Homework 4

Darian-Florian Voda

2022-11-16

Load packages

```
library(mosaic)
library(ggplot2)
library(latex2exp)
library(gridExtra)
library(knitr)
library(MASS)
library(cowplot)
library(PairedData)
```

Exercise 33

A hardware store buys ceramic tiles from three producers, A, B and C. The manager of the hardware store draws a sample of 1000 ceramic tiles per producer and determines the numbers of erroneous ceramic tiles:

- Producer A: 41 erroneous ceramic tiles
- Producer B: 21 erroneous ceramic tiles
- Producer C: 55 erroneous ceramic tiles

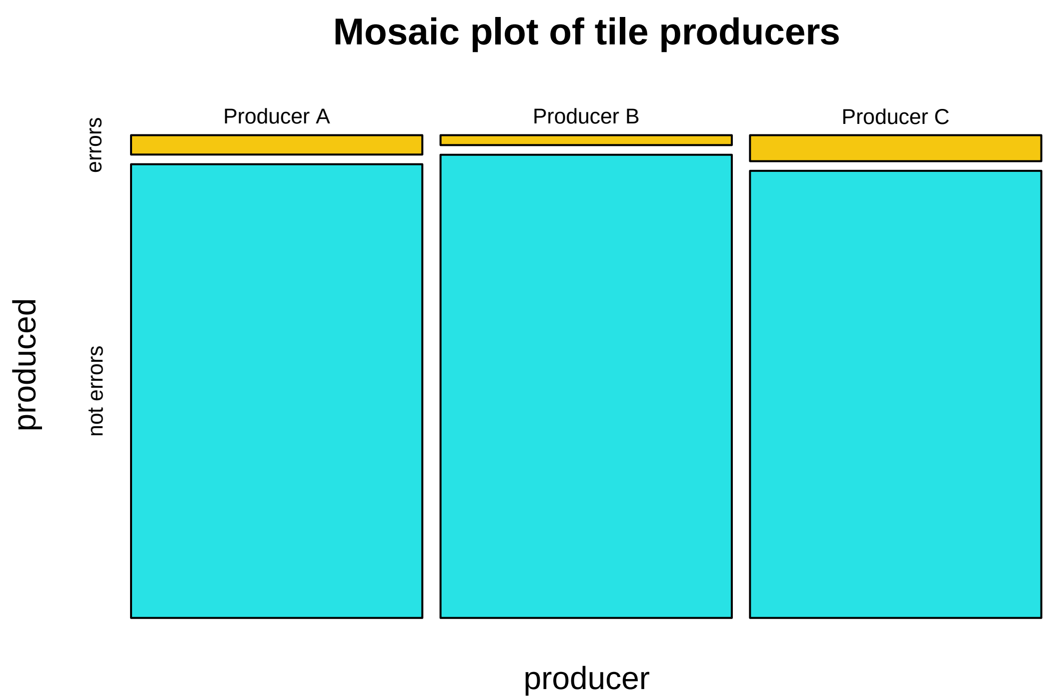
Do the producers deliver ceramic tiles of equal quality? $\alpha = 1\%$? (H_0)

```
####  Exercise 33 ####

data <- matrix(c(41, 21, 55, 1000 - 41, 1000 - 21, 1000 - 55), ncol=2)
dimnames(data) <- list(
  producer = c("Producer A", "Producer B", "Producer C"),
  produced = c("errors", "not errors"))
kable(data)
```

	errors	not errors
Producer A	41	959
Producer B	21	979
Producer C	55	945

```r
mosaicplot(data, col=c(7,5), main="Mosaic plot of tile producers")
```




```r
errors_producer <- sum(data[, "errors"]) / sum(data)
errors_producer
```

```
## [1] 0.039
```

```r
not_errors_producer <- sum(data[, "not errors"]) /
  sum(data)
not_errors_producer
```

```
## [1] 0.961
```

```r
errors_producer + not_errors_producer
```

```
## [1] 1
```

```r
expected <- c(
  (data[, "errors"] + data[, "not errors"]) * errors_producer,
  (data[, "errors"] + data[, "not errors"]) * not_errors_producer)

expected
```

```
## Producer A Producer B Producer C Producer A Producer B Producer C
##         39         39         39        961        961        961
```

```r
table_expected <- matrix(expected, ncol=2)
dimnames(table_expected) <- list(
  producer = c("Producer A", "Producer B", "Producer C"),
  produced = c("errors", "not errors"))
kable(table_expected)
```

	errors	not errors
Producer A	39	961
Producer B	39	961
Producer C	39	961

```r
x <- sum((data - expected)^2 / expected)
x
```

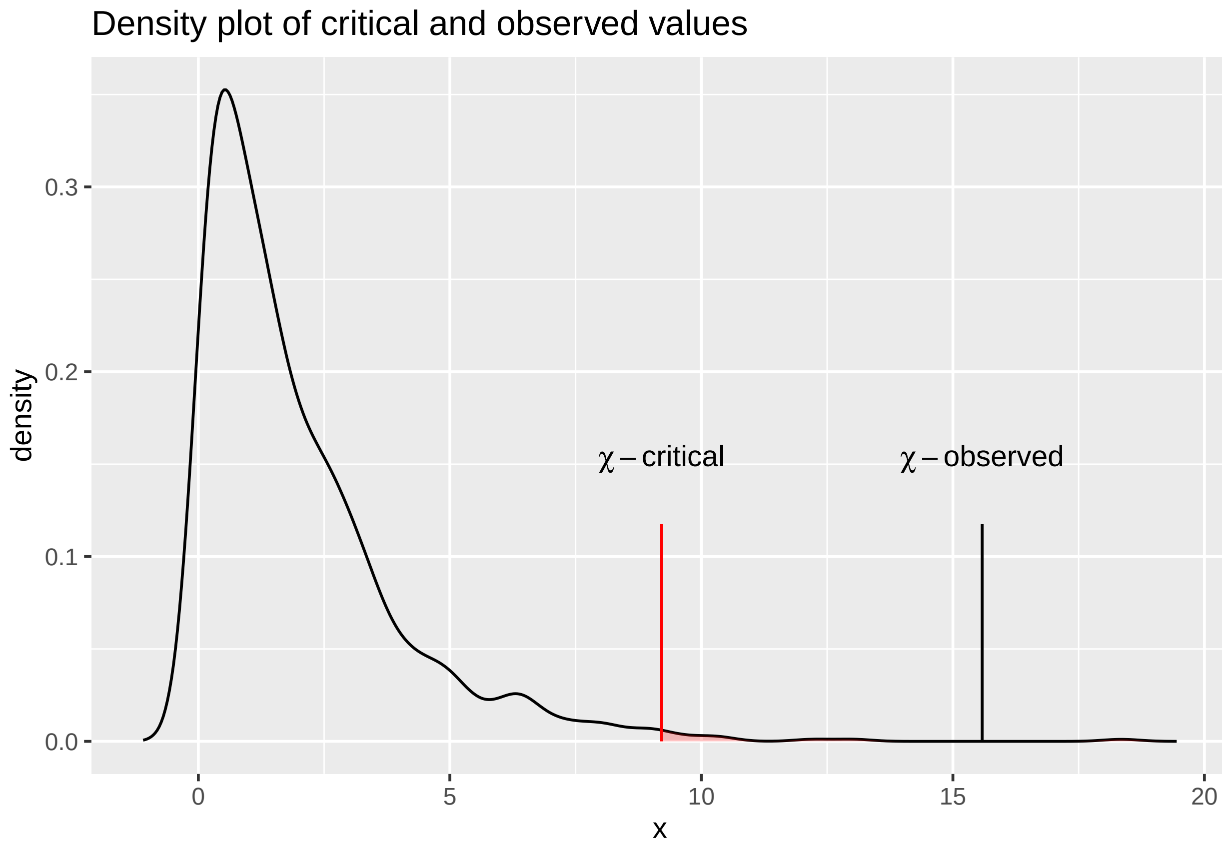
```
## [1] 15.58206
```

```r
alpha = 0.01
df = 2
crit_val <- qchisq(alpha, df, lower.tail = FALSE)
crit_val
```

```
## [1] 9.21034
```

```r
plot_chisq <- function(sim_data, x, crit_val, title = "Density plot of critical and observed values"
) {
  annotation_y_val <- max(sim_data$y) * 1/3
  plt <- ggplot(sim_data, aes(x, y)) + ggtitle(title) +
    labs(y = "density") + geom_line() + geom_area(
      data = subset(sim_data, x >= crit_val),
      fill = "red", alpha = 0.24) +
    annotate("segment", x = crit_val, xend = crit_val,
             y = 0, yend = annotation_y_val,
             color = "red") + annotate("text",
                                        x = crit_val, y = annotation_y_val * 1.3,
                                        label = "chi-critical", parse = TRUE) +
    annotate("segment", x = x, xend = x,
              y = 0, yend = annotation_y_val,
              color = "black") + annotate("text",
                                           x = x, y = annotation_y_val * 1.3, label = "chi-observed
  return(plt)}

sim_data <- data.frame(density(rchisq(1000, df))[c("x", "y")])
plot_chisq(sim_data, x, crit_val)
```




```
p_value <- 1 - pchisq(x, df)
p_value
```

```
## [1] 0.000413427
```

Thus, we can conclude that:

- P-value is smaller than 0.05, so we can Reject H_0
- Our critical value is 9.21, which is less than 15.58, so we can Reject H_0
- Since H_0 is rejected, the producers **do not** deliver ceramic tiles of equal quality

Exercise 34

A hardware store buys plastic bags from three producers, A, B and C. The manager of the hardware store draws draws different samples per producer and determines the numbers of broken plastic bags:

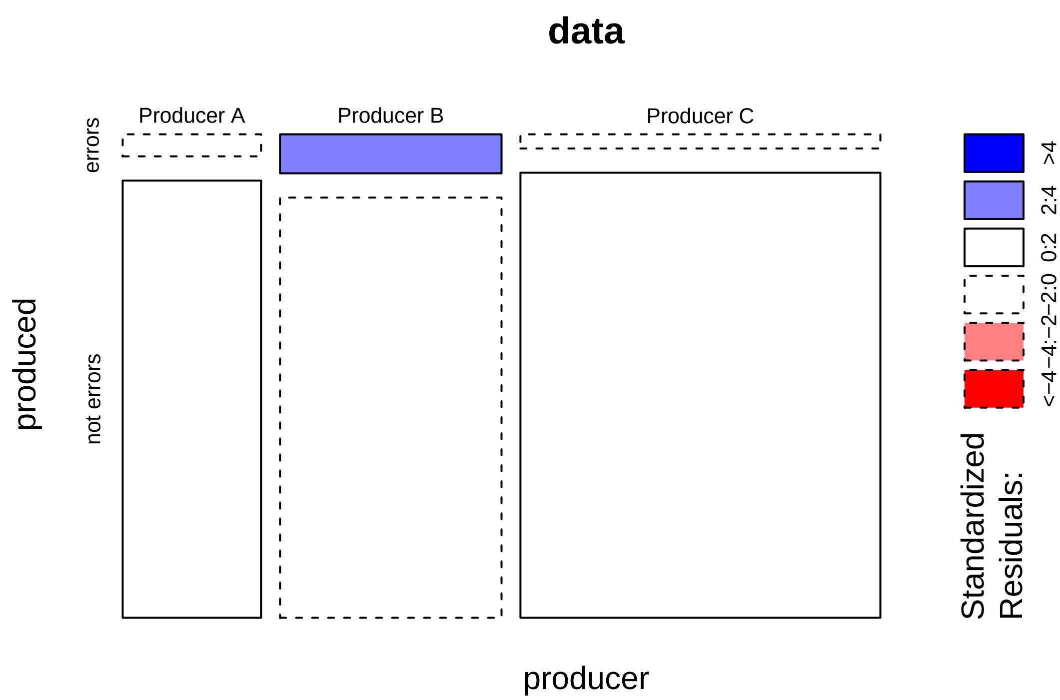
- Producer A: 6 broken plastic bags out of 125 samples
- Producer B: 17 broken plastic bags out of 200 samples
- Producer C: 10 broken plastic bags out of 325 samples

