DATA 252 / DATA 551: Homework 7

1. In class, we used bootstrapping to study the *median* incubation period. Recall that the median is the 50*th* percentile. We might be interested in estimating other percentiles, like the 90*th* percentile, which tells us that 90% of patients will develop symptoms before this time. Use the same dataset (ncov_simple) and answer the following questions.

1.1 What is the 90*th* percentile of the incubation period among patients in the dataset? Recall, in R, you can use the quantile function.

```
# Question 1.1
> # 90th Percentile
> covid19_statictics(Pt = 0.9)
[1] "Percentile"
  90%
6.785
```

1.2 Obtain *B* = 5000 bootstrap samples; record the 90*th* percentile of each bootstrap sample. What is the SD of the 90*th* percentile based on your bootstrap samples?

```
> bootstrap_covid19_statictics(bootstrap_size = 5000,Pt =0.9 )
[1] "SD:"
[1] 1.077051
```

1.3 Compared to the median (we have estimated its SD in class), does the 90*th* percentile have more or less variation?

```
# Q 1.3
> bootstrap_covid19_statictics(bootstrap_size = 5000, statistic =
'median', CI = 95)
[1] "Confidence Interval"
 2.5% 97.5%
 2.92  4.50
[1] "SD:"
[1] 0.3920383
```

90th percentile has high SD.

1.4 Obtain a 95% confidence interval for the 90*th* percentile of the incubation period.

```
bootstrap_covid19_statictics(bootstrap_size = 5000, Pt = 0.9, CI = 95)
[1] "Confidence Interval"
 2.5% 97.5%
5.500 9.104
```

2. Suppose we want to use bootstrapping to study the *maximum* incubation period.

2.1 What is the maximum incubation period among patients in the dataset?

```
# Q 2.1
> max(covid_data)
[1] 11
```

2.2 Obtain *B* = 5000 bootstrap samples and construct a 95% confidence interval for the maximum incubation period.

```
bootstrap_covid19_statictics(bootstrap_size = 5000, maxx=T, CI = 95)
[1] "Confidence Interval"
 2.5% 97.5%
  8.5  11.0
```

2.3 Briefly explain why bootstrapping does not work well for estimating the maximum incubation period.

Ans: It is because, bootstrap sample can’t exceed the 11 days maximum. Which most probability will not be a true a maximum value.

CODE:

```
# Load Data
covid_data = read.csv("./nCoV_simple.csv")$days
print("Data Loaded")


# Sattistics on data without bootstrap
covid19_statictics <- function(statistic = -1, CI = -1, Pt = -1){
  if(CI != -1){
    # Calcaulate and print specified Statistic's CI
    print("Confidence Interval")
    calculate_CI(CI,covid_data)
  }
  if(Pt != -1){
    # Calcaulate and print specified percentile
    print("Percentile")
    print(quantile(covid_data, probs =Pt))  }
  if(statistic != -1){
  # mean or median
  print(statistic)
  print(get(statistic)(covid_data))

  }
  # SD
  print('SD:')
  print(sd(covid_data))
  # plot histogram
```

```
  hist(covid_data)
}

# Do bootstap and get statistics
bootstrap_covid19_statictics <- function(statistic = -1,bootstrap_size, CI = -1, Pt = -1,
maxx = F) {

  bootstap_result = rep(NA,bootstrap_size)
  for(i in 1:bootstrap_size){
  if(statistic != -1)
  {
    bootstap_result[i] = get(statistic)(sample(covid_data, replace = T))
  }
  if(Pt != -1){
      # Calcaulate specified percentile
      bootstap_result[i] = quantile(sample(covid_data, replace = T), Pt)
  }
  if(maxx)
    {
      # Calcaulate max
      bootstap_result[i] = max(sample(covid_data, replace = T))
    }
  }

  if(CI != -1){
  # Calcaulate and print specified Statistic's CI
  print("Confidence Interval")
  calculate_CI(CI,bootstap_result)
  }

  # print Standared Deviation
  print('SD:')
  print(sd(bootstap_result))
  # plot histogram
  hist(bootstap_result)
}



# Calcaulate and print specified Statistic's CI
calculate_CI <- function(CI,data){
  alpha = (100 - CI) / 100
  alpha_half = alpha / 2
  probb = 1 - alpha_half
  print(quantile(data, probs =c(alpha_half, probb) ))
}
# ---------------------------------------------------------------
```

```
# Question 1.1
# 9th Percentile
covid19_statictics(Pt = 0.9)

# Q 1.2
bootstrap_covid19_statictics(bootstrap_size = 5000,Pt =0.9 )

# Q 1.3
bootstrap_covid19_statictics(bootstrap_size = 5000, statistic = 'median', CI = 95)

# Q 1.4
bootstrap_covid19_statictics(bootstrap_size = 5000, Pt = 0.9, CI = 95)


# Q 2.1
max(covid_data)

# Q 2.2
bootstrap_covid19_statictics(bootstrap_size = 5000, maxx=T, CI = 95)

bootstrap_size = 1000
bootstap_result = rep(NA,bootstrap_size)
for(i in 1:bootstrap_size){
  bootstap_result[i] = quantile(sample(covid_data, replace = T), 0.9)
}
```

