Part 2

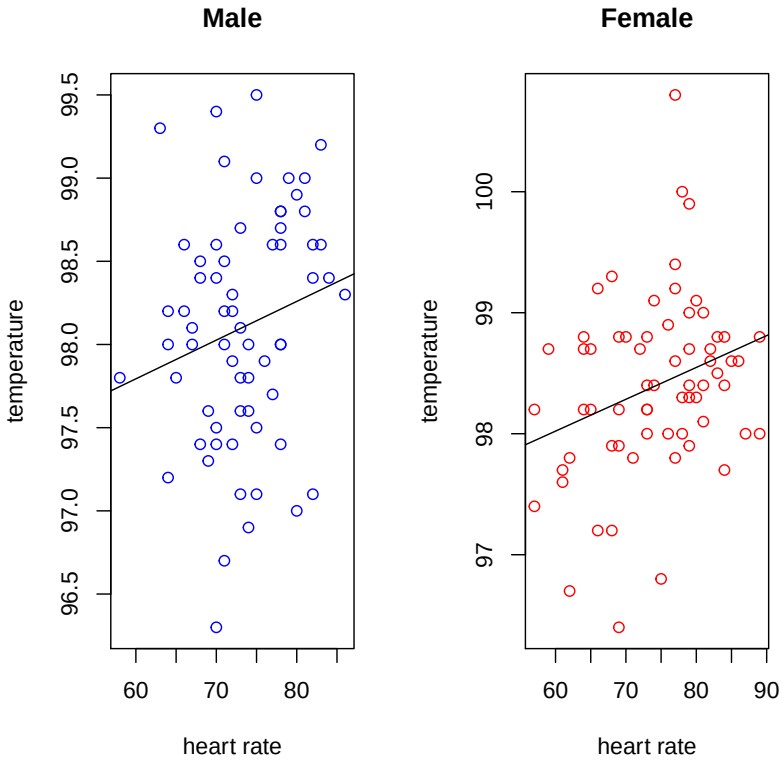
Team 8

May 1, 2018

Contents

Task 102

a)

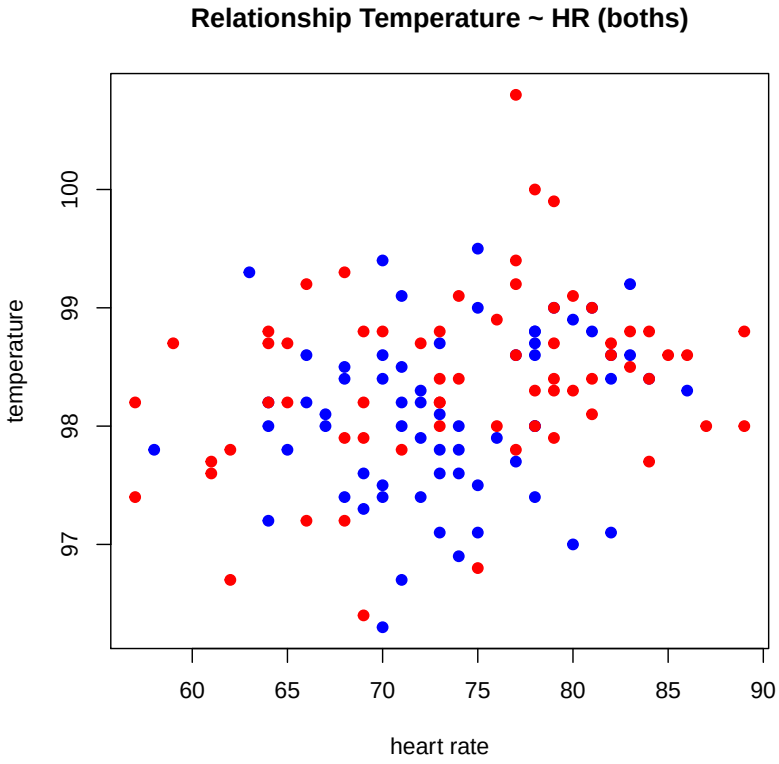


b)

```
> # Strength of relationships is represented by correlation:
> # Men
> cor(data_men[,2],data_men[,1], method="pearson")

[1] 0.1955894

> # rank correlation coefficients
> cor(data_men[,2],data_men[,1], method="kendall")

[1] 0.145074

> cor(data_men[,2],data_men[,1], method="spearman")

[1] 0.2239387

> # Women
> cor(data_women[,2],data_women[,1], method="pearson")

[1] 0.2869312

> # rank correlation coefficients
> cor(data_women[,2],data_women[,1], method="kendall")

[1] 0.1874097

> cor(data_women[,2],data_women[,1], method="spearman")

[1] 0.2655711
```