Do the producers deliver plastic bags of equal quality? $\alpha = 1\%$ and $\alpha = 5\%$ (H_0)

```r
#### Exercise 34 ####
data <- matrix(c(6, 17, 10, 125 - 6, 200 - 17, 325 - 10), ncol=2)
dimnames(data) <- list(
  producer = c("Producer A", "Producer B", "Producer C"),
  produced = c("errors", "not errors"))
kable(data)
```

	errors	not errors
Producer A	6	119
Producer B	17	183
Producer C	10	315

```r
mosaicplot(data, shade=TRUE, off = 5)
```




```r
errors_producer <- sum(data[, "errors"]) / sum(data)
errors_producer
```

```
## [1] 0.05076923
```

```r
not_errors_producer <- sum(data[, "not errors"]) /
  sum(data)
not_errors_producer
```

```
## [1] 0.9492308
```

```r
errors_producer + not_errors_producer
```

```
## [1] 1
```

```r
expected <- c(
  (data[, "errors"] + data[, "not errors"]) * errors_producer,
  (data[, "errors"] + data[, "not errors"]) * not_errors_producer)

expected
```

```
## Producer A Producer B Producer C Producer A Producer B Producer C
##   6.346154  10.153846  16.500000 118.653846 189.846154 308.500000
```

```r
table_expected <- matrix(expected, ncol=2)
dimnames(table_expected) <- list(
  producer = c("Producer A", "Producer B", "Producer C"),
  produced = c("errors", "not errors"))
kable(table_expected)
```

	errors	not errors
Producer A	6.346154	118.6538
Producer B	10.153846	189.8462
Producer C	16.500000	308.5000

```r
x <- sum((data - expected)^2 / expected)
x
```

```
## [1] 7.580301
```

```r
plot_chisq2 <- function(sim_data, x, crit_val, title = ""
) {
  annotation_y_val <- max(sim_data$y) * 1/3
  plt <- ggplot(sim_data, aes(x, y)) + ggtitle(title) +
    labs(y = "density") + geom_line() + geom_area(
      data = subset(sim_data, x >= crit_val),
      fill = "red", alpha = 0.24) +
    annotate("segment", x = crit_val, xend = crit_val,
             y = 0, yend = annotation_y_val,
             color = "red") +
    annotate("text", x = crit_val, y = annotation_y_val,
                                       label = "chi-critical", parse = TRUE) +
    annotate("segment", x = x, xend = x,
             y = 0, yend = annotation_y_val,
             color = "black") + annotate("text",
                                         x = x, y = annotation_y_val* 1.1, label = "chi-observed", pars
  return(plt)}
```

```r
# alpha = 0.01
alpha = 0.01
df = 2
crit_val1 <- qchisq(alpha, df, lower.tail = FALSE)
crit_val1
```

```
## [1] 9.21034
```

```r
sim_data <- data.frame(density(rchisq(125+200+325, df))[c("x", "y")])
plot1 = plot_chisq2(sim_data, x, crit_val1, title="Density plot of alpha=0.01")

p_value <- 1 - pchisq(x, df)
paste(round(p_value,5), "Thus, Reject H0")
```

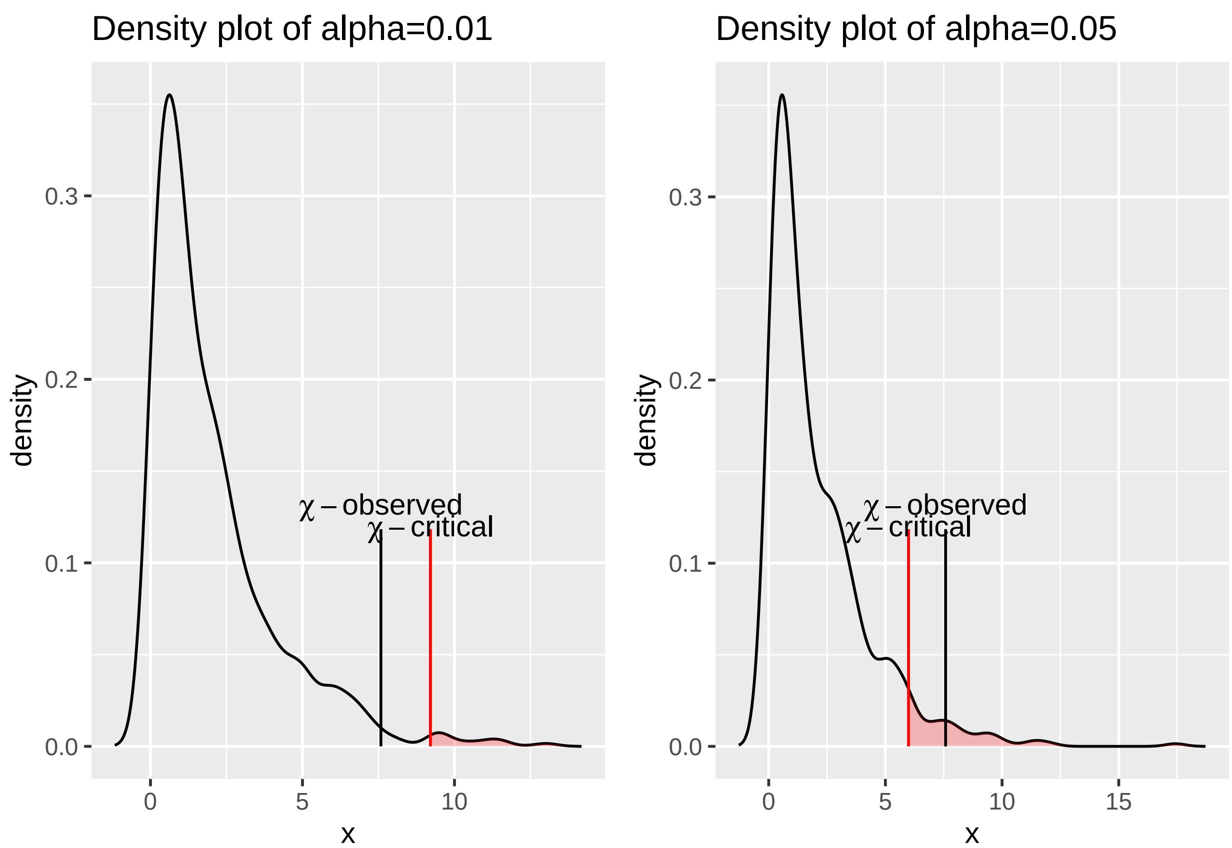
```
## [1] "0.02259 Thus, Reject H0"
```

```r
# alpha = 0.05
alpha = 0.05
df = 2
crit_val2 <- qchisq(alpha, df, lower.tail = FALSE)
crit_val2
```

```
## [1] 5.991465
```

```r
sim_data <- data.frame(density(rchisq(125+200+325, df))[c("x", "y")])
plot2 = plot_chisq2(sim_data, x, crit_val2, title="Density plot of alpha=0.05")

grid.arrange(plot1, plot2, ncol=2)
```




```
p_value <- 1 - pchisq(x, df)
paste(round(p_value,5), "Thus, Reject H0")
```

```
## [1] "0.02259 Thus, Reject H0"
```

```
paste("But")
```

```
## [1] "But"
```

```
paste("Chi-obs =", round(x,5), ", Crit_val at 0.01 =", round(crit_val1,5), ", Crit_val at 0.05 =", rounc
```

```
## [1] "Chi-obs = 7.5803 , Crit_val at 0.01 = 9.21034 , Crit_val at 0.05 = 5.99146"
```

```
((x > crit_val1) && (x > crit_val2))
```

```
## [1] FALSE
```

Thus, we conclude that:

- We cannot reject H_0 because Chi-observed is **LESS** than Chi-critical values for $\alpha = 0.01$ and $\alpha = 0.05$
- Initially, we thought, considering p-values less than 0.05 that our H_0 could be easily rejected
- However, the plotted graphs already showed us that we cannot directly reject the H_0 due to the change of Chi-observed and Chi-Critical values

Exercise 36

Young adults were asked about their satisfaction with their own character and their own family situation.

X.	Family.OK	Family.not.OK
Happy with own character	11	107
Unhappy with own character	60	94

Use a Chi-squared test to determine whether there was a connection between these two variables. (H_0)

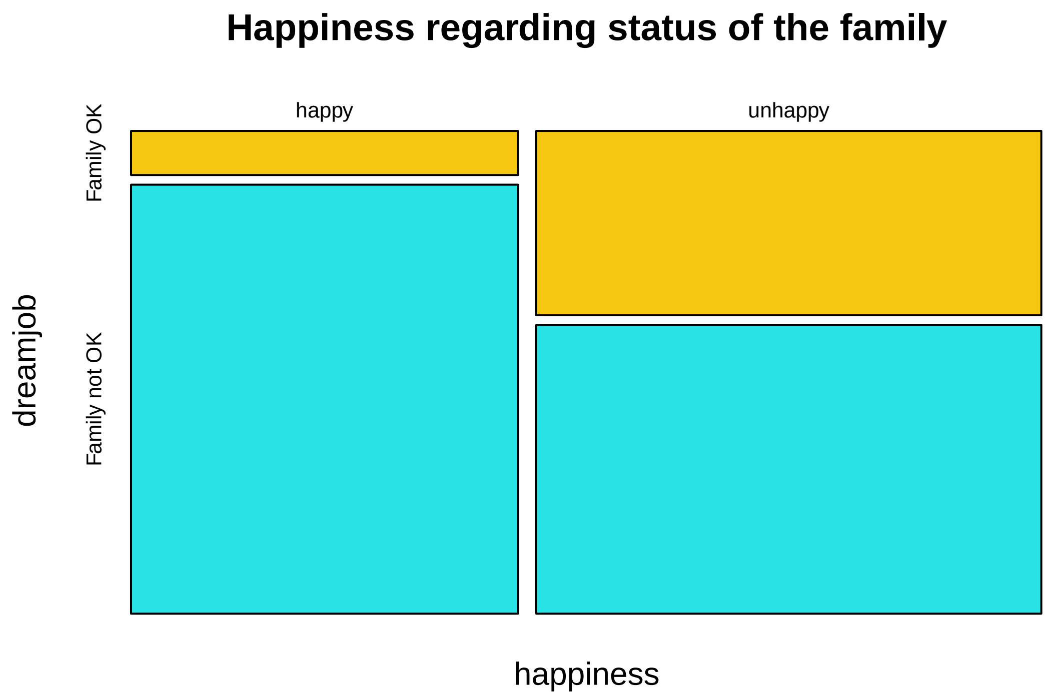
- $\alpha = 1\%$
- $\alpha = 5\%$

```r
#### Exercise 36 ####

data35 <- matrix(c(11, 60, 107, 94), ncol=2)
dimnames(data35) <- list(
  happiness = c("happy", "unhappy"),
  dreamjob = c("Family OK", "Family not OK"))
kable(addmargins(data35))
```

	Family OK	Family not OK	Sum
happy	11	107	118
unhappy	60	94	154
Sum	71	201	272

```r
mosaicplot(data35, col=c(7,5), main="Happiness regarding status of the family")
```




```
numerator <- sum(data35)*(11*94-60*107)^2
numerator
```

```
## [1] 7890446912
```

```
denominator <- sum(data35["happy", ])*sum(data35["unhappy", ])*sum(data35[ ,"Family OK"])*sum(data35[ ,
                                                                    "Family not OK"])
denominator
```

```
## [1] 259332612
```

```
chi_obs <- numerator / denominator
chi_obs
```

```
## [1] 30.42597
```

```
# alpha = 0.05
chi1= qchisq(0.95, 1, ncp=0, lower.tail = TRUE, log.p = FALSE)

# alpha = 0.01
chi2 = qchisq(0.99, 1, ncp=0, lower.tail = TRUE, log.p = FALSE)


paste("Chi-obs =", round(chi_obs,5), ", Crit_val at 0.05 =", round(chi1,5), ", Crit_val at 0.01 =", rou
```

```
## [1] "Chi-obs = 30.42597 , Crit_val at 0.05 = 3.84146 , Crit_val at 0.01 = 6.6349"
```

```
((chi_obs > chi1) && (chi_obs > chi2))
```

```
## [1] TRUE
```

