STAT 210
Applied Statistics and Data Analysis: Homework 5

Juyi Lin

Due on Oct. 9/2022

Question 1 (60 pts)

For this question, use again the data set `human` that we used in HW2. Read the file `Human_data.txt` and store this in an object called human.

a. The body mass index (BMI) is defined as a person's weight in kilograms divided by the square of height in meters. Add a column named `bmi` to the data frame with the value of this index for each subject.

b. Using the function `cut` create a new variable `bmi.fac` in `human` by dividing the subjects into four categories according to the value of `bmi` : below 20 corresponds to underweight, greater than 20 and up to 25 is normal, greater than 25 and up to 30 is overweight and above 30 is obese.

c. Build a contingency table of `Gender` and the factor you created in (b). `Gender` should correspond to the rows of your table.

d. Do a mosaic plot for the table in (c). Comment on what you observe on this graph.

e. Add a margin row and column to the table in (c) with the corresponding totals.

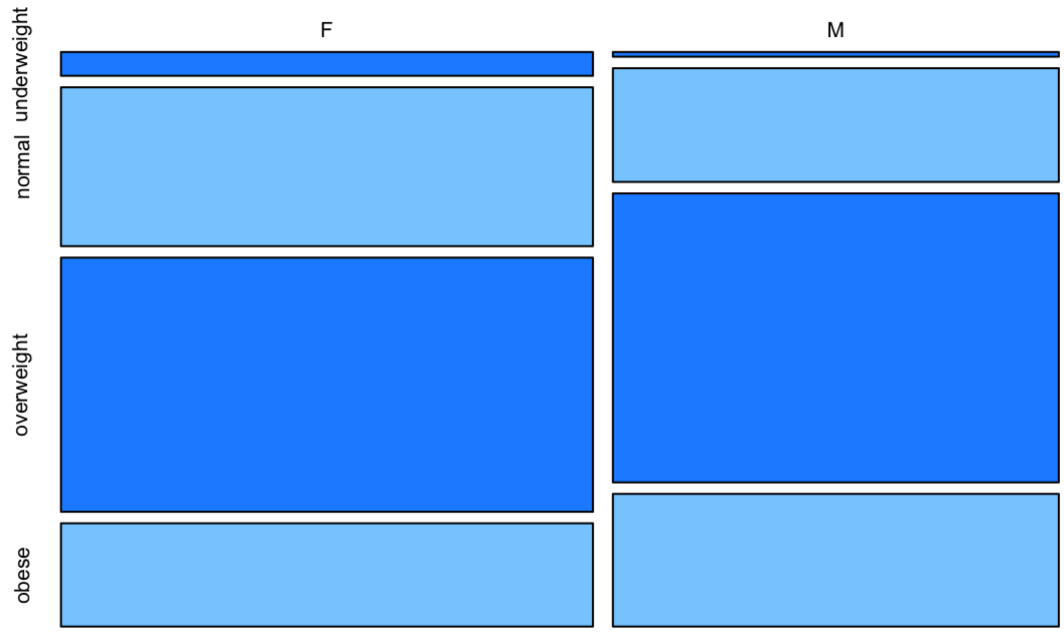
f. Build a table with the proportions with respect to the total number of cases for each gender. Comment on the results.

g. We want to test whether the distribution of the `bmi` categories that you created is the same for the different genders. What test would you use for this and why? What conditions need to be satisfied? Discuss whether they are in this example. Carry out this test and comment on your results.

```
human = read.table("Human_data.txt",header = TRUE)
human <- within(human, bmi <- human$Weight_kg/ (human$Height_cm/100)^2)
human <- within(human, bmi.fac <- cut(human$bmi, c(-Inf,20,25,30,Inf),labels = c("und
erweight","normal","overweight","obese")))
(tab1 = table(human$Gender,human$bmi.fac))
```

```
##
##     underweight normal overweight obese
##   F          12     80        128    52
##   M           2     48        122    56
```

```
mosaicplot(tab1[1:2,1:4], col = c('dodgerblue','skyblue1'),main = 'Gender with bmi.fa
c')
```




```
(human.table <- addmargins(tab1))
```

```
##
##       underweight normal overweight obese Sum
##   F            12     80        128    52 272
##   M             2     48        122    56 228
##   Sum          14    128        250   108 500
```

```
prop.table(tab1)
```

```
##
##     underweight normal overweight obese
##   F       0.024  0.160      0.256 0.104
##   M       0.004  0.096      0.244 0.112
```

```
str(tab1)
```

```
##  'table' int [1:2, 1:4] 12 2 80 48 128 122 52 56
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:2] "F" "M"
##   ..$ : chr [1:4] "underweight" "normal" "overweight" "obese"
```

```
chisq.test(tab1)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab1
## X-squared = 11.653, df = 3, p-value = 0.00867
```