c)

```
> # Compare relationship for males and females
> plot(normtemp[,3],
+      normtemp[,1],
+      col=ifelse(normtemp[2] == 1, "blue", "red"),
+      main="Relationship Temperature ~ HR (boths)",
+      xlab = "heart rate",
+      ylab = "temperature",
+      pch=19)
>
```



Resume

1. Scatterplots are made
2. We can see that boths males and females show the same result: higher heart rate leads to higher temperature overall.
3. The coefficients are calculated
4. Yes, relationship for males appear to be the same as for females.
5. The scatterplot is made

Task 104

a)

Considering that the two samples are distributed normally and have the same variance,

$$\bar{X} - \bar{Y} \sim N(\mu_x - \mu_y, \sigma^2(1/n + 1/m))$$

where n and m denotes the number of observations from each sample. Since $s = m + n$:

$$\bar{X} - \bar{Y} \sim N(\mu_x - \mu_y, \sigma^2(1/n + 1/(s-n)).$$

We know that the confidence interval for the difference between two normally distributed means is

$$\bar{X} - \bar{Y} \pm t_{s-2,1-\alpha/2} S_{\bar{X}-\bar{Y}}$$

where $S_{\bar{X}-\bar{Y}} = S_p\sqrt{1/n + 1/(s-n)}$ and S_p is the root of the pooled variance.

We see that for a given size s we have to minimize $S_{\bar{X}-\bar{Y}}$ in order to minimize the length of the confidence interval.

So let us do that:

$$S_{\bar{X}-\bar{Y}} = S_p\sqrt{\frac{1}{n} + \frac{1}{s-n}} = S_p\sqrt{\frac{s-n+n}{n(s-n)}} = S_p\sqrt{\frac{s}{n(s-n)}}.$$

The expression is minimal, when $\frac{s}{n(s-n)}$ is minimal. FONC gives us:

$$\frac{\partial s(n(s-n)^{-1})}{\partial n} = \frac{2ns - s^2}{(ns - n^2)^2} \stackrel{!}{=} 0$$

$$n = \frac{s}{2}.$$

The second derivative is positive(or n = s/3 gives bigger value), so the expression is minimal when $n = m$, meaning that we should have the same size for both samples in order to get the shortest confidence interval.

b)

In the second part of the problem, we are asked to make the test of H_0 : $\mu_x = \mu_y$ as powerful as possible. There is a relationship between a confidence interval and the power of a test: shortest confidence interval gives the maximum rejection region and consequently the most powerful test. Hence, the most powerful test is when $n = m$.

Task 108

X and Y are independent and normally dustributed, hence their difference is also normally distributed:

$$X - Y \sim N(\mu_x - \mu_y, \sigma_x^2 + \sigma_y^2)$$

$$P(X < Y) = P(X - Y < 0)$$

We standardise to get:

$$\begin{aligned} P(X - Y < 0) &= P(\frac{(x-y) - (\mu_x - \mu_y)}{\sqrt{\sigma_x + \sigma_y}} < \frac{-(\mu_x - \mu_y)}{\sqrt{\sigma_x + \sigma_y}}) \\ &= \Phi(\frac{-(\mu_x - \mu_y)}{\sqrt{\sigma_x + \sigma_y}}) = \Phi(\frac{\mu_x - \mu_y}{\sqrt{\sigma_x + \sigma_y}}) \end{aligned}$$

Task 109

a)

Firstly, assume that X and Y are from normal distribution.