We can conclude that:

- Our H_0 is rejected, our Chi-Observed is bigger than Critical values
- People with OK Families are unhappier than people with **NOT** OK Families. (Strange, but ok)

Exercise 37

70 engineers and 30 sales men applied for a certain job position. The company categorized them into two classes: 'suitable' and 'unsuitable'.

```
data = data.frame(" " = c("Engineer", "Sales man"), "suitable"=c(34, 26), "unsuitable"=c(36, 4))

kable(data)
```

X.	suitable	unsuitable
Engineer	34	36
Sales man	26	4

Use a Chi-squared test to determine whether there was a connection between these two variables. (H_0)

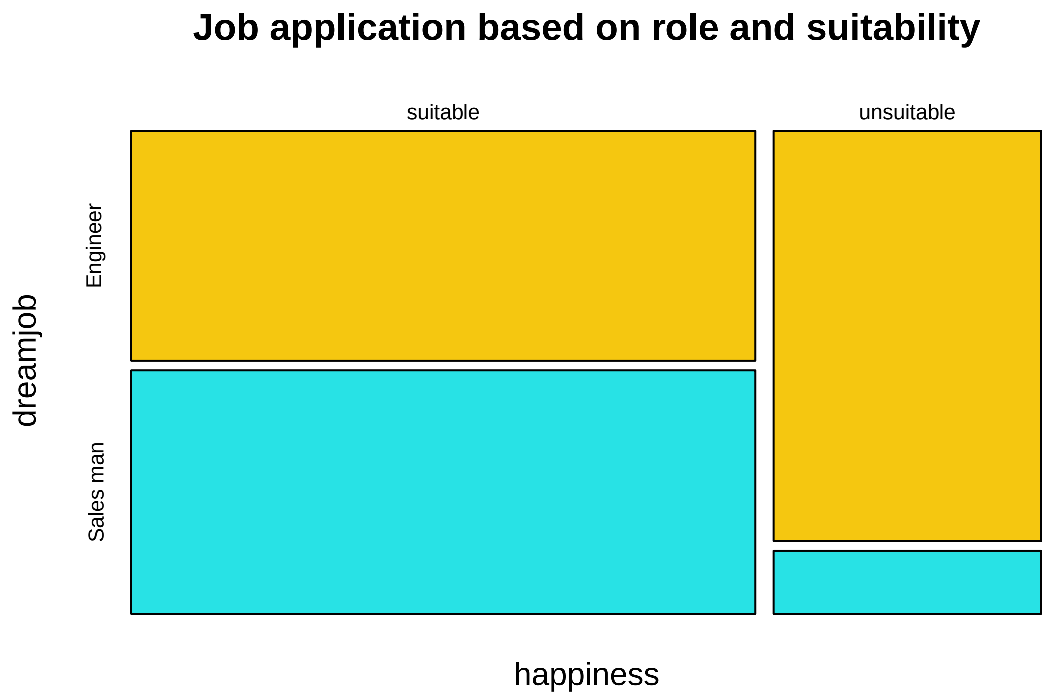
- $\alpha = 1\%$
- $\alpha = 5\$$

```
#### Exercise 37 ####

data35 <- matrix(c(34, 26, 36, 4), ncol=2)
dimnames(data35) <- list(
  happiness = c("suitable", "unsuitable"),
  dreamjob = c("Engineer", "Sales man"))
kable(addmargins(data35))
```

	Engineer	Sales man	Sum
suitable	34	36	70
unsuitable	26	4	30
Sum	60	40	100

```
mosaicplot(data35, col=c(7,5), main="Job application based on role and suitability")
```




```
numerator <- sum(data35)*(34*4-36*26)^2
numerator
```

```
## [1] 6.4e+07
```

```
denominator <- sum(data35["suitable", ])*sum(data35["unsuitable", ])*sum(data35[ ,"Engineer"])*sum(data3
denominator
```

```
## [1] 5040000
```

```
chi_obs <- numerator / denominator
chi_obs
```

```
## [1] 12.69841
```

```
# alpha = 0.05
chi1= qchisq(0.95, 1, ncp=0, lower.tail = TRUE, log.p = FALSE)

# alpha = 0.01
chi2 = qchisq(0.99, 1, ncp=0, lower.tail = TRUE, log.p = FALSE)


paste("Chi-obs =", round(chi_obs,5), "/ Crit value at 0.05 =", round(chi1,5), "/ Crit value at 0.01 =",
```

```
## [1] "Chi-obs = 12.69841 / Crit value at 0.05 = 3.84146 / Crit value at 0.01 = 6.6349"
```

```
((chi_obs > chi1) && (chi_obs > chi2))
```

```
## [1] TRUE
```

Thus, we can conclude that:

- Our H_0 is rejected, Chi-obs value being **greater** than Crit value at 0.05 and 0.01
- People with a Sales man job role are more suitable than Engineers

Exercise 38

70 engineers and 30 sales men applied for a certain job position. The company categorized them into two classes: 'suitable' and 'unsuitable'.

```
data = data.frame(" " = c("Engineer", "Sales man"), "suitable"=c(48, 20), "unsuitable"=c(22, 10))

kable(data)
```

X.	suitable	unsuitable
Engineer	48	22
Sales man	20	10

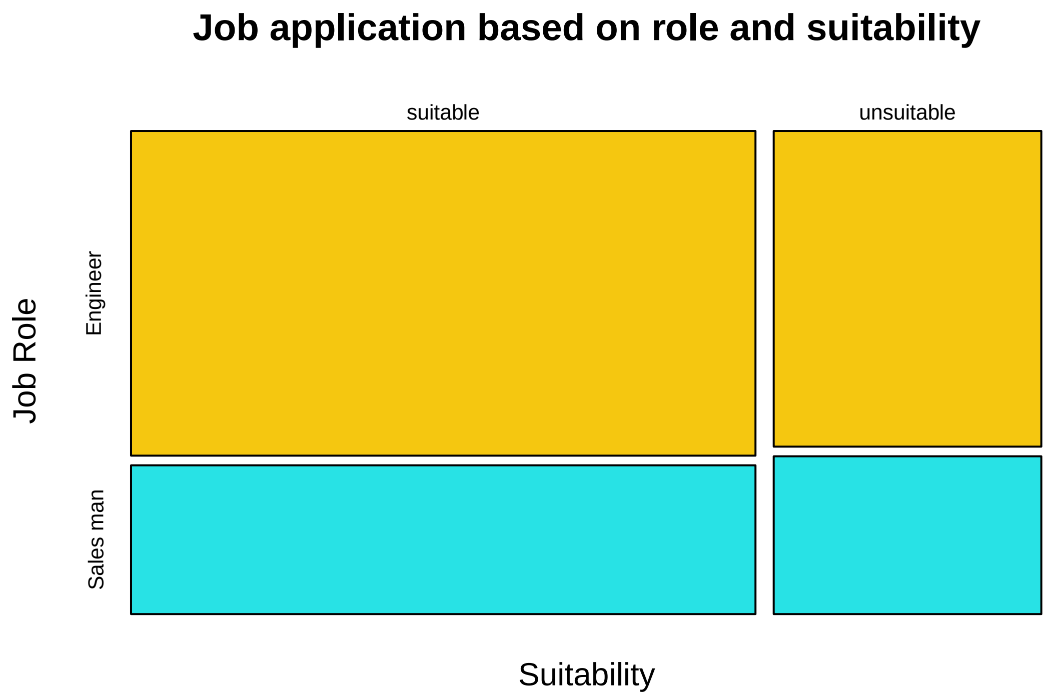
Use a Chi-squared test to determine whether there was a connection between these two variables. (H_0)

- $\alpha = 1\%$
- $\alpha = 5\$$

```
#### Exercise 38 ####
data35 <- matrix(c(48, 20, 22, 10), ncol=2)
dimnames(data35) <- list(
  happiness = c("suitable", "unsuitable"),
  dreamjob = c("Engineer", "Sales man"))
kable(addmargins(data35))
```

	Engineer	Sales man	Sum
suitable	48	22	70
unsuitable	20	10	30
Sum	68	32	100

```
mosaicplot(data35, col=c(7,5), main="Job application based on role and suitability",
           xlab="Suitability", ylab="Job Role")
```




```r
numerator <- sum(data35)*(48*10-20*22)^2
numerator
```

```
## [1] 160000
```

```r
denominator <- sum(data35["suitable", ])*sum(data35["unsuitable", ])*sum(data35[ ,"Engineer"])*sum(data3
denominator
```

```
## [1] 4569600
```

```r
chi_obs <- numerator / denominator
chi_obs
```

```
## [1] 0.03501401
```

```r
# alpha = 0.05
chi1= qchisq(0.95, 1, ncp=0, lower.tail = TRUE, log.p = FALSE)
chi1
```

```
## [1] 3.841459
```

```r
# alpha = 0.01
chi2 = qchisq(0.99, 1, ncp=0, lower.tail = TRUE, log.p = FALSE)
chi2
```

```
## [1] 6.634897
```

```r
paste("Chi-obs =", round(chi_obs,5), "/ Crit value at 0.05 =", round(chi1,5), "/ Crit value at 0.01 =",
```

```
## [1] "Chi-obs = 0.03501 / Crit value at 0.05 = 3.84146 / Crit value at 0.01 = 6.6349"
```

```r
((chi_obs > chi1) && (chi_obs > chi2))
```

```
## [1] FALSE
```

Thus, we can conclude that:

- H_0 is accepted, because Crit values are far way **bigger** than Chi-observed.
- People with a specific job don't have a connection with their job roles of Engineer or Sales man.

Exercise 39

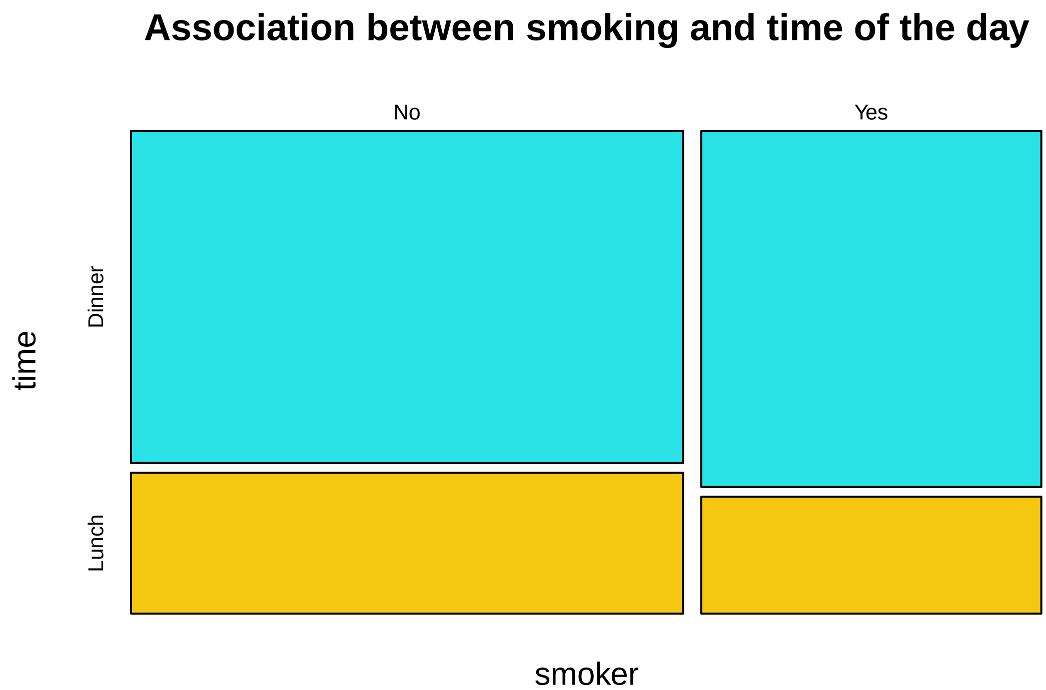
- Use the dataset 'tips'.
- Is there an association between smoking and the time of the day (lunch or dinner)? [H_0 is no association]

```r
#### Exercise 39 ####
tips<-read.csv2("tips.csv")

tally(~smoker | time, data = tips)
```

```
##       time
## smoker Dinner Lunch
##    No     106    45
##    Yes     70    23
```

```r
mosaicplot(smoker ~ time, data = tips, col=c(5,7), main="Association between smoking and time of the da
```




```r
xchisq.test(smoker ~ time, data = tips)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  x
## X-squared = 0.50537, df = 1, p-value = 0.4771
##
##   106       45
## (108.92) ( 42.08)
## [0.054]  [0.139]
## <-0.28>  < 0.45>
##
##    70       23
## ( 67.08) ( 25.92)
## [0.087]  [0.226]
## < 0.36>  <-0.57>
##
## key:
##  observed
##  (expected)
##  [contribution to X-squared]
##  <Pearson residual>
```

```r
paste("p-value is:", 0.4771, "which is greater than 0.05")
```

```
## [1] "p-value is: 0.4771 which is greater than 0.05"
```

