Statistical Methods for Data Science

Mini Project 5

Members:

SHIVA RANGA CHAWALA sxc167630

KRISHNA SINDHU KOTA kxk171030

Contribution:

Shiva Ranga Chawala – Equally contributed

Krishna Sindhu Kota – Equally contributed

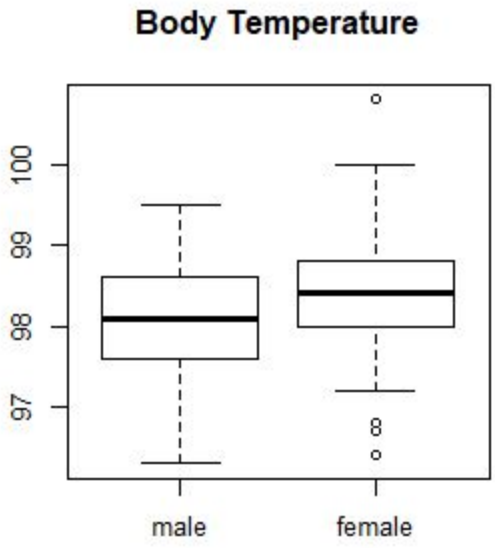
1. **Consider the data stored in bodytemp-heartrate.csv on eLearning, containing measurements of body temperature and heart rate for 65 male (gender = 1) and 65 female (gender = 2) subjects.**
    a. **Do males and females differ in mean body temperature? Answer this question by performing an appropriate analysis of the data, including an exploratory analysis.**

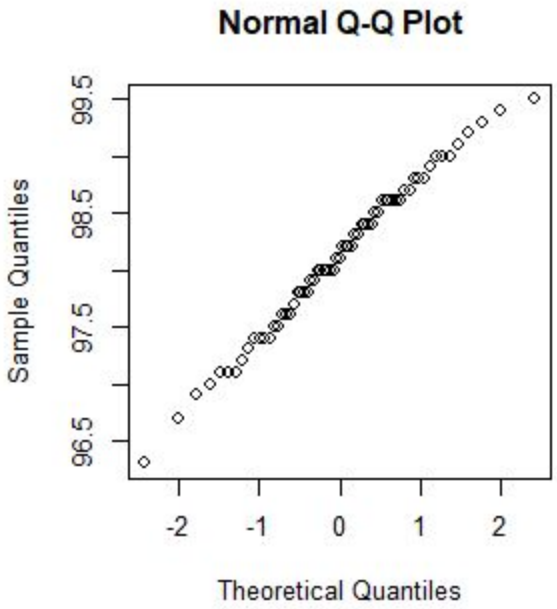
Code:

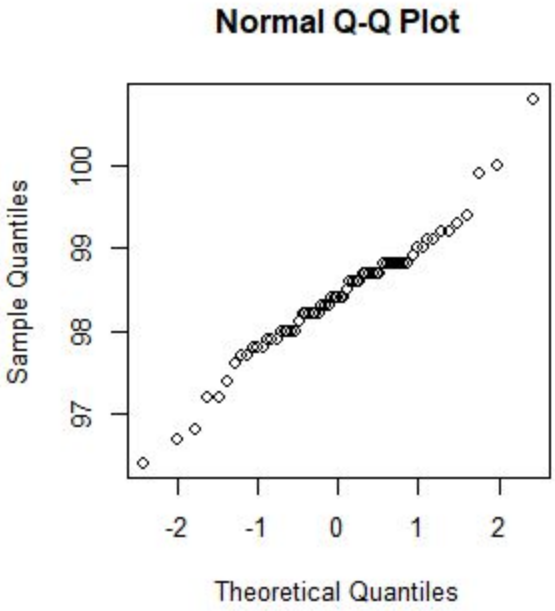
```
#Reading bodytemp-heartrate csv file
data = read.csv(file = "bodytemp-heartrate.csv", sep = ",", header = T)

# Subsetting the data based on the gender
data.male = subset(data, gender == "1")
data.female = subset(data, gender == "2")

# Boxplot between body_temperature and gender
boxplot(body_temperature ~ gender, data = data, names = c("male", "female"), main = "Body Temperature")

#Storing the body_temperatures of male in temp.male and female in temp.female
temp.male = data.male$body_temperature
temp.female = data.female$body_temperature

#QQ plots of body_temperatures of male and female
qqnorm(temp.male)
qqnorm(temp.female)

#Performing t-test for body_temperatures of male and female with two-sided alternative
t.test(temp.male, temp.female, alternative = "two.sided", var.equal = F)
```