PAIRED:

Consider the difference:

$$D_i = X_i - Y_i$$

D_i are independent and with $\mathbb{E}[D_i] = \mu_X - \mu_Y$ and $\text{Var}(D_i) = \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y$ So we have:

$$D_i \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y)$$

Considering: $Cov(X_i, Y_i) = 50$, $\mu_X = \mu_Y = 10$, $\sigma_X^2 = \sigma_Y^2 = 10^2$, $\rho = 0.5$, $i = 1, ..., 25$
We have:

$$\bar{D_i} = (\bar{X} - \bar{Y}) \sim N(\mu_X - \mu_Y, \frac{1}{25} * (100 + 100 - 0.5 * 10 * 10) = N(\mu_X - \mu_Y, 4)$$

Normalizing:

$$Z_i = \frac{\bar{D}_i - (\mu_X - \mu_Y)}{2} \sim N(0,1)$$

The test statistic under H_0 is: $Z = \frac{\bar{X}-\bar{Y}}{2} \sim N(0,1)$

For the power function we obtain:

$$\begin{aligned}
\mathbb{P}\left(|Z| > Z_{\frac{\alpha}{2}}\right) &= 1 - \mathbb{P}\left(-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}\right) \\
&= 1 - \mathbb{P}\left(-Z_{\frac{\alpha}{2}} - \frac{\epsilon}{2} < \frac{\bar{X} - \bar{Y} - \epsilon}{2} < (Z_{\frac{\alpha}{2}} - \frac{\epsilon}{2}\right) \\
&= 1 - (\phi(Z_{\frac{\alpha}{2}} - \frac{\epsilon}{2}) - (1 - (\phi(Z_{\frac{\alpha}{2}} + \frac{\epsilon}{2}))) \\
&= 2 - (\phi(Z_{\frac{\alpha}{2}} - \frac{\epsilon}{2}) - \phi(Z_{\frac{\alpha}{2}} + \frac{\epsilon}{2})
\end{aligned}$$

b)

For the unpaired design we have:

$$(X_i, Y_i) \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$$
$$(\bar{X} - \bar{Y}) \sim N(\mu_X - \mu_Y, \sigma_X^2(\frac{1}{25} + \frac{1}{25}) = 8)$$

The test statistic under H_0 is: $Z = \frac{\bar{X}-\bar{Y}}{\sqrt{(8)}} \sim N(0,1)$

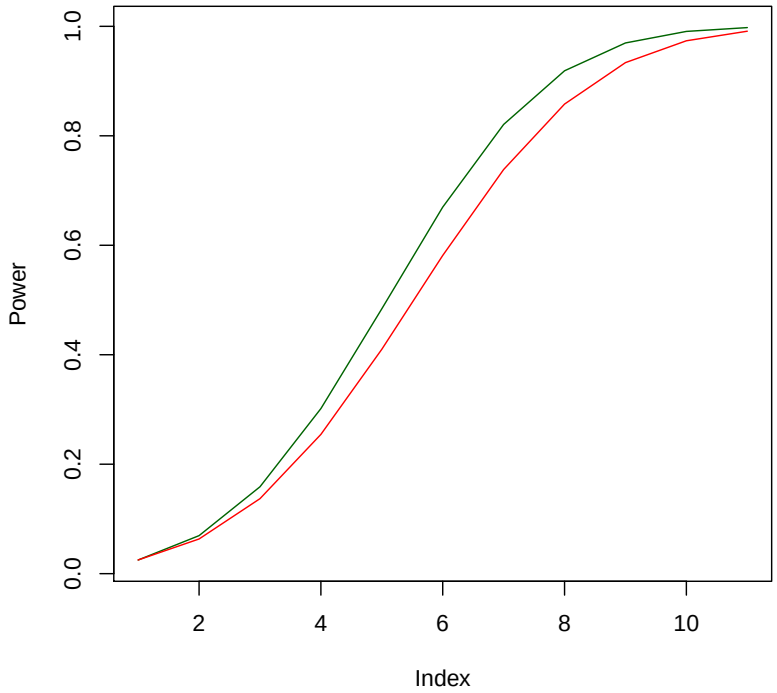
For the power function we obtain:

