Homework 1

Group M: Billo, Pizzignacco, El Gataa, Shahazad

Introduction

Briefly describe the purpose of the homework or tasks to be accomplished.

Block B

CS: Chapter 1, exercises 1.1, 1.6

```
# Placeholder for code for Exercise B1
```

CS: Chapter 3, exercises 3.5, 3.6

```
# Placeholder for code for Exercise B1
```

FSDS: Chapter 2, exercises 2.8, 2.16, 2.21, 2.26, 2.52, 2.53, 2.70

```
# Placeholder for code for Exercise B2
```

FSDS: Chapter 3, exercises 3.18, 3.28, 3.24

```
# Placeholder for code for Exercise B3
```

FSDS: Chapter 4, exercises 4.14, 4.16, 4.48

Exercise 4.14

```
data = read.table("https://stat4ds.rwth-aachen.de/data/Students.dat", header = TRUE)
summary(data)
```

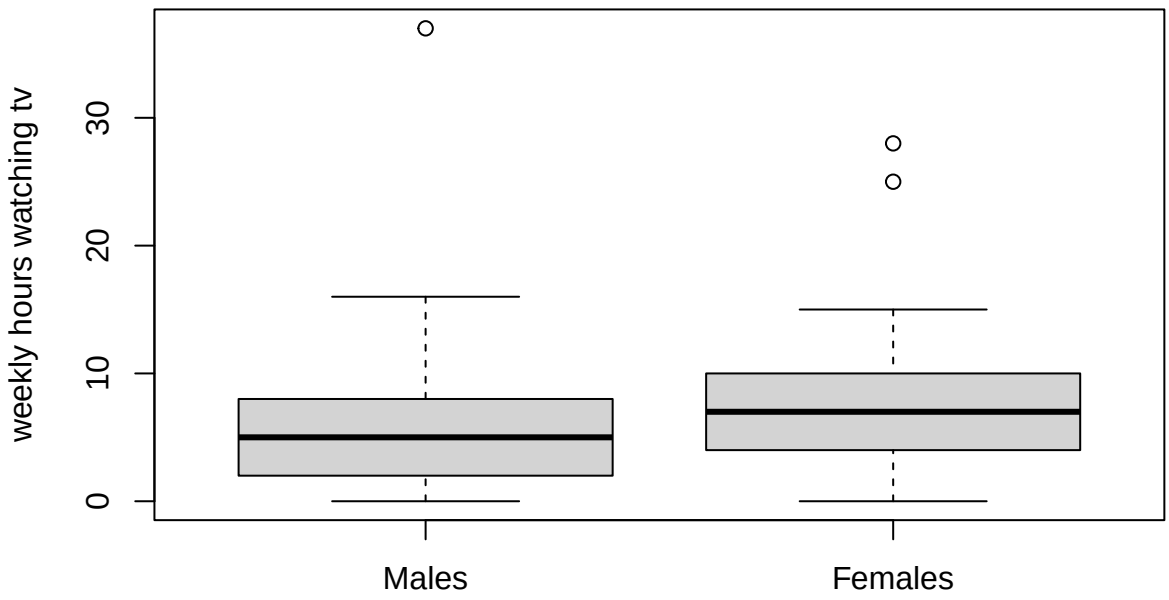
```
##     subject          gender            age            hsgpa
##  Min.   : 1.00   Min.   :0.0000   Min.   :22.00   Min.   :2.000
##  1st Qu.:15.75   1st Qu.:0.0000   1st Qu.:24.00   1st Qu.:3.000
```

```
##  Median :30.50   Median :1.0000   Median :26.50   Median :3.350
##  Mean   :30.50   Mean   :0.5167   Mean   :29.17   Mean   :3.308
##  3rd Qu.:45.25   3rd Qu.:1.0000   3rd Qu.:31.00   3rd Qu.:3.625
##  Max.   :60.00   Max.   :1.0000   Max.   :71.00   Max.   :4.000
##      cogpa           dhome           dres             tv
##  Min.   :2.600   Min.   :   0   Min.   : 0.200   Min.   : 0.000
##  1st Qu.:3.175   1st Qu.: 205   1st Qu.: 1.450   1st Qu.: 3.000
##  Median :3.500   Median : 640   Median : 2.000   Median : 6.000
##  Mean   :3.453   Mean   :1232   Mean   : 3.818   Mean   : 7.267
##  3rd Qu.:3.725   3rd Qu.:1350   3rd Qu.: 5.000   3rd Qu.:10.000
##  Max.   :4.000   Max.   :8000   Max.   :20.000   Max.   :37.000
##      sport             news            aids             veg
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   :0.00
##  1st Qu.: 3.000   1st Qu.: 2.000   1st Qu.: 0.000   1st Qu.:0.00
##  Median : 5.000   Median : 3.000   Median : 0.500   Median :0.00
##  Mean   : 5.483   Mean   : 4.083   Mean   : 1.433   Mean   :0.15
##  3rd Qu.: 7.000   3rd Qu.: 5.250   3rd Qu.: 2.000   3rd Qu.:0.00
##  Max.   :16.000   Max.   :14.000   Max.   :11.000   Max.   :1.00
##      affil          ideol           relig            abor
##  Min.   :1.00   Min.   :1.000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:1.00   1st Qu.:2.000   1st Qu.:0.750   1st Qu.:1.0000
##  Median :2.00   Median :2.000   Median :1.000   Median :1.0000
##  Mean   :2.05   Mean   :3.033   Mean   :1.167   Mean   :0.7833
##  3rd Qu.:3.00   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :3.00   Max.   :7.000   Max.   :3.000   Max.   :1.0000
##      affirm            life
##  Min.   :0.0000   Min.   :1.00
##  1st Qu.:0.0000   1st Qu.:1.00
##  Median :1.0000   Median :1.00
##  Mean   :0.7167   Mean   :1.75
##  3rd Qu.:1.0000   3rd Qu.:3.00
##  Max.   :1.0000   Max.   :3.00
```

```r
#a
x_bar = mean(data$tv)
s = sd(data$tv)
# H0: the mean is 7.2
# H1: the mean is different from 7.2
z = qnorm(0.975) # having n > 60, we can assume normality
SE = s/sqrt(length(data$tv))
CI = x_bar + c(-1, 1)*z*SE
# we can say that 95% of the students on average spend between 5.30 and almost 9 hours watching TV per

#b
only_male = data %>% filter(gender == 0)
only_female = data %>% filter(gender == 1)

boxplot(only_male$tv, only_female$tv, names = c("Males", "Females"), ylab = "weekly hours watching tv")
```