Results:

```
Welch Two Sample t-test

data:  temp.male and temp.female
t = -2.2854, df = 127.51, p-value = 0.02394
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.53964856 -0.03881298
sample estimates:
mean of x mean of y
 98.10462  98.39385
```







Interpretations:

From the above box-plot we can see that the variability for the females is more when compared to males. The values of Q1,Q3 and median are higher for females than that of males and IQR value of males is greater than that of females. The female data contains outliers. We can infer that females have a larger mean than that of males. From the QQ plots we can make a normality assumption and we have no clue about the variances. So we can perform Satterthwaite t-test. As

we are checking the difference in mean body temperature between males and females we can consider null hypothesis H0: → $\mu_{male} = \mu_{female}$ and alternative hypothesis as H1: → $\mu_{male} \neq \mu_{female}$. From the results of the t-test, we can see that the p-value (0.02394) is very less(<0.05), so we can reject the null hypothesis which means that we are accepting the alternative hypothesis H1: → $\mu_{male} \neq \mu_{female}$. We can also observe that the 95% confidence interval is [-0.53964856, -0.03881298]. As 0 doesn't fall in the interval, $\mu_{male} < \mu_{female}$,we can say that there is difference in mean body temperatures between males and females.

(b) Do males and females differ in mean heart rate? Answer this question by performing an appropriate analysis of the data, including an exploratory analysis.

Code:

#Box plot between heart_rate and gender

```
boxplot(heart_rate ~ gender, data = data, names= c("male", "female"), main = "Heart Rate")
```

#Storing heart_rates of males in heart_rate.m and heart_rates of females in heart_rate.f

```
heart_rate.m <- data.male$heart_rate
heart_rate.f <- data.female$heart_rate
```

#Drawing QQ plots of heart_rates of males and females

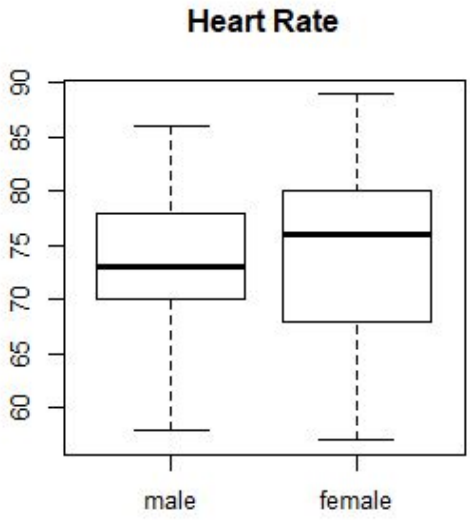
```
qqnorm(heart_rate.m)
qqnorm(heart_rate.f)
```

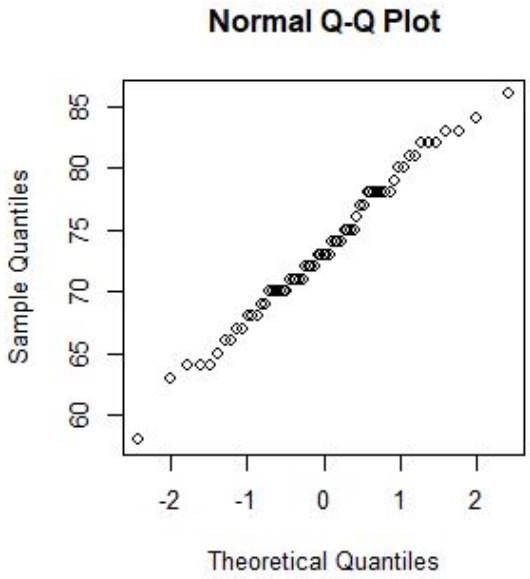
#Performing Satterthwaite t-test of heart rate between males and females for two sided #alternative

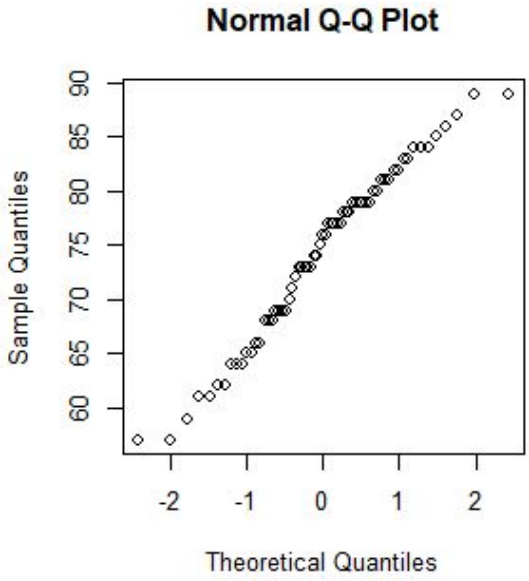
```
t.test(heart_rate.m, heart_rate.f, alternative = "two.sided", var.equal = F)
```