$$\begin{aligned}
\mathbb{P}\left(|Z| > Z_{\frac{\alpha}{2}}\right) &= 1 - \mathbb{P}\left(-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}\right) \\
&= 1 - \mathbb{P}\left(-Z_{\frac{\alpha}{2}} - \frac{\epsilon}{\sqrt{(8)}} < \frac{\bar{X} - \bar{Y} - \epsilon}{\sqrt{(8)}} < (Z_{\frac{\alpha}{2}} - \frac{\epsilon}{\sqrt{(8)}}\right) \\
&= 1 - (\phi(Z_{\frac{\alpha}{2}} - \frac{\epsilon}{\sqrt{(8)}}) - (1 - (\phi(Z_{\frac{\alpha}{2}} + \frac{\epsilon}{\sqrt{(8)}}))) \\
&= 2 - (\phi(Z_{\frac{\alpha}{2}} - \frac{\epsilon}{\sqrt{(8)}}) - \phi(Z_{\frac{\alpha}{2}} + \frac{\epsilon}{\sqrt{(8)}})
\end{aligned}$$

Finally we can plot:

```
> plot(power.t.test(n = 25,
+                   delta = 0:10,
```

```
+                     sd = 10,
+                     sig.level = 0.05,
+                     power = NULL,
+                     type = "paired")$power,
+      type = "l",
+      col = "darkgreen",
+      ylab = "Power")
> lines(power.t.test(n = 25,
+                      delta = 0:10,
+                      sd = 8,
+                      sig.level = 0.05,
+                      power = NULL,
+                      type = "two.sample")$power,
+       col = "red")
```



From the graphs it is obvious that the powers for the paired case are always larger than the ones corresponding to the unpaired case. We conclude that the power of the paired design is greater than the power of the unpaired (independent) design since a correlation is positive. Therefore we might use

pairing in order to obtain more powerful tests and greater precision.

Task 113

Q113

a)

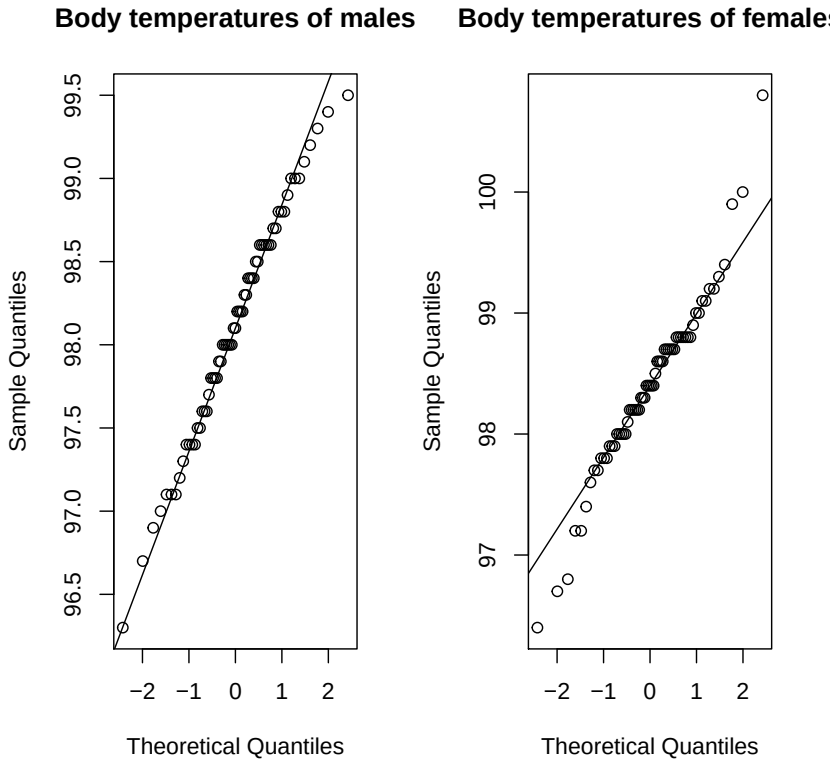
For the confidence interval we used the formula from the lecture notes:

$$(\bar{X} - \bar{Y}) \pm t_{m+n-2,1-\alpha/2} s_{\bar{X}-\bar{Y}}$$

```
> library(UsingR)
> head(normtemp)

  temperature gender hr
1        96.3      1 70
2        96.7      1 71
3        96.9      1 74
4        97.0      1 80
5        97.1      1 73
6        97.1      1 75

> data_men <- subset.data.frame(normtemp,gender == 1, select=c(temperature, hr)
> data_women <- subset.data.frame(normtemp,gender == 2, select=c(temperature, h
> mean1<-mean(data_men[,1])
> sigma1<-sd(data_men[,1])
> mean2<-mean(data_women[,1])
> sigma2<- sd(data_women[,1])
> s_p<-sqrt(((65 - 1) * sigma1^2 + (65 - 1) * sigma2^2) / (65 + 65 - 2))
> s_delta<-s_p*sqrt(1/65 + 1/65)
> mean1-mean2

[1] -0.2892308

> lower<-mean1-mean2-qt(0.975, 65 + 65 - 2)*s_delta
> upper<-mean1-mean2+qt(0.975, 65 + 65 - 2)*s_delta
> interval<-c(lower, upper)
> interval

[1] -0.53963938 -0.03882216
```

We can conclude that with the 95% confidence, the true mean of the population is between -0.54 and -0.04.
So, the difference of mean has no high deviation, and we can conclude, that female body temperature is slightly higher than males.
Now, let's draw the qqplots for males and females body temperature to see how much our data fits to normal distribution.

```
> par(mfrow=c(1,2))
> qqnorm(data_men[,1],main="Body temperatures of males")
> qqline(data_men[,1])
> qqnorm(data_women[,1],main="Body temperatures of females")
> qqline(data_women[,1])
```



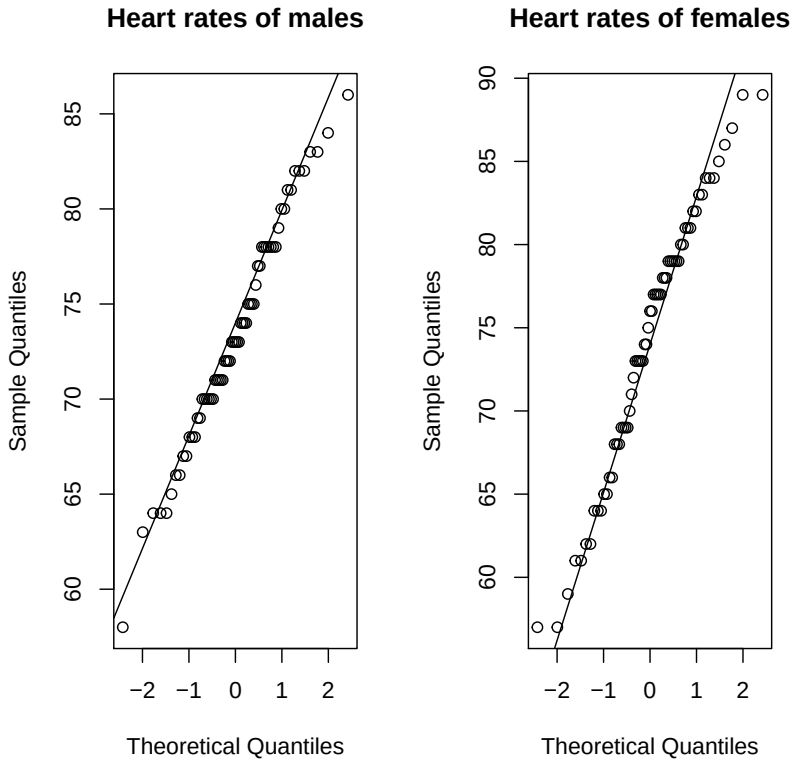
According to the qqplots we can conclude that body temperature are approximately normal distributed for males and females, but for females we have heavier tails. So, use of normal approximation is reasonable in this case.

b)
We are doing the same procedure but with variable heart rate.

```
> mean1<-mean(data_men[,2])
> sigma1<-sd(data_men[,2])
> mean2<-mean(data_women[,2])
> sigma2<- sd(data_women[,2])
> s_p<-sqrt(((65 - 1) * sigma1^2 + (65 - 1) * sigma2^2) / (65 + 65 - 2))
> s_delta<-s_p*sqrt(1/65 + 1/65)
> mean1-mean2

[1] -0.7846154

> lower<-mean1-mean2-qt(0.975, 65 + 65 - 2)*s_delta
> upper<-mean1-mean2+qt(0.975, 65 + 65 - 2)*s_delta
> interval<-c(lower, upper)
> interval

[1] -3.241461  1.672230
```