assuming -independent populations -equal variances -normality -> weakest assumption, barely 30 observations –> use t student's distribution

```
x_bar_male = mean(only_male$tv)
x_bar_female = mean(only_female$tv)
s_male = sd(only_male$tv)
s_female = sd(only_female$tv)
n_male = length(only_male$tv)
n_female = length(only_female$tv)

# assuming equal variance
s_p = ((n_male -1)*(s_male^2) + (n_female - 1)*(s_female^2))/(n_male + n_female -1)
# t = (x_bar_male - x_bar_female)/(s_p*(1/n_male + 1/n_female))
SE_diff = (x_bar_male - x_bar_female)/((s_p^2)*(1/n_male + 1/n_female))
mean_diff = x_bar_male - x_bar_female
z = qt(0.975, n_male + n_female - 2)
CI_diff = mean_diff + c(-1, 1)*z*SE_diff

# Test for equality of means
t.test(only_male$tv, only_female$tv, conf.level = 0.05)
```

```
##
##  Welch Two Sample t-test
##
## data:  only_male$tv and only_female$tv
## t = -0.84995, df = 56.249, p-value = 0.399
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 5 percent confidence interval:
##  -1.593836 -1.373906
## sample estimates:
## mean of x mean of y
##  6.500000  7.983871
```

We can not state that the variances are significantly different betweent the 2 groups However the confidence interval seems to suggest that females watch slightly more tv than males.

```
# assuming unequal variance
SE_diff = sqrt((s_male^2)/n_male + (s_female^2)/n_female)
# t = (x_bar_male - x_bar_female)/(s_p*(1/n_male + 1/n_female))
mean_diff = x_bar_male - x_bar_female
z = qt(0.975, n_male + n_female - 2)
CI_diff_uvars = mean_diff + c(-1, 1)*z*SE_diff
```

Now we the interval crosses 0?? how is such a different result possible?

