**Q1. a)** This is the case of independent samples with unknown population variances, where one sample is collection of 65 male body temperatures and other sample is collection of 65 female body temperatures. First, we will do the boxplot analysis of these two samples:

# Read the input data file

> temp\_hrate <- read.csv(file = "bodytemp-heartrate.csv", header=TRUE, sep=",")

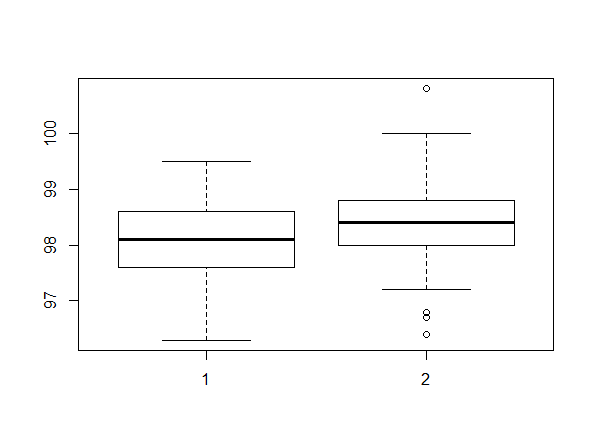
# male body temperatures

> male\_temp <- temp\_hrate$body\_temperature[temp\_hrate$gender == 1]

# female body temperatures

> female\_temp <- temp\_hrate$body\_temperature[temp\_hrate$gender == 2]

> boxplot(male\_temp, female\_temp)



In the figure above, (1) represents boxplot for male temperatures and (2) for female temperatures. It can be seen from the boxplots above that the two samples have different variances, different median and different IQR.

> var(male\_temp)

[1] 0.4882596

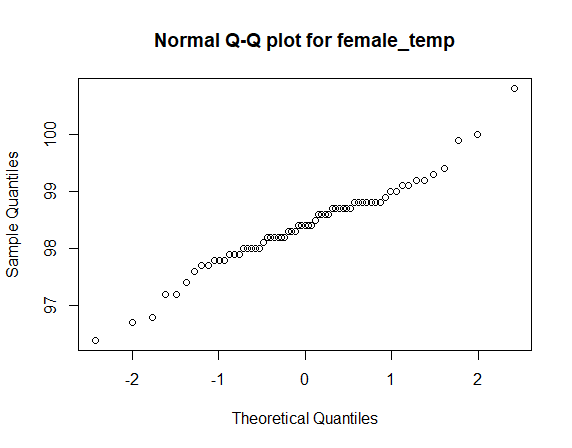
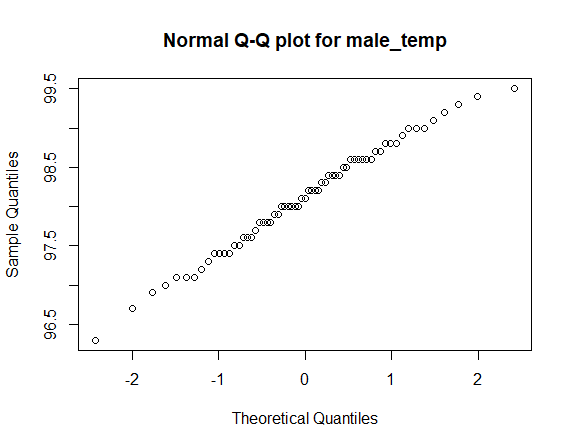
> var(female\_temp)

[1] 0.552774

Now, let’s check the normality assumption for the two samples by using qqnorm:

> qqnorm(male\_temp, main = "Normal Q-Q plot for male\_temp")

> qqnorm(female\_temp, main = "Normal Q-Q plot for female\_temp")

We can see from the above qqnorm plots that the two samples can be assumed to be coming from normal populations. Hence, we can check if male and female mean body temperatures differ by using t-test for difference in mean of two populations (Satterwhaite’s approximation since 2 sample variances are not equal).

> t.test(male\_temp, female\_temp, alternative = "two.sided", conf.level = 0.95,var.equal = FALSE)

Welch Two Sample t-test

data: male\_temp and female\_temp

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

Here, we had chosen the null hypothesis, H0 = “Difference in mean of male and female body temperatures is 0” and alternative hypothesis, H1 = “Difference in mean of male and female body temperatures is not 0”. Also, we are considering the 95% confidence interval and hence alpha = 0.05. We can see from the above t-test that the the p-value = 0.02 < alpha, hence the null hypothesis can be rejected. So, we can conclude that male and females differ in mean body temperatures.