In this case wee see that the difference mean is -0.78 (it means that females have slightly higher heart rate than males), the deviation interval for mean is higher in this case (-3.24, 1.67).

```
> par(mfrow=c(1,2))
> qqnorm(data_men[,2],main="Heart rates of males")
> qqline(data_men[,2])
> qqnorm(data_women[,2],main="Heart rates of females")
> qqline(data_women[,2])
```



On the qqplot it is visible, both variables are approximately normally distributed. So, we can use normal approximation. And found confidence interval is reasonable.

c)

Firstly we will do parametric t-test(comparing 2 means) for the 2 independent samples:

H_0: the difference of body temepatures between males and females is equal to 0;
H_A: the difference of body temepatures between males and females is not equal to 0.

```
> t.test(data_men[,1],data_women[,1])

        Welch Two Sample t-test

data:  data_men[, 1] and data_women[, 1]
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
```

P-value = 0.02394, so we reject H_0 that the difference of body temepatures between males and females is equal to 0.
Now the same t-test but for heart rate:

H_0: the difference of heart rates between males and females is equal to 0;
H_A: the difference of heart rates between males and females is not equal to 0.

```
> t.test(data_men[,2],data_women[,2])

        Welch Two Sample t-test

data:  data_men[, 2] and data_women[, 2]
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
```

P-value = 0.5287, hence we accept H_0 that the difference of heart rates between males and females is equal to 0.

Now we will use nonparametric test.

```
> wilcox.test(data_men[,1],data_women[,1])

        Wilcoxon rank sum test with continuity correction

data:  data_men[, 1] and data_women[, 1]
W = 1637, p-value = 0.02676
alternative hypothesis: true location shift is not equal to 0
```

We can make the same conclusion as with parametric test aqbove, that the difference of body temepatures between males and females is not equal to 0.

```
> wilcox.test(data_men[,2],data_women[,2])

        Wilcoxon rank sum test with continuity correction

data:  data_men[, 2] and data_women[, 2]
W = 1927.5, p-value = 0.3898
alternative hypothesis: true location shift is not equal to 0
```

And here we see the same result as above in case of parametric case. We will accept H_0, there is no difference in heart rate between males and females.

Task 114

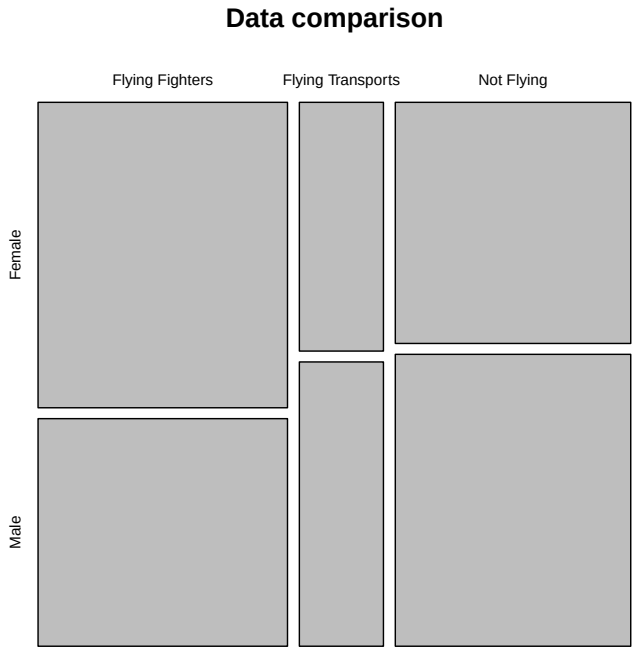
H_0 - "There are no differences", H_1 - "There are differences".

