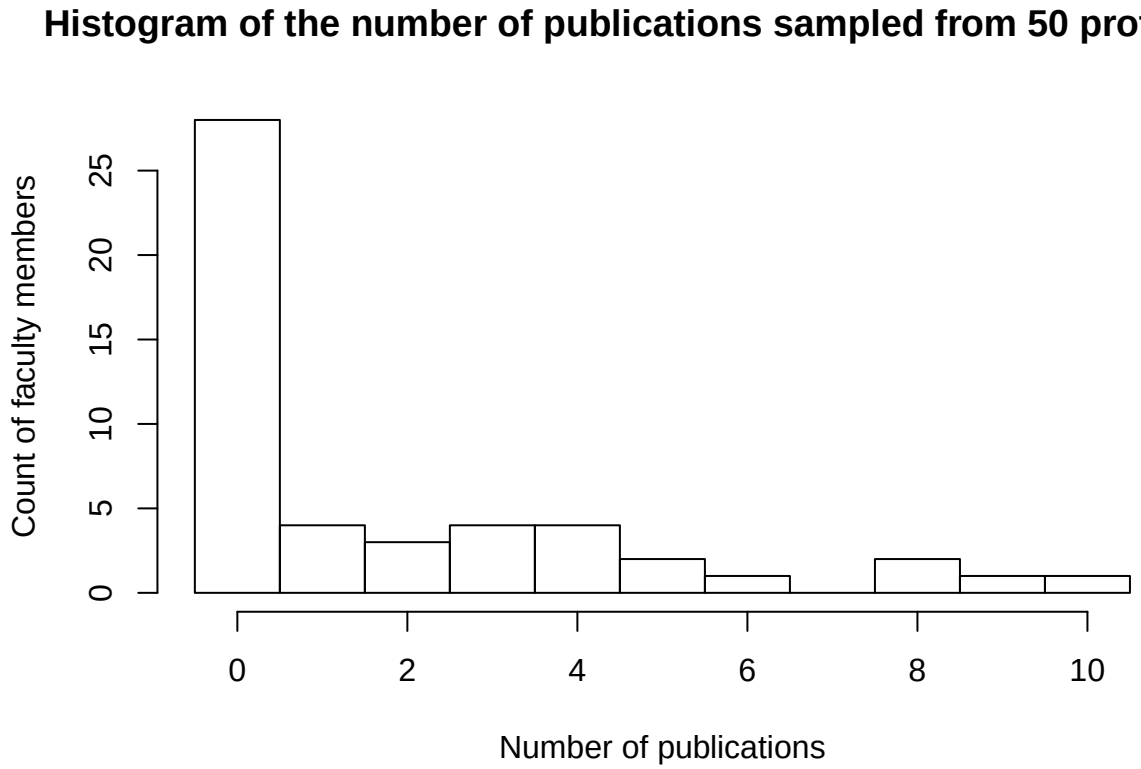
Problem 3

A university has 807 faculty members. For each faculty member, the number of refereed publications was recorded. A frequency table for number of refereed publications is given below for a simple random sample of 50 faculty members.

(a) Plot the data using a histogram. Describe the shape of the data.

```
publi = rep(0:10, times = c(28,4,3,4,4,2,1,0,2,1,1))

hist(publi, breaks = seq(-0.5,10.5,1), xlab = "Number of publications", ylab = "Count of faculty members
     main = "Histogram of the number of publications sampled from 50 professors ")
```



The data is not normal at all and it is very skewed to the right.

(b) Estimate the mean number of publications per faculty member, and give the standard error for your estimate.

```
n = 50
N = 807
mean_publi = mean(publi)
print(mean_publi)
```

```
## [1] 1.78
```

```
SE_mean_publi = sqrt(1-n/N)*sd(publi)/sqrt(n)
print(SE_mean_publi)
```

```
## [1] 0.3674151
```

The estimated mean number of publications per faculty member is 1.78

The standard error for my estimate is 0.367

(c) Do you think that y_bar from part (b) will be approximately normally distributed? Why or why not?