Thus, we conclude that:

- H_0 is accepted, which says that there is ***no*** association between smoking and the time of the day (lunch or dinner)
- We can also see from the mosaic plot that there is only a small difference between lunch and dinner smokers/non-smokers

Exercise 41

- Use the dataset 'ICM'.
- Is there an association between gender and education? [H_0 is no association]

```r
#### Exercise 41 ####
ICM<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/ICM.txt",
                stringsAsFactors=F)
inspect(ICM)
```

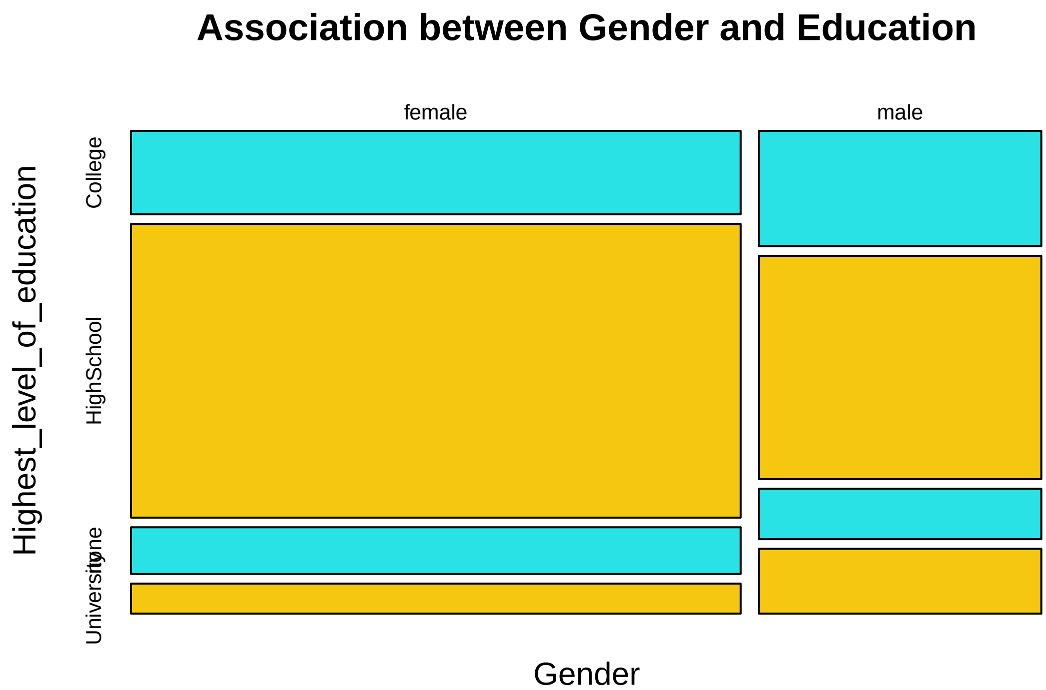
```
##
## categorical variables:
##                         name     class levels   n missing
## 1                     Gender character      2 199       0
## 2              Englishfluent character      2 199       0
## 3               Germanfluent character      2 199       0
## 4                  Transport character      4 199       0
## 5  Highest_level_of_education character      4 199       0
## 6                Do_you_smoke character      2 199       0
## 7           Socialmediahours character      4 199       0
## 8            Timewithfriends character      5 199       0
## 9                        Pet character      2 199       0
## 10                  Siblings character      2 199       0
## 11                  Children character      2 199       0
## 12        Relationshipstatus character      4 199       0
##                                 distribution
## 1  female (68.3%), male (31.7%)
## 2  yes (87.9%), no (12.1%)
## 3  no (58.3%), yes (41.7%)
## 4  Car (39.7%), PublicTransport (32.2%) ...
## 5  HighSchool (59.8%), College (20.6%) ...
## 6  No (84.9%), Yes (15.1%)
## 7  1.5-3hrs/day (44.2%) ...
## 8  2-5hrs/week (30.2%) ...
## 9  No (52.3%), Yes (47.7%)
## 10 Yes (85.4%), No (14.6%)
## 11 No (84.9%), Yes (15.1%)
## 12 Single (45.2%), Relationship (41.2%) ...
##
## quantitative variables:
```

```
##                           name   class       min       Q1     median         Q3
## 1                        ï..ID integer  1.0000000 52.500000 103.000000 155.500000
## 2                          Age integer 16.0000000 19.000000  20.000000  25.000000
## 3              Activitieshours integer  5.0000000 10.000000  10.000000  20.000000
## 4                 NegativeMood numeric  0.0000000  1.000000   1.545455   2.363636
## 5                 PositiveMood numeric  0.0000000  1.791667   2.333333   2.833333
## 6                 Mentalhealth numeric  0.1666667  2.000000   2.500000   3.000000
## 7                Socialization numeric  0.5000000  1.833333   2.666667   3.000000
## 8                     Activity numeric  0.4000000  2.200000   2.600000   3.000000
## 9                SocialSupport numeric  0.3333333  2.000000   3.000000   3.333333
## 10 Communication_open_direct numeric  1.4615385  3.538462   3.846154   4.076923
## 11                         OHS numeric  2.2413793  3.586207   4.275862   4.862069
##           max       mean         sd   n missing
## 1  209.000000 103.889447 59.9994768 199       0
## 2   87.000000  24.979899 10.9128595 199       0
## 3   50.000000  16.507538 11.4697095 199       0
## 4    4.000000   1.683693  0.8948584 194       5
## 5    4.000000   2.272959  0.8355765 196       3
## 6    4.000000   2.447811  0.7964411 198       1
## 7    4.000000   2.512090  0.7543263 193       6
## 8    4.000000   2.627411  0.6832246 197       2
## 9    4.000000   2.670017  0.8863537 199       0
## 10   4.846154   3.746066  0.5413436 176      23
## 11   5.655172   4.204801  0.7764805 181      18
```

```r
tally(~Gender | Highest_level_of_education, data = ICM)
```

```
##          Highest_level_of_education
## Gender    College HighSchool none University
##   female      25         88   14          9
##   male        16         31    7          9
```

```r
mosaicplot(Gender ~ Highest_level_of_education, data = ICM, col=c(5,7), main="Association between Gende
```




```
xchisq.test(Gender ~ Highest_level_of_education, data=ICM)
```

```
##
##  Pearson's Chi-squared test
##
## data:  x
## X-squared = 5.584, df = 3, p-value = 0.1337
##
##    25       88       14        9
## (28.02)  (81.33)  (14.35)  (12.30)
## [0.3255] [0.5476] [0.0086] [0.8861]
## <-0.571> < 0.740> <-0.093> <-0.941>
##
##    16       31        7        9
## (12.98)  (37.67)  ( 6.65)  ( 5.70)
## [0.7027] [1.1821] [0.0186] [1.9128]
## < 0.838> <-1.087> < 0.136> < 1.383>
##
## key:
##  observed
##  (expected)
##  [contribution to X-squared]
##  <Pearson residual>
```

```r
paste("p-value is:", 0.1337, "which is greater than 0.05")
```

```
## [1] "p-value is: 0.1337 which is greater than 0.05"
```

Thus, we conclude that:

- H_0 is accepted, since it is greater than confidence interval of 0.05 which says that there is ***no*** association between Gender and Education
- We can also see from the mosaic plot that there is only a small difference between Gender and Education Level

Exercise 42

- Use the dataset 'ICM'.
- Is there an association between education and smoking? [H_0 is no association]

```r
#### Exercise 42 ####
ICM<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/ICM.txt",
                stringsAsFactors=F)
inspect(ICM)
```

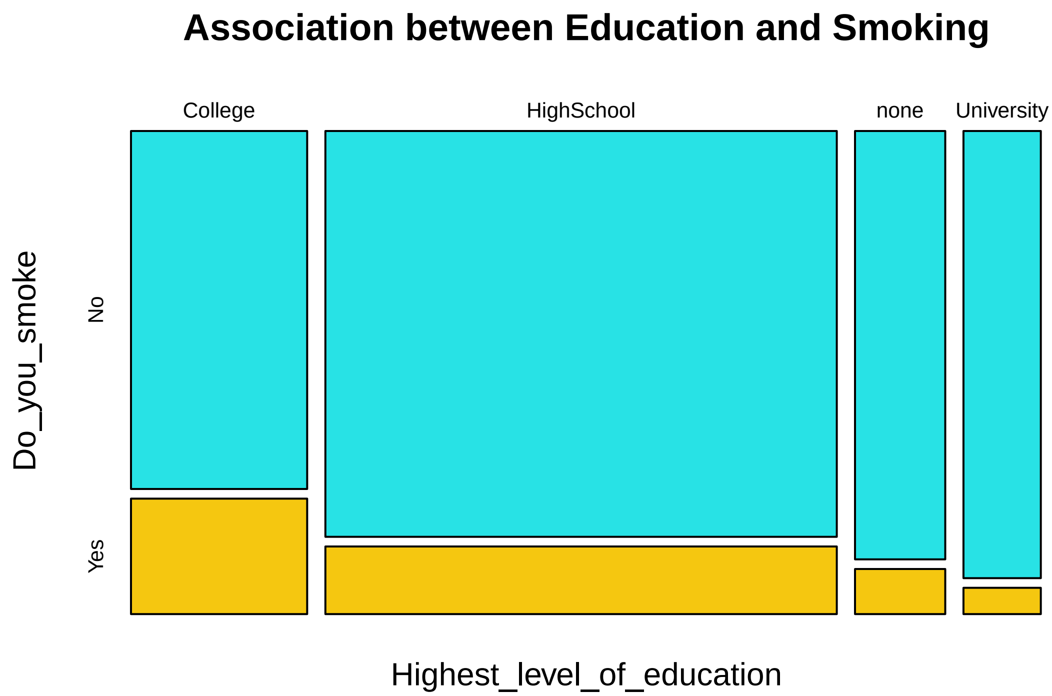
```
##
## categorical variables:
##                        name     class levels   n missing
## 1                    Gender character      2 199       0
## 2             Englishfluent character      2 199       0
## 3              Germanfluent character      2 199       0
## 4                 Transport character      4 199       0
## 5  Highest_level_of_education character      4 199       0
## 6              Do_you_smoke character      2 199       0
## 7          Socialmediahours character      4 199       0
## 8           Timewithfriends character      5 199       0
## 9                       Pet character      2 199       0
## 10                 Siblings character      2 199       0
## 11                 Children character      2 199       0
## 12       Relationshipstatus character      4 199       0
##                                    distribution
## 1  female (68.3%), male (31.7%)
## 2  yes (87.9%), no (12.1%)
## 3  no (58.3%), yes (41.7%)
## 4  Car (39.7%), PublicTransport (32.2%) ...
## 5  HighSchool (59.8%), College (20.6%) ...
## 6  No (84.9%), Yes (15.1%)
## 7  1.5-3hrs/day (44.2%) ...
## 8  2-5hrs/week (30.2%) ...
## 9  No (52.3%), Yes (47.7%)
## 10 Yes (85.4%), No (14.6%)
## 11 No (84.9%), Yes (15.1%)
## 12 Single (45.2%), Relationship (41.2%) ...
##
## quantitative variables:
```

```
##                          name   class        min        Q1     median         Q3
## 1                       ï..ID integer  1.0000000 52.500000 103.000000 155.500000
## 2                         Age integer 16.0000000 19.000000  20.000000  25.000000
## 3             Activitieshours integer  5.0000000 10.000000  10.000000  20.000000
## 4                NegativeMood numeric  0.0000000  1.000000   1.545455   2.363636
## 5                PositiveMood numeric  0.0000000  1.791667   2.333333   2.833333
## 6                Mentalhealth numeric  0.1666667  2.000000   2.500000   3.000000
## 7               Socialization numeric  0.5000000  1.833333   2.666667   3.000000
## 8                    Activity numeric  0.4000000  2.200000   2.600000   3.000000
## 9               SocialSupport numeric  0.3333333  2.000000   3.000000   3.333333
## 10 Communication_open_direct numeric  1.4615385  3.538462   3.846154   4.076923
## 11                        OHS numeric  2.2413793  3.586207   4.275862   4.862069
##           max       mean         sd   n missing
## 1  209.000000 103.889447 59.9994768 199       0
## 2   87.000000  24.979899 10.9128595 199       0
## 3   50.000000  16.507538 11.4697095 199       0
## 4    4.000000   1.683693  0.8948584 194       5
## 5    4.000000   2.272959  0.8355765 196       3
## 6    4.000000   2.447811  0.7964411 198       1
## 7    4.000000   2.512090  0.7543263 193       6
## 8    4.000000   2.627411  0.6832246 197       2
## 9    4.000000   2.670017  0.8863537 199       0
## 10   4.846154   3.746066  0.5413436 176      23
## 11   5.655172   4.204801  0.7764805 181      18
```

```
tally(~Highest_level_of_education | Do_you_smoke, data = ICM)
```

```
##                            Do_you_smoke
## Highest_level_of_education  No Yes
##                 College     31  10
##                 HighSchool 102  17
##                 none        19   2
##                 University  17   1
```

```
mosaicplot(Highest_level_of_education ~ Do_you_smoke, data = ICM, col=c(5,7), main="Association between
```




```r
xchisq.test(Highest_level_of_education ~ Do_you_smoke, data=ICM)
```

```
##
##  Pearson's Chi-squared test
##
## data:  x
## X-squared = 4.6163, df = 3, p-value = 0.2021
##
##    31       10
## ( 34.82) (   6.18)
## [0.4189] [2.3598]
## <-0.647> < 1.536>
##
##   102       17
## (101.06) ( 17.94)
## [0.0087] [0.0492]
## < 0.093> <-0.222>
##
##    19        2
## ( 17.83) (   3.17)
## [0.0762] [0.4293]
## < 0.276> <-0.655>
##
##    17        1
## ( 15.29) (   2.71)
```

```
## [0.1921] [1.0821]
## < 0.438> <-1.040>
##
## key:
##  observed
##  (expected)
##  [contribution to X-squared]
##  <Pearson residual>
```

```
paste("p-value is:", 0.2021, "which is greater than 0.05")
```

```
## [1] "p-value is: 0.2021 which is greater than 0.05"
```