The following is modified from the last question on Exam 1 (some numbers have been changed). Re-do this question. The grade (out of 10 points) you receive here will be added to your exam grade. If your current exam grade is 95 and above (good job!), you can skip this part, and will receive 5 points extra credit automatically. This part will not be counted in the homework grade. **You must work on this question independently, without communicating with any other student.**

Suppose the incubation period, T , of a certain disease can be modeled by an exponential distribution with a mean of 4.5 days.

(1) What is the chance that a given person develops symptoms within 5 days of contracting the disease (i.e., $P(T \leq 5)$)?

```
(2)     > prob_of_showing_symptoms = pexp(5,rate = 1/4.5) # 0.670807
(3)     > prob_of_showing_symptoms
(4)     [1] 0.670807
```

(2) In real life, we usually cannot observe the actual incubation periods T for each person, but can only observe whether or not a person develops any symptoms in, say, 5 days. Simulate 100 patients who have contracted the disease, with a value of 1 representing that the patient has symptoms within 5 days, and a value of 0 representing that the patient has no symptom within 5 days (i.e., you'll be generating a vector of length 100 with entries 0 or 1). Copy-paste this vector here.

(3) Based on your answer in (2), how many patients, out of the 100 patients, have developed symptoms within 5 days?

Method 1 using sample:

hundered_patents_y_n = sample(c(0,1), size = 100, replace = TRUE, prob = c(prob_of_not_showing_symptoms,prob_of_showing_symptoms))

```
> hundered_patents_y_n
  [1] 1 1 0 0 1 1 0 1 1 0 1 0 1 1 1 1 1 0 0 0 1 1 0 1 0 0 1 1 0 1 0 1 0 1 0 1
1 1 1 0 0
 [42] 1 1 1 0 0 0 0 1 0 0 1 1 1 0 0 1 0 1 1 0 0 1 1 0 1 0 1 1 1 1 1 0 1 1 1 1
0 1 1 0 1
 [83] 1 1 1 1 1 1 0 1 0 1 1 1 1 0 1 1 1 0

(4)     sum(hundered_patents_y_n)
(5)     [1] 71
```

Method 2 using binomial:

```
vector = rbinom(100,size = 1, prob = prob_of_showing_symptoms)
> vector
  [1] 1 0 0 1 1 1 1 0 1 0 0 0 1 1 1 1 0 1 1 1 0 0 0 0 1 0 0 1 1 1 1 1 1 1 1 1
0 0 1 1 1 1 1 1 1 0 1
 [48] 1 1 1 1 1 1 1 1 1 0 1 1 1 1 0 1 1 1 1 1 1 1 0 0 1 0 1 1 1 0 1 1 0 1 1 0
0 1 1 1 0 0 1 0 0 0 1
 [95] 1 0 1 1 1 1

> sum(vector)
[1] 69
```

(6) Run a simulation to estimate the probability that, among 100 people who have contracted the disease, 80 or more people will develop symptoms within 5 days.