Also, from t-test, our 95% confidence interval is [-0.5396, -0.0388] which consists of all negative values (i.e. 0 doesn’t lie in this interval). Hence, we can say that mean of female body temperatures is generally more than mean of male body temperatures.

**b)** This is again the case of independent samples with unknown population means, where one sample is collection of 65 male heartrates and other sample is collection of 65 female heartrates. First, we will do the boxplot analysis of these two samples:

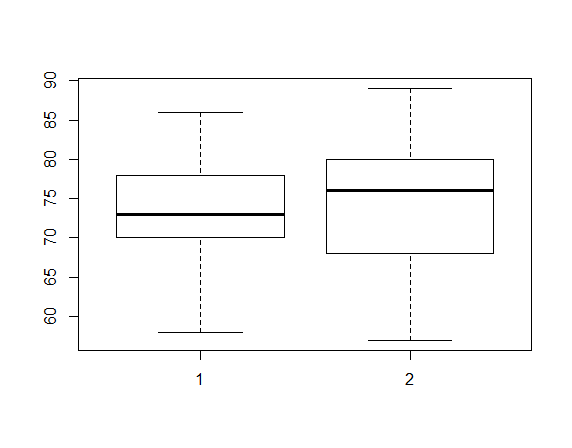
# Heart rates of 65 males

> male\_hrate <- temp\_hrate$heart\_rate[temp\_hrate$gender == 1]

# Heart rates of 65 females

> female\_hrate <- temp\_hrate$heart\_rate[temp\_hrate$gender == 2]

> boxplot(male\_hrate, female\_hrate)



In the figure above, (1) represents boxplot for male heart rates and (2) for female heart rates. It can be seen from the boxplots above that the two samples have different variances, different median and different IQR.

> var(male\_hrate)

[1] 34.51779

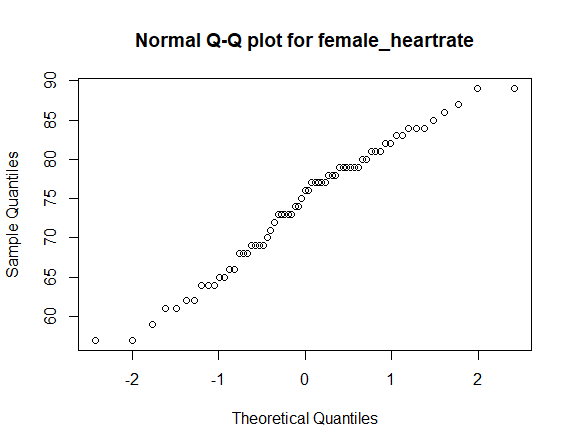
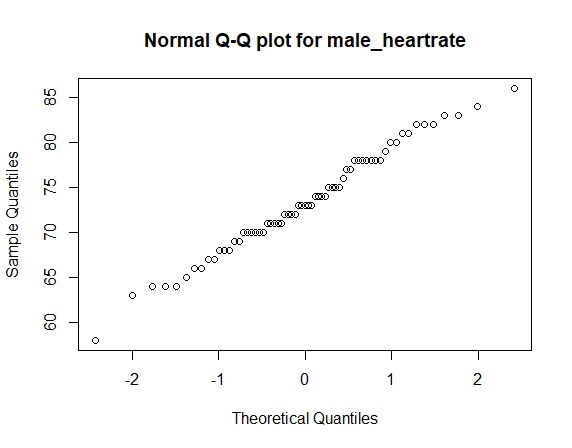
> var(female\_hrate)

[1] 65.69471

Now, let’s check the normality assumption for the two samples by using qqnorm:

> qqnorm(male\_hrate, main = "Normal Q-Q plot for male\_heartrate")

> qqnorm(female\_hrate, main = "Normal Q-Q plot for female\_heartrate")



Seeing the qqnorm-plots above, we will assume the two samples coming from normal populations. Hence, we can check if male and female mean heart rates differ by using t-test for difference in mean of two populations (Satterwhaite’s approximation since 2 sample variances are not equal).