Thus, we conclude that:

- H_0 is accepted, since it is greater than confidence interval of 0.05 which says that there is ***no*** association between Education and Smoking
- We can also see from the mosaic plot that there is only a small difference between Education Level and Smoking

Exercise 43

- Use the dataset 'ICM'.
- Is there an association between the transport used to get to work and the time spent with social media?

```
#### Exercise 43 ####
ICM<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/ICM.txt",
                stringsAsFactors=F)
inspect(ICM)
```

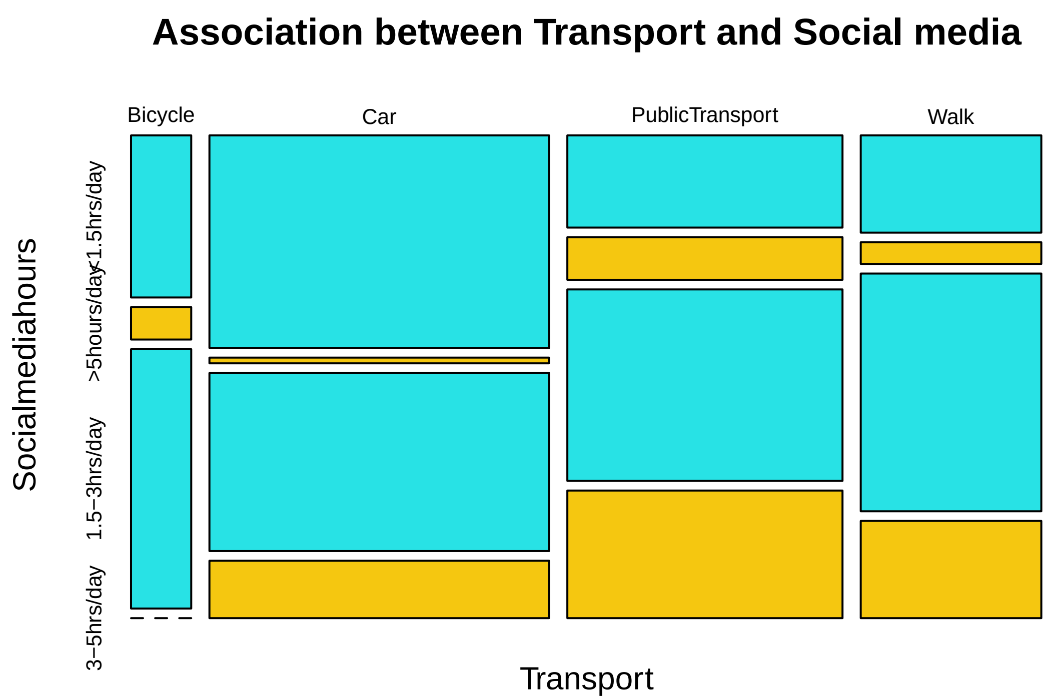
```
##
## categorical variables:
##                         name     class levels   n missing
## 1                     Gender character      2 199       0
## 2              Englishfluent character      2 199       0
## 3               Germanfluent character      2 199       0
## 4                  Transport character      4 199       0
## 5  Highest_level_of_education character      4 199       0
## 6                Do_you_smoke character      2 199       0
## 7           Socialmediahours character      4 199       0
## 8            Timewithfriends character      5 199       0
## 9                        Pet character      2 199       0
## 10                  Siblings character      2 199       0
## 11                  Children character      2 199       0
## 12        Relationshipstatus character      4 199       0
##                                  distribution
## 1  female (68.3%), male (31.7%)
## 2  yes (87.9%), no (12.1%)
## 3  no (58.3%), yes (41.7%)
## 4  Car (39.7%), PublicTransport (32.2%) ...
## 5  HighSchool (59.8%), College (20.6%) ...
```

```
## 6  No (84.9%), Yes (15.1%)
## 7  1.5-3hrs/day (44.2%) ...
## 8  2-5hrs/week (30.2%) ...
## 9  No (52.3%), Yes (47.7%)
## 10 Yes (85.4%), No (14.6%)
## 11 No (84.9%), Yes (15.1%)
## 12 Single (45.2%), Relationship (41.2%) ...
##
## quantitative variables:
##                         name   class        min        Q1     median         Q3
## 1                      ï..ID integer  1.0000000 52.500000 103.000000 155.500000
## 2                        Age integer 16.0000000 19.000000  20.000000  25.000000
## 3             Activitieshours integer  5.0000000 10.000000  10.000000  20.000000
## 4               NegativeMood numeric  0.0000000  1.000000   1.545455   2.363636
## 5               PositiveMood numeric  0.0000000  1.791667   2.333333   2.833333
## 6               Mentalhealth numeric  0.1666667  2.000000   2.500000   3.000000
## 7              Socialization numeric  0.5000000  1.833333   2.666667   3.000000
## 8                   Activity numeric  0.4000000  2.200000   2.600000   3.000000
## 9              SocialSupport numeric  0.3333333  2.000000   3.000000   3.333333
## 10 Communication_open_direct numeric  1.4615385  3.538462   3.846154   4.076923
## 11                       OHS numeric  2.2413793  3.586207   4.275862   4.862069
##           max       mean         sd   n missing
## 1  209.000000 103.889447 59.9994768 199       0
## 2   87.000000  24.979899 10.9128595 199       0
## 3   50.000000  16.507538 11.4697095 199       0
## 4    4.000000   1.683693  0.8948584 194       5
## 5    4.000000   2.272959  0.8355765 196       3
## 6    4.000000   2.447811  0.7964411 198       1
## 7    4.000000   2.512090  0.7543263 193       6
## 8    4.000000   2.627411  0.6832246 197       2
## 9    4.000000   2.670017  0.8863537 199       0
## 10   4.846154   3.746066  0.5413436 176      23
## 11   5.655172   4.204801  0.7764805 181      18
```

```
tally(~Transport | Socialmediahours, data = ICM)
```

```
##                   Socialmediahours
## Transport          <1.5hrs/day >5hours/day 1.5-3hrs/day 3-5hrs/day
##   Bicycle                    5           1            8          0
##   Car                       37           1           31         10
##   PublicTransport           13           6           27         18
##   Walk                       9           2           22          9
```

```
mosaicplot(Transport ~ Socialmediahours, data = ICM, col=c(5,7), main="Association between Transport and
```




```
xchisq.test(Transport ~ Socialmediahours, data=ICM)
```

```
##
##  Pearson's Chi-squared test
##
## data:  x
## X-squared = 23.478, df = 9, p-value = 0.005208
##
##     5        1        8        0
## ( 4.50)  ( 0.70)  ( 6.19)  ( 2.60)
## [0.0550] [0.1249] [0.5286] [2.6030]
## < 0.234> < 0.353> < 0.727> <-1.613>
##
##    37        1       31       10
## (25.41)  ( 3.97)  (34.93)  (14.69)
## [5.2897] [2.2217] [0.4432] [1.4965]
## < 2.300> <-1.491> <-0.666> <-1.223>
##
##    13        6       27       18
## (20.58)  ( 3.22)  (28.30)  (11.90)
## [2.7936] [2.4098] [0.0599] [3.1275]
## <-1.671> < 1.552> <-0.245> < 1.768>
##
##     9        2       22        9
## (13.51)  ( 2.11)  (18.57)  ( 7.81)
```

```
## [1.5042] [0.0058] [0.6324] [0.1816]
## <-1.226> <-0.076> < 0.795> < 0.426>
##
## key:
##  observed
##  (expected)
##  [contribution to X-squared]
##  <Pearson residual>
```

```
paste("p-value is:", 0.0052, "which is less than 0.05")
```

```
## [1] "p-value is: 0.0052 which is less than 0.05"
```

Thus, we conclude that:

- H_0 is rejected, since it is less than confidence interval of 0.05 which says that there is an association between Transport and Social Media
- We can also see from the mosaic plot that there is only an important difference between Transport types and Social media use per day

Exercise 45

- Use the dataset 'diet paired'.
- Is there a statistically significant difference between the body weight of the patients before the diet and after the diet? [H_0 assumes that there is no difference (identical) of body weight before and after diet]

```
#### Exercise 45 ####
diet<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/diet_pair
                stringsAsFactors=F)
head(diet, 3)
```

```
##   ï..Patient before_diet after_diet
## 1          1        86.2       83.4
## 2          2        92.7       85.8
## 3          3       102.1       98.3
```

```
inspect(diet)
```