Results:

```
Welch Two Sample t-test
data:  heart_rate.m and heart_rate.f
t = -0.63191, df = 116.7, p-value = 0.5287
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.243732  1.674501
sample estimates:
mean of x mean of y
 73.36923  74.15385
```







Interpretations:

From the above box-plot, we can see that the variability for the females is more when compared to males. Median is higher for females than that of males and IQR value of females is greater than that of males.The value for Q1 is lesser for females than males and Q3 is greater for females. From the QQ plots we can make a normality assumption and we have no clue about the variances. So we can perform Satterthwaite t-test. As we are checking the difference in mean heart rate between males and females we can consider null hypothesis H0: $\rightarrow \mu_{male} = \mu_{female}$ and alternative hypothesis as H1: $\rightarrow \mu_{male} \neq \mu_{female}$. From the results of the t-test, we can see that the p-value (0.5287) is greater than 0.05 , so we accept the null hypothesis which means that we are

rejecting the alternative hypothesis H1: $\rightarrow \mu_{male} \neq \mu_{female}$. We can also observe that the 95% confidence interval is [-3.243732 1.674501]. As 0 falls in the interval, we can say that there is no difference in mean heart rates between males and females.

(c) Is there a linear relationship between body temperature and heart rate? Does this relationship depend on gender? Answer these questions by performing an appropriate analysis of the data, including an exploratory analysis.

Code:

#Scatter plots between body temperature and heart rate for males and females
plot(body_temperature ~ heart_rate, data = data.male, col = "blue")
points(body_temperature ~ heart_rate, data = data.female, col = "red")

#Correlation between body temperatures and heart rates for males and females
cor(data.male$body_temperature, data.male$heart_rate)
cor(data.female$body_temperature, data.female$heart_rate)

#Linear model for body temperature and heart rate for males and females
model.m <- lm(body_temperature ~ heart_rate, data = data.male)
model.f <- lm(body_temperature ~ heart_rate, data = data.female)

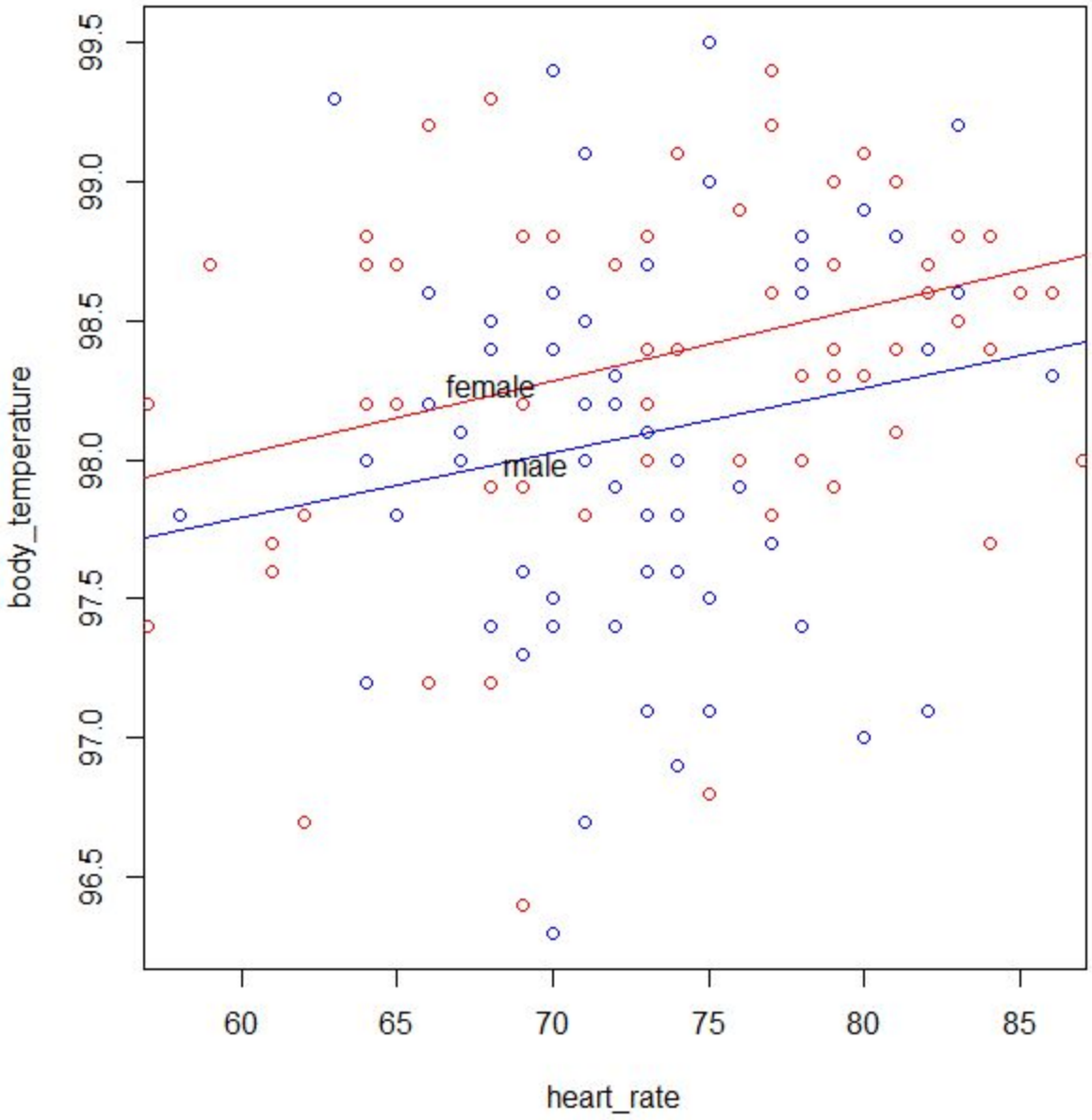
#Adding model to the same scatter plot
abline(model.m, col = "blue")
abline(model.f, col = "red")
text(locator(), labels = c("male", "female"))