So let us have a closer look on the data.

```
> father <- matrix(c(51,14,38,38,16,46),nrow=3)
> colnames(father) <- c("Female", "Male")
> rownames(father) <- c("Flying Fighters", "Flying Transports", "Not Flying")
> father

                 Female Male
Flying Fighters      51   38
Flying Transports    14   16
Not Flying           38   46
```

So that is the data and graphically we can illustrate it with mosaicplots:

```
> mosaicplot(father, main="Data comparison")
```



We see that non-flying people in the data have indeed less girls than boys born. Let us check, if this impression is statistically significant.
We can now to try to investigate the data by applying the Pearson's Chi-square test for homogeneity. We will test the table overall and again each group against each other.

```
> chisq.test(father)

        Pearson's Chi-squared test

data:  father
X-squared = 2.7504, df = 2, p-value = 0.2528
```

We stick to the null that there is no difference between the 3 groups.

```
> # flying fighters vs flying transport
> chisq.test(father[1:2,1:2])

        Pearson's Chi-squared test with Yates' continuity correction

data:  father[1:2, 1:2]
X-squared = 0.63997, df = 1, p-value = 0.4237
```

We cannot reject the null. There is no significant difference between the two categories of pilots.

```
> # flying fighters vs not flying
> chisq.test(father[c(1,3),1:2])

        Pearson's Chi-squared test with Yates' continuity correction

data:  father[c(1, 3), 1:2]
X-squared = 2.0585, df = 1, p-value = 0.1514
```

We also cannot reject the null. There is no significant difference between flying fighters and the group of non-flyers.

```
> # flying transport vs not flying
> chisq.test(father[2:3,1:2])

        Pearson's Chi-squared test with Yates' continuity correction

data:  father[2:3, 1:2]
X-squared = 8.3629e-31, df = 1, p-value = 1
```

We also cannot reject the null. There is no difference between transporting flyers and the group of non-flyers.

Let us now have a look on the flyers (sum of pilots for transports and fights) together vs the non-flyers:

```
> fly_nonfly <- rbind(colSums(father[1:2,]), father[3,])
> colnames(fly_nonfly) <- c("Female", "Male")
> rownames(fly_nonfly) <- c("Pilots", "Not Flying")
> fly_nonfly

           Female Male
Pilots         65   54
Not Flying     38   46

> # pilots vs non-flyers
> chisq.test(fly_nonfly)

        Pearson's Chi-squared test with Yates' continuity correction

data:  fly_nonfly
X-squared = 1.3796, df = 1, p-value = 0.2402
```

We see also here we cannot find a significant difference. Thus, we stick to the null hypothesis that there are no differences between flyers and non-flyers wrt to there offsprings.

The exercise also asks to compare it with the overall sex ratio of the United states which is $105,37/100$. From this ratio we can calculate the probability to be a boy/girl in the US and find out what the expected numbers of girls/boys should be. We can test again with a the same test procedures as above.

```
> p_boy <- 105.37/205.37
> p_girl <- 1- p_boy
> # likelihood to be boy/girl in the US
> c(p_boy,p_girl)

[1] 0.513074 0.486926
```

Given this "theoretical" probabilities, we can test again with the chi-squared test.

```
> chisq.test(father,correct = F, p=c(p_boy,p_girl))

        Pearson's Chi-squared test

data:  father
X-squared = 2.7504, df = 2, p-value = 0.2528
```

Also compared to the US sex ratio, our data doesn't differ significantly. We don't reject the null. Our data seems to be conistent with the overall US sex ratio.

Task 115

We use the Chi-squared Test for independence, which tests $H_0 : \pi_{ij} = \pi i.\pi_{.j}$ against π_{ij}'s are free, where i's are numbers of the column and j's are - rows. The test has the following form:

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n} \sim \chi^2_{(J-1)(I-1)}$$

```
> A <- matrix(c(8,15,13,14,19,15,15,4,7,3,1,4),3,4)
> A
```

```
     [,1] [,2] [,3] [,4]
[1,]    8   14   15    3
[2,]   15   19    4    1
[3,]   13   15    7    4
```

```
> chisq.test(A)

        Pearson's Chi-squared test

data:  A
X-squared = 12.183, df = 6, p-value = 0.058
```

We have p-value to be larger, than 0.05, thus we do not reject the hypothesis that there is no dependence between courses and grades at 5% significance level.