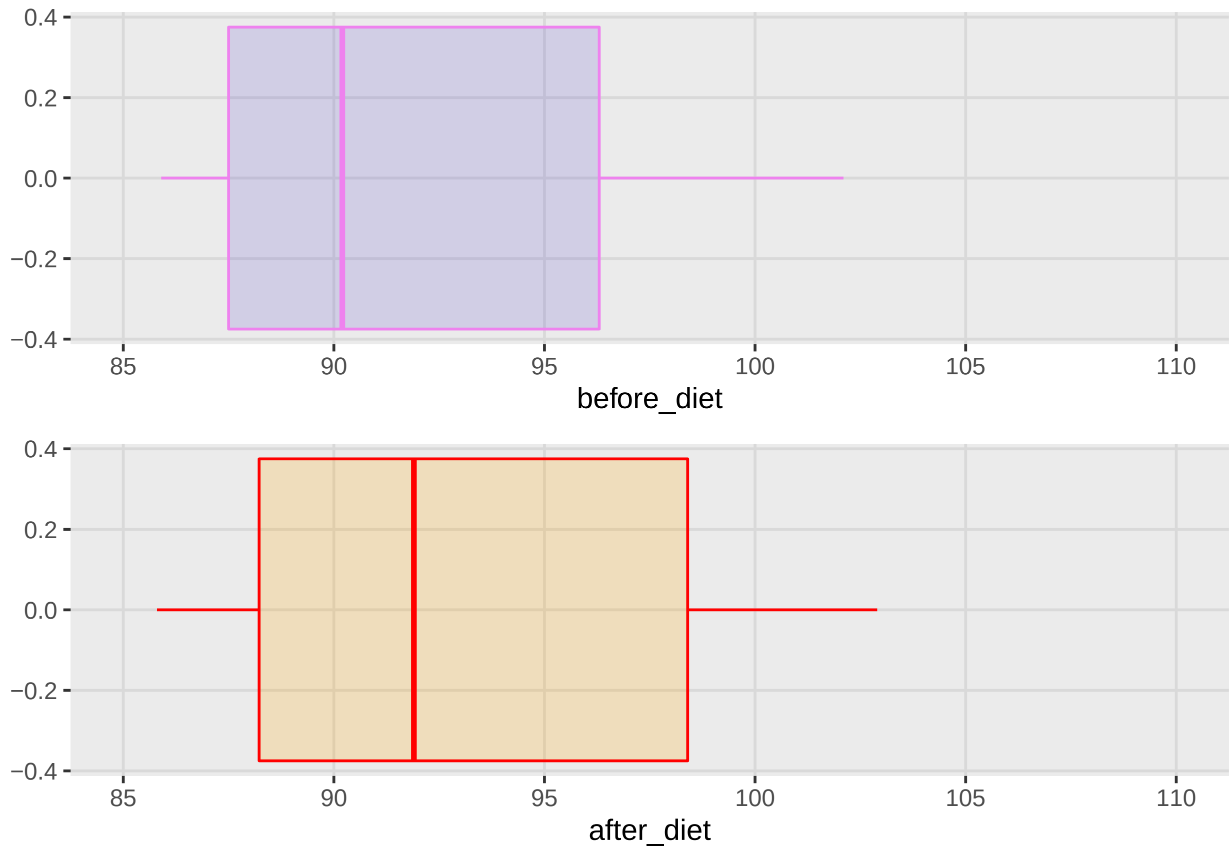
```
##
## quantitative variables:
##          name   class  min     Q1 median     Q3   max  mean       sd  n missing
## 1  ï..Patient integer  1.0  3.250   5.50  7.750  10.0  5.50 3.027650 10       0
## 2 before_diet numeric 85.9 88.100  91.45 97.575 110.2 93.90 7.823611 10       0
## 3  after_diet numeric 83.4 86.125  89.85 96.900 102.9 91.22 6.807969 10       0
```

```
bp <- ggplot(diet, aes(x=before_diet, color=before_diet)) +
  geom_boxplot(color="violet", varwidth = TRUE, fill="slateblue", alpha=0.2) +
  theme(legend.position = "none")+
  background_grid(major = "xy", minor = "none")+
```

```
  xlim(85,110)

bp2 = ggplot(diet, aes(x=after_diet, color=before_diet)) +
  geom_boxplot(color="red", varwidth = TRUE, fill="orange", alpha=0.2) +
  theme(legend.position = "none")+
  background_grid(major = "xy", minor = "none")+
  xlim(85, 110)

grid.arrange(bp, bp2, nrow=2)
```




```
wilcox.test(diet$before_diet, diet$after_diet, paired=TRUE)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  diet$before_diet and diet$after_diet
## V = 48, p-value = 0.03711
## alternative hypothesis: true location shift is not equal to 0
```

```
paste("p-value is:", 0.037, "which is less than 0.05")
```

```
## [1] "p-value is: 0.037 which is less than 0.05"
```

Thus, we conclude that:

- H_0 is rejected, since it is less than confidence interval of 0.05 which says that there is a difference of weights before and after diet
- We can also see from the box plot that there is only a change in medians

Exercise 46

- Use the dataset 'OHS 2020 paired'.
- Is there a statistically significant difference between the happiness of the students between the three time points? [H_0 assumes that there is no difference (identical) of happiness between three time points]

```
#### Exercise 46 ####
students<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/OHS_2
                     stringsAsFactors=F)
head(students, 3)
```

```
##    ï..Name OHS_1 OHS_2 OHS_3
## 1 Jennifer    NA   4.8   5.2
## 2    Tanja   4.6   4.8    NA
## 3    Heike   3.7   3.8   4.5
```

```
inspect(students)
```

```
##
## categorical variables:
##      name     class levels  n missing
## 1 ï..Name character     17 21       0
##                                distribution
## 1 Denise (9.5%), Florian (9.5%) ...
##
## quantitative variables:
##    name   class min   Q1 median  Q3 max     mean        sd  n missing
## 1 OHS_1 numeric 3.7 4.55   4.75 5.1 5.6 4.770000 0.5161599 20       1
## 2 OHS_2 numeric 3.8 4.70   4.90 5.4 5.8 4.928571 0.5514915 21       0
## 3 OHS_3 numeric 4.1 4.60   4.90 5.3 5.9 4.968421 0.4546704 19       2
```

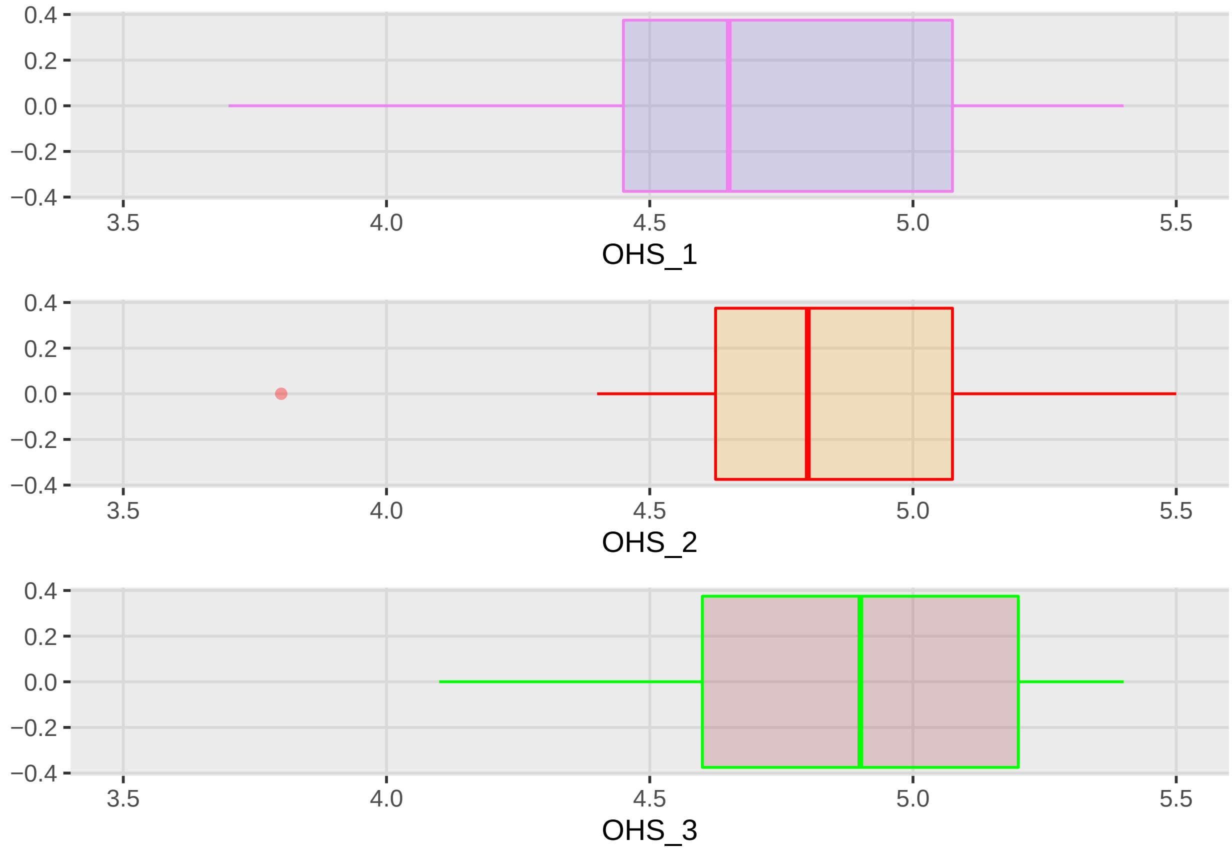
```
bp <- ggplot(students, aes(x=OHS_1, color=before_diet)) +
  geom_boxplot(color="violet", varwidth = TRUE, fill="slateblue", alpha=0.2) +
  theme(legend.position = "none")+
  background_grid(major = "xy", minor = "none")+
  xlim(3.5, 5.5)

bp2 = ggplot(students, aes(x=OHS_2, color=before_diet)) +
  geom_boxplot(color="red", varwidth = TRUE, fill="orange", alpha=0.2) +
  theme(legend.position = "none")+
  background_grid(major = "xy", minor = "none")+
  xlim(3.5, 5.5)

bp3 = ggplot(students, aes(x=OHS_3, color=before_diet)) +
  geom_boxplot(color="green", varwidth = TRUE, fill="brown", alpha=0.2) +
  theme(legend.position = "none")+
  background_grid(major = "xy", minor = "none")+
```

```
  xlim(3.5, 5.5)

grid.arrange(bp, bp2, bp3, nrow=3)
```




```
## Remove NA vals
oh1 = students$OHS_1
oh1 = na.omit(oh1)
oh1
```

```
##  [1] 4.6 3.7 4.6 4.2 4.6 5.0 4.3 5.1 5.2 4.9 4.4 4.6 5.1 4.8 4.7 5.1 3.9 5.6 5.6
## [20] 5.4
## attr(,"na.action")
## [1] 1
## attr(,"class")
## [1] "omit"
```

```
length(oh1)
```

```
## [1] 20
```

```
oh2 = students$OHS_2
oh2 = na.omit(oh2)
oh2
```

```
##  [1] 4.8 4.8 3.8 5.0 4.6 5.4 4.4 5.1 4.8 5.8 4.7 4.9 4.7 5.1 5.4 5.0 4.5 3.8 5.8
## [20] 5.6 5.5
```

```
length(oh2)
```

```
## [1] 21
```

```
oh3 = students$OHS_3
oh3 = na.omit(oh3)
oh3
```

```
##  [1] 5.2 4.5 4.9 4.6 5.3 4.1 4.9 4.6 5.9 4.7 5.3 4.6 5.0 5.4 5.1 4.9 4.4 5.6 5.4
## attr(,"na.action")
## [1]  2 20
## attr(,"class")
## [1] "omit"
```

```
length(oh3)
```

```
## [1] 19
```

```
## Normalize the vectors length

## Create random variables to remove

remove_element = function(vec){
ran1 = sample(min(vec):max(vec),1)
ran1
ran1 = match(c(ran1),vec)
ran1
vec = vec[-ran1]
return(vec)
}

oh1 = remove_element(oh1)
oh2 = remove_element(oh2)
oh2 = remove_element(oh2)
length(oh1)
```

```
## [1] 19
```

```
length(oh2)
```

```
## [1] 19
```

```
length(oh3)
```

```
## [1] 19
```

```r
# Test Oh1 & Oh2
wilcox.test(oh1, oh2, paired=TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  oh1 and oh2
## V = 40.5, p-value = 0.09245
## alternative hypothesis: true location shift is not equal to 0
```

```r
paste("p-value is:", 0.058, "which is greater than 0.05")
```

```
## [1] "p-value is: 0.058 which is greater than 0.05"
```

```r
# Plot Oh1 & Oh2
pd1 <- paired(oh1, oh2)
pl1 = plot(pd1, type = "profile") + theme_bw()

# Test Oh1 & Oh3
wilcox.test(oh1, oh3, paired=TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  oh1 and oh3
## V = 37, p-value = 0.2006
## alternative hypothesis: true location shift is not equal to 0
```

```r
paste("p-value is:", 0.018, "which is less than 0.05")
```

```
## [1] "p-value is: 0.018 which is less than 0.05"
```

```r
# Plot Oh1 & Oh3
pd2 <- paired(oh1, oh3)
pl2 = plot(pd2, type = "profile") + theme_bw()

# Test Oh2 & Oh3
wilcox.test(oh2, oh3, paired=TRUE)
```