Exercise 4.16

```
data = read.table("https://stat4ds.rwth-aachen.de/data/Substance.dat", header = TRUE)

# compare alcohol users and non-users
alpha = 0.05
n = sum(data$count)
# find the total number of students that have or haven't used alcohol
N_alcohol_total = sum(data[data$alcohol == "yes", 4])
N_NOalcohol_total = sum(data[data$alcohol == "no", 4])
N_marijuana = sum(data[data$alcohol == "yes" & data$marijuana == "yes", 4])
N_NOalcohol_marijuana = sum(data[data$alcohol == "no" & data$marijuana == "yes", 4])

pi_1_hat = N_marijuana/N_alcohol_total
pi_2_hat = N_NOalcohol_marijuana/N_NOalcohol_total

# by hand
z = qnorm(1-alpha/2)
SE = (pi_1_hat*(1 - pi_1_hat)/N_alcohol_total + pi_2_hat*(1 - pi_2_hat)/N_NOalcohol_total)
CI_prop = (pi_1_hat - pi_2_hat) + c(-1, 1)*z*sqrt(SE)

#with software
prop.test(c(955, 5), c(1949, 327), conf.level = alpha, correct = F)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(955, 5) out of c(1949, 327)
## X-squared = 258.73, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
```

```
## 5 percent confidence interval:
##  0.4738766 0.4755321
## sample estimates:
##     prop 1     prop 2
## 0.48999487 0.01529052
```

Interpretation: there seems to be a significant mean difference in the use of marijuana between students that used alcohol before and students who didn't.

Exercise 4.48

Given

$$\text{Given} \quad SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

With 95% likelihood, find:

$$SE \leq \frac{1}{\sqrt{n}} \quad \text{when } \hat{p} = 0.5$$

This is where SE is maximized.

$$SE = \sqrt{\frac{0.5 \times (0.5)}{n}} = \frac{\sqrt{0.25}}{\sqrt{n}} = \frac{0.5}{\sqrt{n}} = \frac{1}{2\sqrt{n}}$$

Example: When $\hat{p} \neq 0.5$

$$SE = \sqrt{\frac{0.3 \times (1-0.3)}{n}} = \sqrt{\frac{0.21}{n}} = \frac{0.458}{\sqrt{n}} \approx \frac{0.91}{2\sqrt{n}}$$

Example: When $\hat{p} = 0.2$

$$SE = \sqrt{\frac{0.2 \times (0.8)}{n}} = \sqrt{\frac{0.16}{n}} = \frac{0.4}{\sqrt{n}} = \frac{0.8}{2\sqrt{n}}$$

Example: When $\hat{p} = 0.7$

$$SE = \sqrt{\frac{0.7 \times (0.3)}{n}} = \sqrt{\frac{0.21}{n}} = \frac{0.458}{\sqrt{n}} \approx \frac{0.91}{2\sqrt{n}}$$

Observation

Note that when both $\hat{p} < 0.5$ and $\hat{p} > 0.5$ the numerator decreases, along with the standard error.

For Maximum Standard Error

Set the maximum SE to be within margin M.

$$\frac{1}{\sqrt{n}} = M$$

$$\Rightarrow \frac{1}{M} = \sqrt{n} \Rightarrow n = \frac{1}{M^2}$$

As long as $n \geq \frac{1}{M^2}$, our error will be within M.

FSDS: Chapter 5, exercises 5.2, 5.12, 5.50

```
# Placeholder for code for Exercise B5
```