Results:

> cor(data.male$body_temperature, data.male$heart_rate)
[1] 0.1955894
> cor(data.female$body_temperature, data.female$heart_rate)
[1] 0.2869312

Interpretations:

Yes, there is linear relationship between body temperature and heart rate for both males and females which can be observed from the scatter plot. The strength of linear relationship can be measured by correlation. The correlation value between body temperature and heart rate for males is 0.1955894 and that of females is 0.2869312. We can observe that the correlation value for females is greater than males and the correlation values are very less(<0.3) which means that they have a weak positive linear relationship. From the scatter plot, we can see that the points are more scattered. The relationship depends on body temperature and heart rate classified by gender. From the plot, we can observe that the slope of females is slightly higher than that of males.



2. The goal of this exercise to see how large n should be for the large-sample and the bootstrap percentile method confidence intervals for the mean of an exponential population to be accurate. To be specific, let X1, . . . , Xn represent a random sample from an exponential (λ) distribution. Note that this distribution is skewed and its mean is μ = 1/λ. We can construct two confidence intervals for μ — one the large-sample z-interval (interval 1) and the other a bootstrap percentile method interval (interval 2). We would like to investigate their accuracy, i.e., how close their estimated coverage probabilities are to the assumed nominal level of confidence, for various combinations of (n, λ). This investigation

will focus on $1 - \alpha = 0.95$, $\lambda = 0.01, 0.1, 1, 10$ and $n = 5, 10, 30, 100$. Thus, we have a total of $4 * 4 = 16$ combinations of (n, λ) to investigate.

(a) For a given setting, compute Monte Carlo estimates of coverage probabilities of the two intervals by simulating appropriate data, using them to construct the two confidence intervals, and repeating the process 5000 times.
(b) Repeat (a) for the remaining combinations of (n;). Present an appropriate summary of the results.