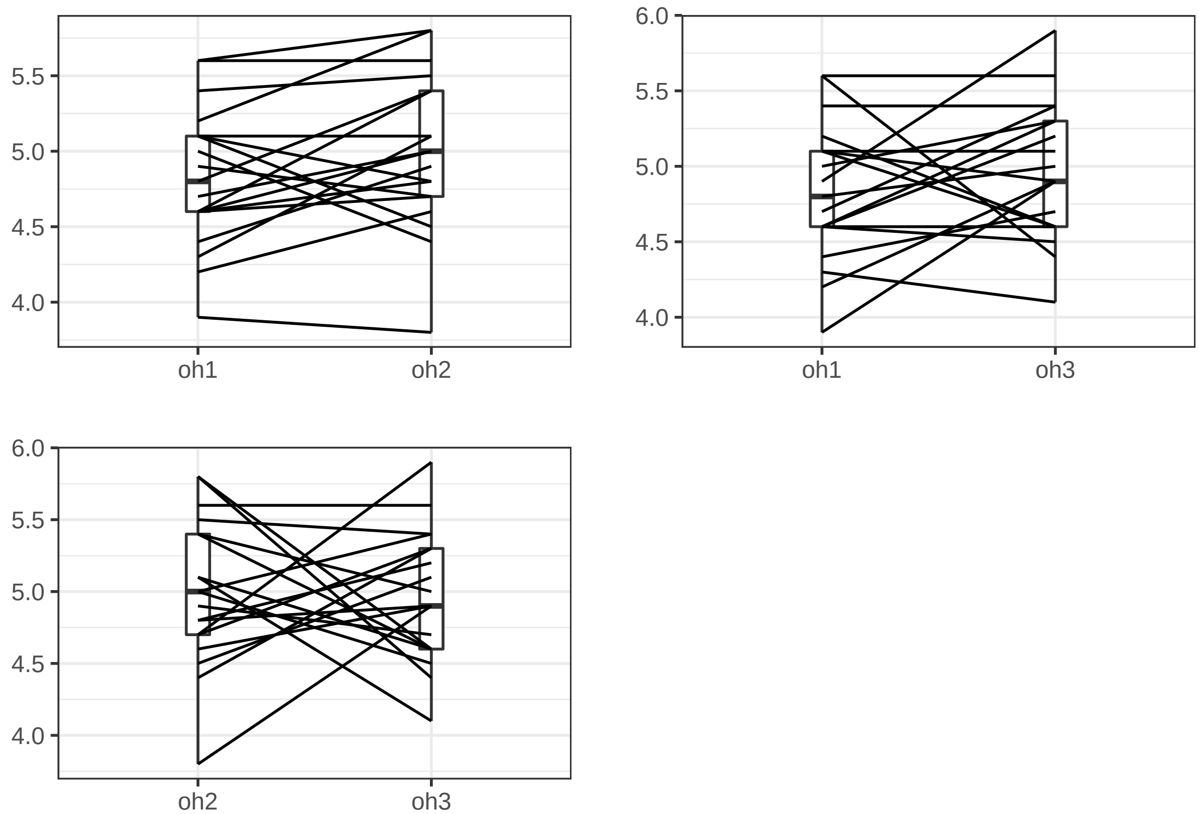
```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  oh2 and oh3
## V = 87.5, p-value = 0.9479
## alternative hypothesis: true location shift is not equal to 0
```

```r
paste("p-value is:", 0.947, "which is greater than 0.05")
```

```
## [1] "p-value is: 0.947 which is greater than 0.05"
```

```r
# Plot Oh2 & Oh3
pd3 <- paired(oh2, oh3)
pl3 = plot(pd3, type = "profile") + theme_bw()

grid.arrange(pl1, pl2, pl3, ncol=2)
```



Thus, we conclude that:

- H_0 is rejected, since for OH1 and OH3 p-value it is less than confidence interval of 0.05 which says that there is a difference of happiness between these three points
- We can also see from the box plots and the pairplots that there is a change in median and values

Exercise 49

- Use the data set 'ICM'.
- Without assuming the data to have normal distribution, decide at .05 significance level if the Communication style (open and direct) of students with siblings and students without siblings in ICM have identical data distribution. [H_0]

```r
#### Exercise 49 ####
ICM<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/ICM.txt",
                stringsAsFactors=F)

head(ICM)
```

```
##   ï..ID Gender Age Englishfluent Germanfluent       Transport
## 1    75 female  22           yes           no PublicTransport
## 2    90 female  22           yes           no PublicTransport
## 3   173 female  37           yes          yes             Car
## 4   189 female  17           yes          yes             Car
## 5   100 female  19           yes          yes            Walk
## 6   155 female  16           yes           no            Walk
##   Highest_level_of_education Do_you_smoke Socialmediahours Timewithfriends Pet
## 1                    College           No     1.5-3hrs/day     2-5hrs/week  No
## 2                    College           No     1.5-3hrs/day     2-5hrs/week  No
## 3                 University           No      <1.5hrs/day    5-10hrs/week Yes
## 4                       none           No     1.5-3hrs/day   10-20hrs/week Yes
## 5                 HighSchool           No       3-5hrs/day     >20hrs/week  No
## 6                       none           No     1.5-3hrs/day   10-20hrs/week  No
##   Siblings Children Relationshipstatus Activitieshours NegativeMood
## 1      Yes       No       Relationship              10           NA
## 2      Yes       No       Relationship              10           NA
## 3       No      Yes       Relationship              20           NA
## 4      Yes       No             Single              40     4.000000
## 5      Yes       No             Single              20     2.818182
## 6      Yes       No             Single              10     2.454545
##   PositiveMood Mentalhealth Socialization Activity SocialSupport
## 1           NA    2.6666667            NA      2.8     4.0000000
## 2           NA    2.6666667            NA      2.8     4.0000000
## 3           NA    3.5000000            NA      3.4     2.3333333
## 4    0.0000000    1.0000000           1.0      3.2     0.6666667
## 5    0.3333333    0.8333333           2.5      1.2     2.3333333
## 6    0.3333333    1.6666667           2.5      2.6     1.3333333
##   Communication_open_direct      OHS
## 1                        NA 4.586207
## 2                        NA 4.586207
## 3                  3.384615 5.103448
## 4                  3.615385 3.137931
## 5                  3.153846 2.758621
## 6                  3.461538 3.586207
```

```
inspect(ICM)
```

```
##
## categorical variables:
##                          name     class levels   n missing
## 1                      Gender character      2 199       0
## 2               Englishfluent character      2 199       0
## 3                Germanfluent character      2 199       0
## 4                   Transport character      4 199       0
## 5  Highest_level_of_education character      4 199       0
## 6                Do_you_smoke character      2 199       0
## 7            Socialmediahours character      4 199       0
## 8             Timewithfriends character      5 199       0
## 9                         Pet character      2 199       0
## 10                   Siblings character      2 199       0
## 11                   Children character      2 199       0
## 12         Relationshipstatus character      4 199       0
##                                    distribution
```

```
## 1  female (68.3%), male (31.7%)
## 2  yes (87.9%), no (12.1%)
## 3  no (58.3%), yes (41.7%)
## 4  Car (39.7%), PublicTransport (32.2%) ...
## 5  HighSchool (59.8%), College (20.6%) ...
## 6  No (84.9%), Yes (15.1%)
## 7  1.5-3hrs/day (44.2%) ...
## 8  2-5hrs/week (30.2%) ...
## 9  No (52.3%), Yes (47.7%)
## 10 Yes (85.4%), No (14.6%)
## 11 No (84.9%), Yes (15.1%)
## 12 Single (45.2%), Relationship (41.2%) ...
##
## quantitative variables:
##                         name   class        min        Q1     median         Q3
## 1                      ï..ID integer  1.0000000 52.500000 103.000000 155.500000
## 2                        Age integer 16.0000000 19.000000  20.000000  25.000000
## 3             Activitieshours integer  5.0000000 10.000000  10.000000  20.000000
## 4                NegativeMood numeric  0.0000000  1.000000   1.545455   2.363636
## 5                PositiveMood numeric  0.0000000  1.791667   2.333333   2.833333
## 6                Mentalhealth numeric  0.1666667  2.000000   2.500000   3.000000
## 7               Socialization numeric  0.5000000  1.833333   2.666667   3.000000
## 8                    Activity numeric  0.4000000  2.200000   2.600000   3.000000
## 9               SocialSupport numeric  0.3333333  2.000000   3.000000   3.333333
## 10 Communication_open_direct numeric  1.4615385  3.538462   3.846154   4.076923
## 11                        OHS numeric  2.2413793  3.586207   4.275862   4.862069
##           max       mean         sd   n missing
## 1  209.000000 103.889447 59.9994768 199       0
## 2   87.000000  24.979899 10.9128595 199       0
## 3   50.000000  16.507538 11.4697095 199       0
## 4    4.000000   1.683693  0.8948584 194       5
## 5    4.000000   2.272959  0.8355765 196       3
## 6    4.000000   2.447811  0.7964411 198       1
## 7    4.000000   2.512090  0.7543263 193       6
## 8    4.000000   2.627411  0.6832246 197       2
## 9    4.000000   2.670017  0.8863537 199       0
## 10   4.846154   3.746066  0.5413436 176      23
## 11   5.655172   4.204801  0.7764805 181      18
```