Method 1 using sample:

prob_of_showing_symptoms = pexp(5,rate = 1/4.5) # 0.670807
prob_of_not_showing_symptoms = 1 - prob_of_showing_symptoms # 0.329193
nsim = 10000000
result = rep(NA,nsim)
for(i in 1:nsim){
  hundered_patents_y_n = sample(c(0,1), size = 100, replace = TRUE, prob = c(prob_of_not_showing_symptoms,prob_of_showing_symptoms))
  result[i] = sum(hundered_patents_y_n)
}

mean(result>=80)
0.003106

Method 2 using binomial:

```
nsim = 10000000
result = rep(NA,nsim)
for(i in 1:nsim){
  result[i] = rbinom(1,size = 100, prob = prob_of_showing_symptoms)
}

mean(result>=80)
  0.0030372
```

Method 3:

```
> nsim = 10000000
> result = rbinom(nsim,size = 100, prob = prob_of_showing_symptoms)
> mean(result>=80)
[1] 0.0030585
```

(7) Suppose the number of new cases contracting the disease today follows a Poisson distribution with a mean of 100 people. Modify your simulation in (4) to estimate the probability that, among people who have contracted the disease today, 80 or more people will show symptoms within 5 days.

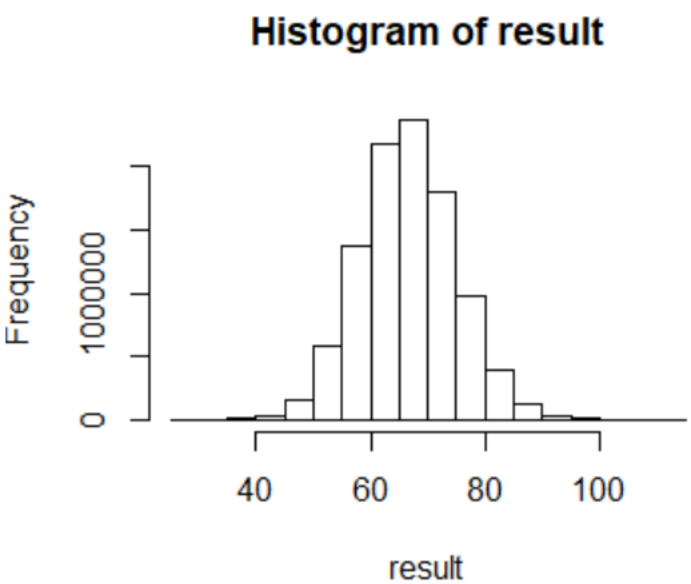
(8) Generate a histogram of the distribution of the number of people that will show symptoms within 5 days, under the setup in (5). Take a screenshot of the histogram to include here.

Method 1 using sample:

```
prob_of_showing_symptoms = pexp(5,rate = 1/4.5) # 0.670807
prob_of_not_showing_symptoms = 1 - prob_of_showing_symptoms # 0.329193
nsim = 10000000
result = rep(NA,nsim)
for(i in 1:nsim){
  n_newpeople = rpois(1,lambda = 100)
  patents_y_n = sample(c(0,1), size = n_newpeople, replace = TRUE, prob =
c(prob_of_not_showing_symptoms,prob_of_showing_symptoms))
  result[i] = sum(patents_y_n)
}

mean(result>80)
0.0539275

hist(result)
```



Method 2 using binomial:

```
nsim = 10000000
result = rep(NA,nsim)
prob_of_showing_symptoms = pexp(5,rate = 1/4.5) # 0.670807
for(i in 1:nsim){
  n_newpeople = rpois(1,lambda = 100)
  result[i] = rbinom(1,size = n_newpeople, prob = prob_of_showing_symptoms)
}

mean(result>=80)
0.0676643
```

hist(result)