(a),(b)

```
library(boot) #importing boot lib
lambda = c(0.01,0.1,1,10) #lambda values
n = c(5,10,30,100) #n values
alpha = 0.05 #alpha value for 95% CI
monte = 5000 #no. Of runs

mean_x = function(x, indices) #function we used for boot call
{
  result = mean(x[indices])
  return(result)
}

for(i in n)
{
  for(j in lambda)
  {
    count1=0
    count2=0

    for (k in 1:monte)
    {
      x = rexp(i, j) #generating x values with exponential distribution
      mean.x = 1/j #population mean
      var.x = var(x)
      standard.error = sqrt(var.x/i)
    CI.large = mean(x) + c(-1,1) * qnorm(1-alpha/2) * standard.error #large sample z-interval CI
      if(CI.large[1]<mean.x && CI.large[2]>mean.x)
      {
        count1=count1+1 #increasing count if mean lies in the CI
      }

      CI.boot = boot(x, mean_x, R=999, sim = "ordinary", stype = "i")
      boot.percentile = sort(CI.boot$t)[c(25,975)] #Bootstrap percentile CI
      if(boot.percentile[1]<mean.x && boot.percentile[2]>mean.x)
      {
```

```
    count2=count2+1 #increasing count if mean lies in the CI
   }
  }
  coverage.prob.large = count1/monte #coverage prob for large sample z-interal
  coverage.prob.boot = count2/monte #coverage prob for bootstrap percentile
  print(paste("Value of n:", i))
  print(paste("Value of lambda", j))
  print(paste("Coverage probability of large sample z-interval", coverage.prob.large))
  print(paste("Coverage probability of percentile bootstrap", coverage.prob.boot))
 }
}
```

Results:

#	n	lambda	Large sample coverage probability	Percentile bootstrap coverage probability
1	5	0.01	0.8144	0.7852
2	5	0.1	0.809	0.7856
3	5	1	0.8144	0.7908
4	5	10	0.808	0.7834
5	10	0.01	0.8678	0.8664
6	10	0.1	0.8734	0.8652
7	10	1	0.8672	0.8636
8	10	10	0.871	0.8654
9	30	0.01	0.9258	0.923
10	30	0.1	0.9188	0.9216
11	30	1	0.9216	0.9204
12	30	10	0.9176	0.9196
13	100	0.01	0.929	0.9322
14	100	0.1	0.9348	0.937
15	100	1	0.9322	0.9338
16	100	10	0.939	0.9382

(c) Interpret all the results. Be sure to answer the following questions: In case of the large-sample interval, how large n is needed for the interval to be accurate? Likewise, in case of the bootstrap interval, how large n is needed for the interval to be accurate? Do these answers depend on lambda? Can we say that one method is more accurate than the other? Which interval would you recommend? Provide justification for all your conclusions.

Interpretation:

- We can see that as the value of 'n' increases the accuracy for both the estimators increases.
- If we consider 0.90 as accuracy level then we can conclude that "n" value should be 30 for both large sample z-interval and percentile bootstrap interval to be accurate.
- For larger values of n we can see that bootstrap percentile method gives higher accuracy irrespective of lambda.
- For smaller values of n we can observe that large sample z-interval method gives higher accuracy irrespective of lambda.
- As n increases and at a particular lambda value, the coverage probability of both percentile bootstrap and large sample z interval increases.
- When n is small, large sample z-interval is more accurate when compared to bootstrap percentile method.
- When n is large, bootstrap percentile CI is more accurate when compared to large sample z-interval.
- We cannot say that one method is accurate than another, because when n is small the large sample z-interval is more accurate and when n is large bootstrap percentile interval is more accurate.
- When n is fixed and lambda changes we can see that there is no consistent change in values of coverage probabilities, from which we can infer that coverage probabilities doesn't depend on lambda values. Below is the table for the same:

#	n	lambda	Large sample coverage probability	Percentile bootstrap coverage probability
1	5	0.01	0.8144	0.7852
2	5	0.1	0.809	0.7856
3	5	1	0.8144	0.7908
4	5	10	0.808	0.7834

(d) Do your conclusions in (c) depend on the specific values of lambda that were fixed in advance? Explain.

As the accuracy is not depending up on the lambda values, we can say that the specific values of lambda that were fixed in advance doesn't affect conclusions. We can observe from the below table that even if the lambda values are different i.e., not fixed, we are observing the same trend which is when n value is small, the large sample z interval coverage probabilities are more when compared to the percentile bootstrap coverage probability and as n increases, the percentile bootstrap coverage probabilities are exceeding the large sample z interval coverage probabilities. So we can deduct that, the conclusions we derived is not affected by the fixed values of lambda.

#	n	lambda	Large sample coverage probability	Percentile bootstrap coverage probability
1	5	0.05	0.8102	0.7856
2	5	30	0.8076	0.781
3	10	0.05	0.8664	0.8608
4	10	30	0.8684	0.8628
5	30	0.05	0.915	0.9156
6	30	30	0.913	0.9148
7	100	0.05	0.9464	0.9474
8	100	30	0.9446	0.9448