Task 116

We can build up in R our exercise in order to use the Pearson's Chi-Square test statistic.
First we create the matrix:

```
> matrice <- matrix(c(14,16,8,2,133,180,93,81,12,14,12,1,241,285,139,153,11,6,8
> rownames(matrice)<-c("a PB such","a NPB such","and FB I","and NFB I","the PB
> colnames(matrice)<-c("Sense and Sensibility","Emma","Sandition I","Sandition
> matrice
```

```
           Sense and Sensibility Emma Sandition I Sandition II
a PB such                     14   16           8            2
a NPB such                   133  180          93           81
and FB I                      12   14          12            1
and NFB I                    241  285         139          153
the PB on                     11    6           8           17
the NPB on                   259  265         221          204
```

Now we use the test for different words:

```
> chisq.test(matrice[ , 1 : 2])

        Pearson's Chi-squared test

data:  matrice[, 1:2]
X-squared = 6.1744, df = 5, p-value = 0.2896
```

With level of significance of 5%, the $\chi^2_6 = 12.59$, therefore we accept the null: there is no difference between the first two words taken into consideration (sense and sensibility and Emma)

```
> chisq.test(matrice[ , 1 : 3])

        Pearson's Chi-squared test

data:  matrice[, 1:3]
X-squared = 23.287, df = 10, p-value = 0.009735
```

With level of significance of 5%, the $\chi^2_{10} = 18.31$, therefore we reject the null: there is difference between the first three words taken into consideration.

```
> chisq.test(matrice[ , 3 : 4])

        Pearson's Chi-squared test

data:  matrice[, 3:4]
X-squared = 17.774, df = 5, p-value = 0.003244
```

With level of significance of 5%, the $\chi^2_5 = 11.07$, therefore we reject null: there is difference between the third and fourth words taken into consideration.

```
> chisq.test(matrice[ , 1 : 4])

        Pearson's Chi-squared test

data:  matrice[, 1:4]
X-squared = 52.436, df = 15, p-value = 4.783e-06
```

Task 117

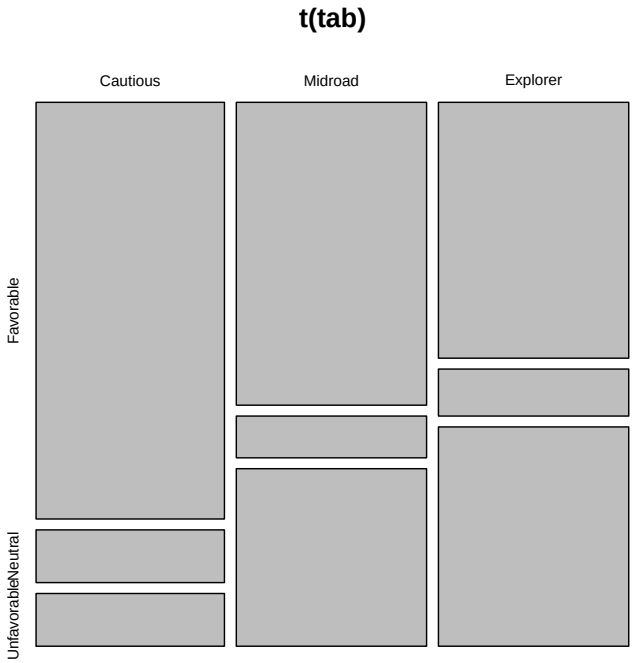
We will use Chi-Squared Test of Independence. Firstly, we need to formulate hypothesis:

$H_0 : \pi_{ij} = \pi_i \pi_j, \quad i = 1, 2, 3, j = 1, 2, 3$ (rows and columns classifications are independent)
$H_A : \pi_{ij}$ are free

```
> tab <- matrix(c(79, 58, 49, 10, 8, 9, 10, 34, 42), nrow = 3, byrow = TRUE)
> rownames(tab) <- c("Favorable", "Neutral", "Unfavorable")
> colnames(tab) <- c("Cautious", "Midroad", "Explorer")
> tab

            Cautious Midroad Explorer
Favorable         79      58       49
Neutral           10       8        9
Unfavorable       10      34       42

> chisq.test(tab)

        Pearson's Chi-squared test

data:  tab
X-squared = 27.289, df = 4, p-value = 1.737e-05
```

According to p-value, which is less than 0.05, we reject H_0, so we conclude there is the relationship between personality type and attitude towards small cars. And we can say that the preference to the small cars is depent on the personality of driver. If we have a look at frequencies plot:

```
> mosaicplot(t(tab))
```



We can notice that small cars are more favorable for cautious people. And favorability decreases for middle-of-the-road and explorers.