> t.test(male\_hrate, female\_hrate, alternative = "two.sided", conf.level = 0.95,var.equal = FALSE)

Welch Two Sample t-test

data: male\_hrate and female\_hrate

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

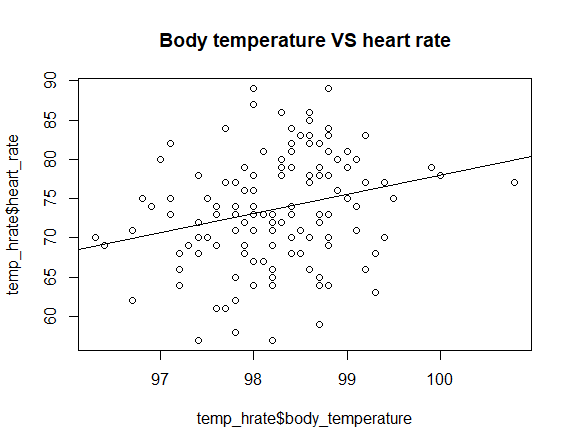
Here, we had chosen the null hypothesis, H0 = “Difference in mean of male and female heart rates is 0” and alternative hypothesis, H1 = “Difference in mean of male and female heart rates is not 0”. Also, we are considering the 95% confidence interval and hence alpha = 0.05. We can see from the above t-test that the p-value = 0.5287 > alpha, hence the null hypothesis can be accepted. So, we can conclude that males and females don’t differ in mean heart rates.

Also, from t-test, our 95% confidence interval is [-3.2437, 1.6745] which shows that the difference of mean heart rate of males and females can be zero because 0 lies in this interval. Hence, we can say that mean female heart rate generally equals mean male heart rate.

**c)** To check if there is any linear relationship between body temperature and heart rate, we can plot the data and check:

> plot(temp\_hrate$body\_temperature, temp\_hrate$heart\_rate, main="Body temperature VS heart rate")

> abline(lm(temp\_hrate$heart\_rate ~ temp\_hrate$body\_temperature))



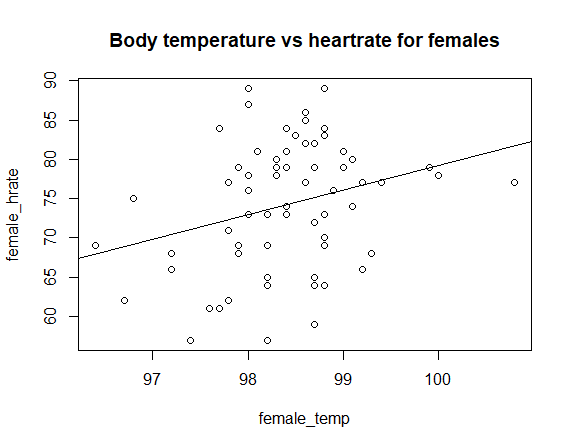
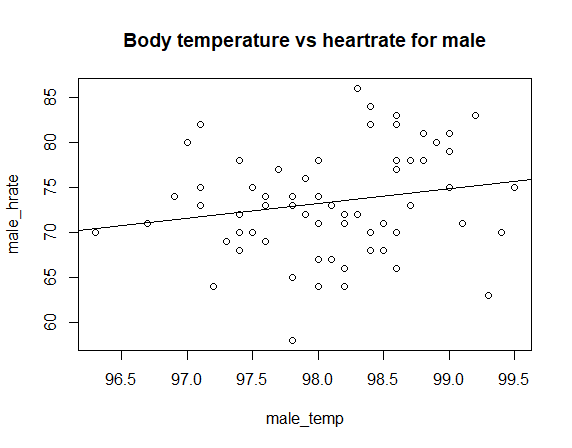
We can see from the plot that there is positive linear relation between heart rate and body temperature. Now we can check if this relationship depends on gender by plotting heart rate and temperatures for males and females separately:

> plot(male\_temp, male\_hrate, main="Body temperature vs heartrate for male")

> abline(lm(male\_hrate ~ male\_temp))

> plot(female\_temp, female\_hrate, main="Body temperature vs heartrate for females")

> abline(lm(female\_hrate ~ female\_temp))



We can see from the above plots that there is positive linear relationship between heart rate and body temperature whether its male or female.

**Q2. a)** We know that we can create a random sample (say x) of size ‘n’ from an exponentially distributed population as below:

x <- rexp(n, lambda) ;

where n = number of observations, lambda = rate or frequency of events in exponential distribution.

We need to find two confidence intervals for a sample mean in this question. First confidence interval is large-sample z-interval, which can be calculated as below in R:

Conf\_int1 <- mean(x) + c(-1, 1) \* z-score(alpha/2)\*sd(x)/sqrt(n) = mean(x) + c(-1, 1) \* qnorm (1-alpha/2)\*sd(x)/sqrt(n)

For our question, alpha = 0.05 in all cases.