I observe that there are more obese males than females and fewer underweight males than females on this graph. I observe that overweight people have most portion no matter females or males.

I use Chi-Squared Test. Because the chi-square test is a non-parametric test, and the non-parametric test does not have the assumptions of specific parameters and overall normal distribution. Condition : $p < 0.01$, so that bmi is not the same for different genders. BMI has a relation with the genders.

Question 2 (40 pts)

Are newborn babies more likely to be boys than girls?

a. In the city of Comala, out of 5235 babies born, 2705 were boys. Is this evidence that boys are more common than girls? State clearly the statistical procedure that you are using to answer this question. Describe the assumptions that you make. Are they reasonable in this case? Carry out this procedure and discuss your results.

```
tmp = c(rep(c('B'), times=2705) , rep(c('G'), times=5235-2705) )
df <- data.frame(id=c(1:5235),sex=tmp)
(prfs = xtabs( ~ sex, data=df))
```

```
## sex
##    B    G
## 2705 2530
```

```
str(prfs)
```

```
##  'xtabs' int [1:2(1d)] 2705 2530
##  - attr(*, "dimnames")=List of 1
##   ..$ sex: chr [1:2] "B" "G"
##  - attr(*, "call")= language xtabs(formula = ~sex, data = df)
```

```
chisq.test(prfs)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  prfs
## X-squared = 5.85, df = 1, p-value = 0.01558
```

```
#tab <- matrix(c(7, 5, 14, 19, 3, 2, 17, 6, 12), ncol=3, byrow=TRUE)
```

We use Chi-Squared Test.

H0: In the general, newborn babies are not more likely to be boys than girls. H1 : In the general, newborn babies are more likely to be boys than girls. p >0.01， accept H0， newborn babies are not more likely to be boys than girls. if we choose p<0.05, newborn babies are more likely to be boys than girls.

b. In the city of Macondo, out of 3765 babies born, 1905 were boys. Is there evidence that the frequency of boys is different in these two cities? Again, state clearly the statistical procedure that you are using to answer this question. Describe the assumptions that you make. Are they reasonable in this case? Carry out this procedure and discuss your results.

```
tab = matrix(c(2705, 5235-2705,1905,3765-1905),ncol=2,byrow=TRUE)
#boy = c(2705, 1905)
#girl = c(5235-2705,3765-1905)
#label = c("boys","grils")
colnames(tab) = c('boy','girl')
rownames(tab) = c('Comala','Macondo')
#df <- data.frame(labels= label, boys=boy,girls=girl  )
prfs = as.table(tab)
str(prfs)
```

```
##  'table' num [1:2, 1:2] 2705 1905 2530 1860
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:2] "Comala" "Macondo"
##   ..$ : chr [1:2] "boy" "girl"
```

```
print(prfs)
```

```
##          boy girl
## Comala  2705 2530
## Macondo 1905 1860
```

```
chisq.test(prfs)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  prfs
## X-squared = 0.9682, df = 1, p-value = 0.3251
```

The χ2 test can also be used to test for independence of categorical variables in contingency tables. We use table to produce the contingency table for these two variables.

We use Chi-Squared Test.

H0: there is no evidence that the frequency of boys is different in these two cities. H1 : there is evidence that the frequency of boys is different in these two cities.

p>0.05, we accept H0, there is no evidence that the frequency of boys is different in these two cities.