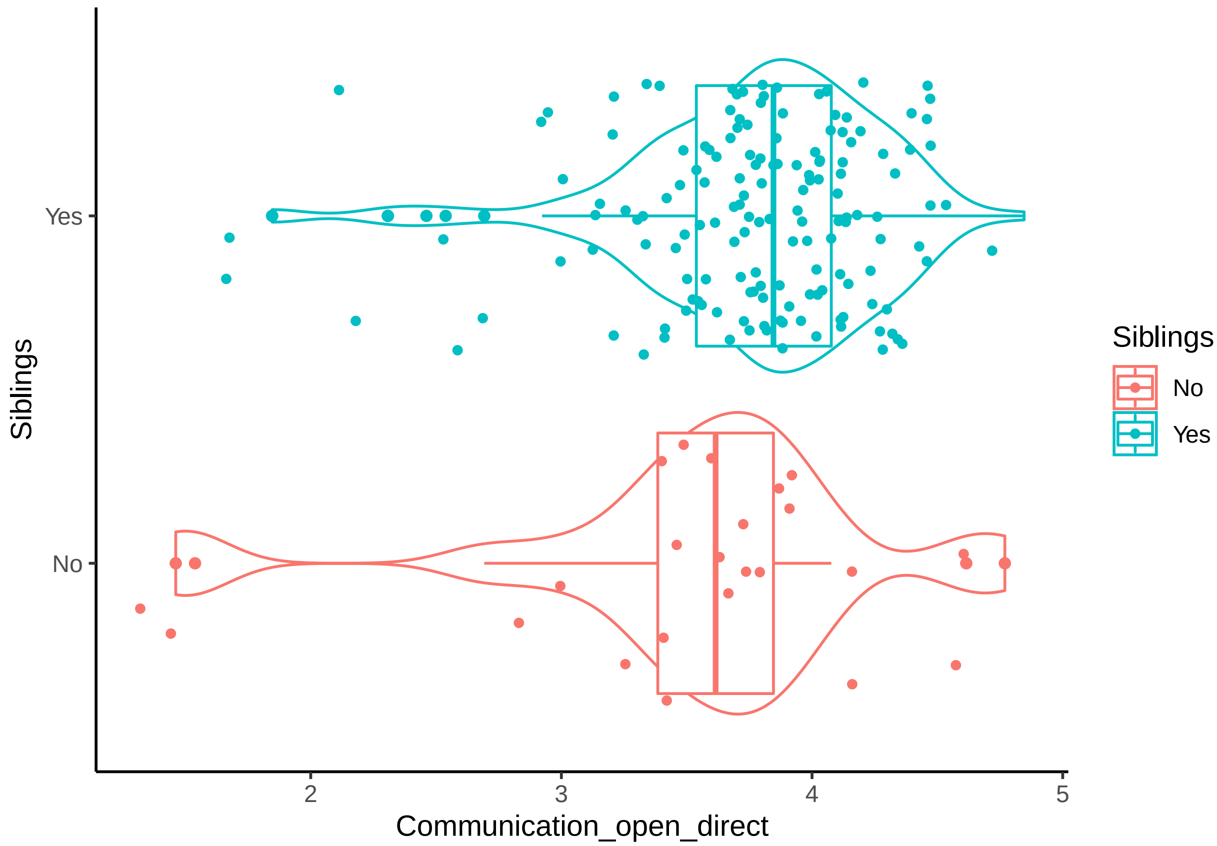
```r
wilcox.res <- wilcox.test(Communication_open_direct ~ Siblings, data=ICM)
wilcox.res
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Communication_open_direct by Siblings
## W = 1266.5, p-value = 0.03032
## alternative hypothesis: true location shift is not equal to 0
```

```r
paste("p-value is:", 0.0303, "which is less than 0.05")
```

```
## [1] "p-value is: 0.0303 which is less than 0.05"
```

```r
ggplot(ICM, aes(x=Communication_open_direct, y=Siblings, color=Siblings)) +
  geom_violin(fill="white", alpha=0.4) +
  geom_boxplot()+
  background_grid(major = "xy", minor = "none")+
  geom_jitter(shape=16, position=position_jitter(0.2))+
  scale_fill_brewer(palette="Blues") + theme_classic()
```



Thus, we conclude that:

- H_0 is rejected, since p-value is **0.0303** which is less than confidence interval of 0.05 which says that there is a difference of communication between students with siblings and without
- We can also see from the mosaic+box plots that there is a change in median and values

Exercise 50

- Use the data set 'ICM'.
- Without assuming the data to have normal distribution, decide at .05 significance level if the mental health of students with children and students without children in ICM have identical data distribution. [H_0]

```r
#### Exercise 50 ####
ICM<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/ICM.txt",
                 stringsAsFactors=F)

head(ICM)
```

```
##   ï..ID Gender Age Englishfluent Germanfluent       Transport
## 1    75 female  22           yes           no PublicTransport
## 2    90 female  22           yes           no PublicTransport
## 3   173 female  37           yes          yes             Car
## 4   189 female  17           yes          yes             Car
## 5   100 female  19           yes          yes            Walk
## 6   155 female  16           yes           no            Walk
##   Highest_level_of_education Do_you_smoke Socialmediahours Timewithfriends Pet
## 1                    College           No     1.5-3hrs/day     2-5hrs/week  No
## 2                    College           No     1.5-3hrs/day     2-5hrs/week  No
## 3                 University           No      <1.5hrs/day    5-10hrs/week Yes
## 4                       none           No     1.5-3hrs/day   10-20hrs/week Yes
## 5                 HighSchool           No       3-5hrs/day     >20hrs/week  No
## 6                       none           No     1.5-3hrs/day   10-20hrs/week  No
##   Siblings Children Relationshipstatus Activitieshours NegativeMood
## 1      Yes       No       Relationship              10           NA
## 2      Yes       No       Relationship              10           NA
## 3       No      Yes       Relationship              20           NA
## 4      Yes       No             Single              40     4.000000
## 5      Yes       No             Single              20     2.818182
## 6      Yes       No             Single              10     2.454545
##   PositiveMood Mentalhealth Socialization Activity SocialSupport
## 1           NA    2.6666667            NA      2.8     4.0000000
## 2           NA    2.6666667            NA      2.8     4.0000000
## 3           NA    3.5000000            NA      3.4     2.3333333
## 4    0.0000000    1.0000000           1.0      3.2     0.6666667
## 5    0.3333333    0.8333333           2.5      1.2     2.3333333
## 6    0.3333333    1.6666667           2.5      2.6     1.3333333
##   Communication_open_direct      OHS
## 1                        NA 4.586207
## 2                        NA 4.586207
## 3                  3.384615 5.103448
## 4                  3.615385 3.137931
## 5                  3.153846 2.758621
## 6                  3.461538 3.586207
```

```
inspect(ICM)
```

```
##
## categorical variables:
##                          name     class levels   n missing
## 1                      Gender character      2 199       0
## 2               Englishfluent character      2 199       0
## 3                Germanfluent character      2 199       0
## 4                   Transport character      4 199       0
## 5  Highest_level_of_education character      4 199       0
## 6                Do_you_smoke character      2 199       0
## 7            Socialmediahours character      4 199       0
## 8             Timewithfriends character      5 199       0
## 9                         Pet character      2 199       0
## 10                   Siblings character      2 199       0
## 11                   Children character      2 199       0
## 12         Relationshipstatus character      4 199       0
##                                    distribution
```

```
## 1  female (68.3%), male (31.7%)
## 2  yes (87.9%), no (12.1%)
## 3  no (58.3%), yes (41.7%)
## 4  Car (39.7%), PublicTransport (32.2%) ...
## 5  HighSchool (59.8%), College (20.6%) ...
## 6  No (84.9%), Yes (15.1%)
## 7  1.5-3hrs/day (44.2%) ...
## 8  2-5hrs/week (30.2%) ...
## 9  No (52.3%), Yes (47.7%)
## 10 Yes (85.4%), No (14.6%)
## 11 No (84.9%), Yes (15.1%)
## 12 Single (45.2%), Relationship (41.2%) ...
##
## quantitative variables:
##                         name   class       min        Q1     median         Q3
## 1                      ï..ID integer 1.0000000 52.500000 103.000000 155.500000
## 2                        Age integer 16.0000000 19.000000  20.000000  25.000000
## 3            Activitieshours integer 5.0000000 10.000000  10.000000  20.000000
## 4               NegativeMood numeric 0.0000000  1.000000   1.545455   2.363636
## 5               PositiveMood numeric 0.0000000  1.791667   2.333333   2.833333
## 6               Mentalhealth numeric 0.1666667  2.000000   2.500000   3.000000
## 7              Socialization numeric 0.5000000  1.833333   2.666667   3.000000
## 8                   Activity numeric 0.4000000  2.200000   2.600000   3.000000
## 9              SocialSupport numeric 0.3333333  2.000000   3.000000   3.333333
## 10 Communication_open_direct numeric 1.4615385  3.538462   3.846154   4.076923
## 11                       OHS numeric 2.2413793  3.586207   4.275862   4.862069
##           max       mean         sd   n missing
## 1  209.000000 103.889447 59.9994768 199       0
## 2   87.000000  24.979899 10.9128595 199       0
## 3   50.000000  16.507538 11.4697095 199       0
## 4    4.000000   1.683693  0.8948584 194       5
## 5    4.000000   2.272959  0.8355765 196       3
## 6    4.000000   2.447811  0.7964411 198       1
## 7    4.000000   2.512090  0.7543263 193       6
## 8    4.000000   2.627411  0.6832246 197       2
## 9    4.000000   2.670017  0.8863537 199       0
## 10   4.846154   3.746066  0.5413436 176      23
## 11   5.655172   4.204801  0.7764805 181      18
```

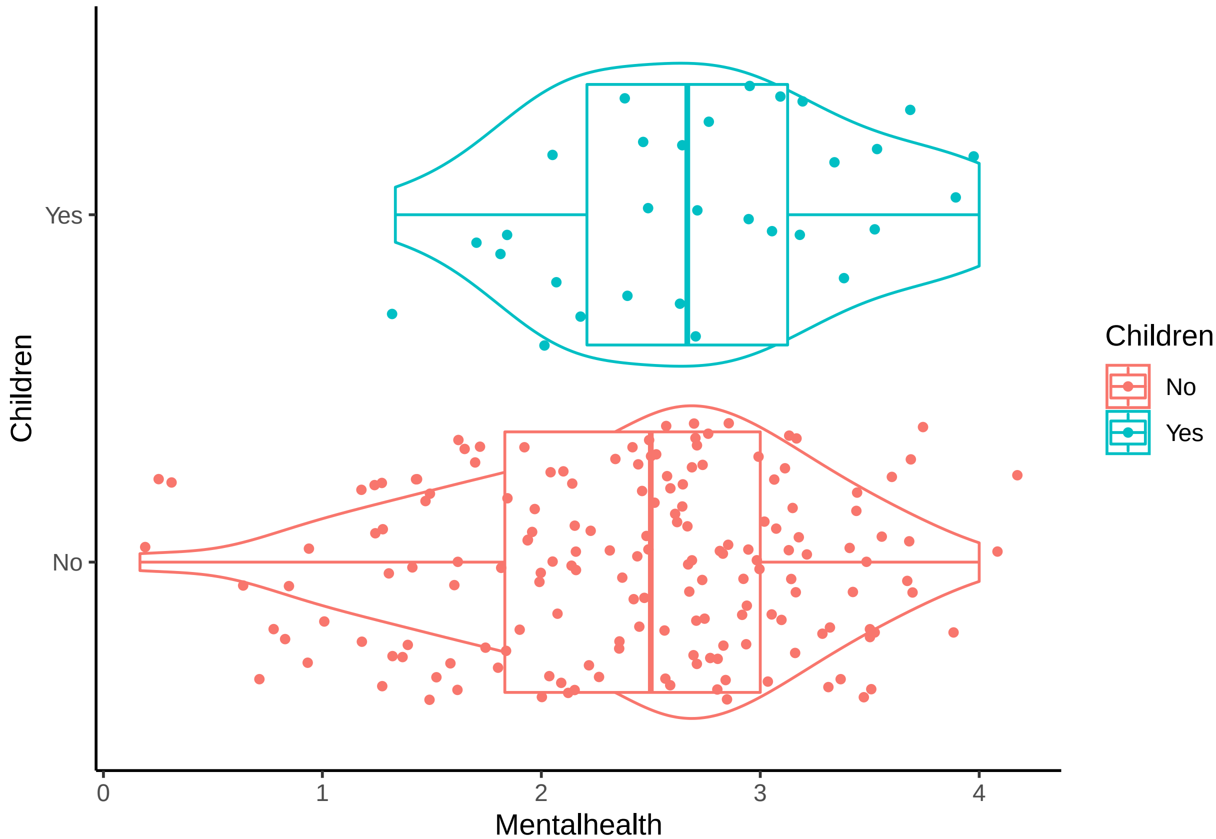
```r
wilcox.res <- wilcox.test(Mentalhealth ~ Children, data=ICM)
wilcox.res
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Mentalhealth by Children
## W = 2032.5, p-value = 0.09124
## alternative hypothesis: true location shift is not equal to 0
```

```r
paste("p-value is:", 0.0912, "which is greater than 0.05")
```

```
## [1] "p-value is: 0.0912 which is greater than 0.05"
```

```r
ggplot(ICM, aes(x=Mentalhealth, y=Children, color=Children)) +
  geom_violin(fill="white", alpha=0.4) +
  geom_boxplot()+
  background_grid(major = "xy", minor = "none")+
  geom_jitter(shape=16, position=position_jitter(0.2))+
  scale_fill_brewer(palette="Blues") + theme_classic()
```



- H_0 is accepted, since p-value is **0.0912** which is greater than confidence interval of 0.05 which says that there is no difference of mental health between students with children and without
- We can also see from the mosaic+box plots that there is not a big change in median and values