Second confidence interval is calculated by parametric bootstrap method as below:

nboot <- 1000

#Distribution of sample mean using parametric bootstrap

mean.pboot.dist <- replicate(nboot, mean.star(x))

#95 % CI using parametric percentile bootstrap

conf\_int2 <- sort(mean.pboot.dist)[c(25, 975)]

where we define mean.star function to calculate resamples and return mean as below:

mean.star <- function(x) {

n <- length(x)

xbar <- mean(x)

# For resample, lambda = 1/xbar

xstar <- rexp(n, 1/xbar)

meanstar <- mean(xstar)

return(meanstar)

}

The above two methods for C.I. calculation can be combined in one function (defined below) to return both type of C.I’s for each sample.

conf\_intervals <- function(n, lambda) {

x <- rexp(n, lambda)

## 1st Confidence interval - large z-sample C.I. (alpha = 0.05)

conf\_int1 <- mean(x) + c(-1, 1)\*qnorm(1 - (0.05/2))\*sd(x)/sqrt(n)

##2nd Confidence interval - parametric percentile bootstrap C.I. (alpha = 0.05)

nboot <- 1000

mean.pboot.dist <- replicate(nboot, mean.star(x))

conf\_int2 <- sort(mean.pboot.dist)[c(25, 975)]

# Return both C.I.’s as a single matrix(1X4), where 1st two values are for 1st C.I. and last two values are for 2nd C.I.

return(c(conf\_int1, conf\_int2))

}

Then the process can be repeated 5000 times for Monte-Carlo simulations and we can check the proportion of accuracy of each method for C.I.

Now, for example, for the first case, we set n = 5, lambda = 0.01

> n <- 5

> lambda <- 0.01

> mean\_exp <- 1/lambda

> n\_sim <- 5000

# Finding the both C.I.’s 5000 times by monte-carlo simulations

> conf\_int\_mc <- replicate(n\_sim, conf\_intervals(n,lambda))

# Proportion of times when 1st C.I is accurate for lambda=0.01, i.e. number of times when mean\_exp = 1/lambda lies in # calculated C.I.

> coverage\_prob\_conf\_int1 <- mean( (mean\_exp >= conf\_int\_mc[1,])\*(mean\_exp <= conf\_int\_mc[2,]) )

> coverage\_prob\_conf\_int1

[1] 0.8192

# Proportion of times when 2nd C.I is accurate for lambda=0.01, i.e. number of times when mean\_exp = 1/lambda lies in # calculated C.I.

> coverage\_prob \_conf\_int2 <- mean( (mean\_exp >= conf\_int\_mc[3,])\*(mean\_exp <= conf\_int\_mc[4,]) )

> coverage\_prob\_conf\_int2

[1] 0.9002

So, we can say that for a given ‘n’ and ‘lambda’, among the 5000 times, 81.92% of times the confidence interval1 is accurate and 90.02 % of times the confidence interval2 is accurate. It’s not good because confidence interval should have been correct atleast 95% of the times (as 1-alpha = 0.95)

**b).** As calculated in part a), we can calculate the coverage probabilities for different values of ‘n’ and ‘lambda’.

Coverage probabilities for confidence interval 1:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | n = 5 | n = 10 | n = 30 | n = 100 |
| Lambda= 0.01 | 0.8192 | 0.8758 | 0.9156 | 0.938 |
| Lambda= 0.1 | 0.8024 | 0.864 | 0.9124 | 0.9354 |
| Lambda= 1 | 0.8144 | 0.875 | 0.9176 | 0.9338 |
| Lambda= 10 | 0.8204 | 0.8666 | 0.918 | 0.9392 |

Coverage probabilities for confidence interval 2:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | n = 5 | n = 10 | n = 30 | n = 100 |
| Lambda= 0.01 | 0.9002 | 0.9254 | 0.9394 | 0.9444 |
| Lambda= 0.1 | 0.8946 | 0.9158 | 0.9346 | 0.9454 |
| Lambda= 1 | 0.8952 | 0.9232 | 0.9386 | 0.9398 |
| Lambda= 10 | 0.8994 | 0.917 | 0.9378 | 0.9486 |

**c)** More the value of ‘n’, more will be the accuracy of both the intervals.

Even for n = 1000, lambda = 1, coverage probability for interval 1 is 0.9496 and for interval 2 is 0.9492. Hence,