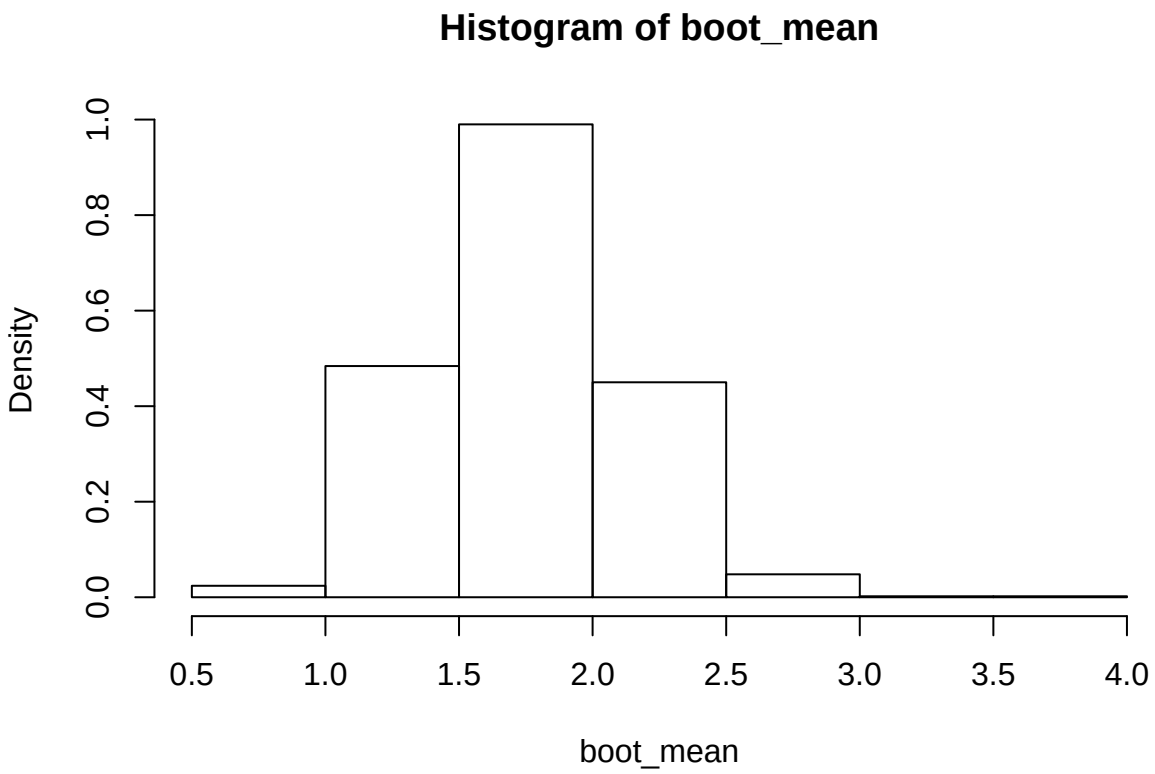
Although the population distribution of the number of publications seems very skewed to the right, our sample size n = 50 seems sufficiently large to ensure that y_bar is normally distributed.

Let's use bootstrap to verify this assumption:

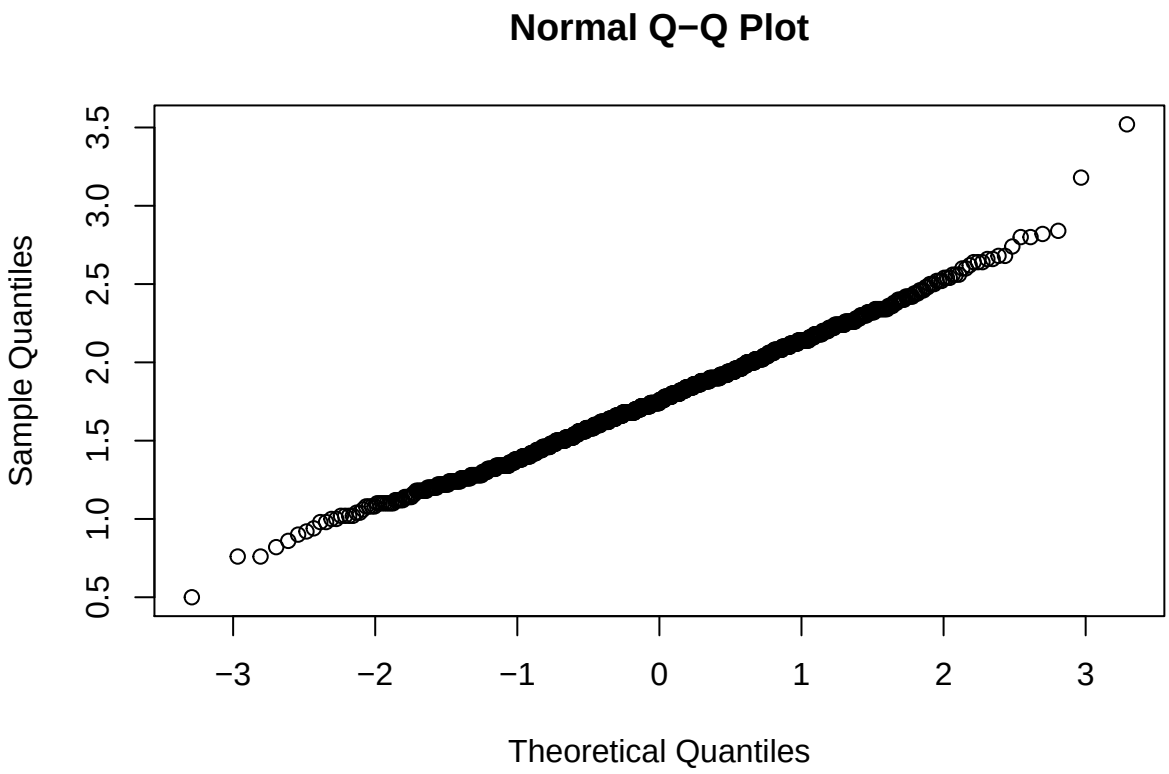
```
#Use bootstrap to check normality
boot_mean = c()

for (i in 1:1000){
boot_mean[i] = mean(sample(publi, 50, replace = T))
}

hist(boot_mean, breaks = 10, freq=F)
```




```
qqnorm(boot_mean)
```



Using bootstrap, we see that the distribution of y_bar seems approximately normal. This is confirmed by the QQplot that shows an approximately straight line.

Estimate the proportion of faculty members with no publications, and give a 95% confidence interval.

```
alpha = 0.05

prop_publi = (publi == 0)

mean_prop_publi = mean(prop_publi)
print(mean_prop_publi)
```

```
## [1] 0.56
```

```
SE_prop_publi = sqrt((1-n/N) * mean_prop_publi * (1-mean_prop_publi)/(n-1))

CI <- mean_prop_publi + c(-1,1) * qnorm(1-alpha/2) * SE_prop_publi
print(CI)
```

```
## [1] 0.4253887 0.6946113
```

The estimated proportion of faculty members with no publications is 0.56.

A 95% confidence interval for this estimate is [0.425 , 0.695]

Problem 4

The manager of a mail-in lottery game took a random sample of 1000 entries from the last few contests, and found that 175 of those came from the South.

(a) Find a 95% confidence interval for the percentage of all entries that come from the South.

```
n = 1000

alpha_south = 0.05

prop_south = 175/n

# In absence of better information we assume that population is large: fpc = 1
SE_prop_south = sqrt(prop_south*(1-prop_south)/(n-1))

CI_south <- prop_south + c(-1,1) * qnorm(1-alpha_south/2) * SE_prop_south
print(CI_south)
```

```
## [1] 0.151438 0.198562
```

[15.1% , 19.9%] is a 95% confidence interval for the percentage of all entries that come from the South.

(b) According to the Statistical Abstract of the United States, 30.9% of the U.S. population lives in states that the lottery manager considered to be in the South. Is there evidence from your interval in part (a) that the percentage of entries from the South differs from the percentage of persons living in the South?

30.9% is not included in the confidence interval [15.1% , 19.9%] thus there is evidence that the percentage of entries from the South differs from the percentage of persons living in the South.

Problem 5

Suppose we wish to take a simple random sample of the 580 children served by a family medical practice, to estimate the proportion who are overdue for a vaccination. What sample size would be necessary to estimate the proportion with 95% confidence and margin of error 0.08?

```
N = 580

alpha_child = 0.05

z_alpha = qnorm(1-alpha_child/2)

e = 0.08

# In absence of better information we use Var = 0.25 (maximum possible)
Var_child = 0.25
SE_child = sqrt(Var_child)
```

```r
n_0 = (z_alpha*SE_child/e)^2

n = n_0 / (1+n_0/N)
ceiling(n)
```

```
## [1] 120
```

Therefore a sample size of 120 children would be necessary to estimate the proportion with 95% confidence and margin of error 0.08.