STAT 210
Applied Statistics and Data Analysis:
Homework 5
Solution

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 (a) 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.

Solution

(a) Start by reading the data and looking at the structure of the data frame

```
human <- read.table('Human_data.txt', header = T)
str(human)
```

```
## 'data.frame':    500 obs. of  10 variables:
##  $ Index         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Gender        : chr  "M" "F" "M" "F" ...
##  $ age           : int  22 33 46 24 37 31 38 38 21 31 ...
##  $ Ocupation     : chr  "Nothing" "Nothing" "Work" "student" ...
##  $ Head_size     : num  34.4 28 27 24.8 30.1 26.6 25.6 25.6 27.6 23.6 ...
##  $ Height_cm     : num  206 163 162 156 173 ...
##  $ Weight_kg     : num  105.3 71.3 94.7 56 103.3 ...
##  $ Salary        : int  0 0 19268 2034 14829 10586 11272 13048 2068 12326 ...
##  $ blood_type    : int  4 4 4 3 2 3 4 2 1 3 ...
##  $ Sugar_in_blood: num  95.2 83.5 92.7 95.8 114.1 ...
```

We now create the `bmi` variable in the `human` data frame.

```
human <- within(human, {bmi = Weight_kg/(Height_cm/100)^2})
```

(b) We use the function `cut` to create a new variable:

```
human <- within(human, {bmi.fac = cut(bmi, breaks = c(0, 20, 25, 30, 40),
                labels = c('underweight','normal','overweight','obese'))})
```

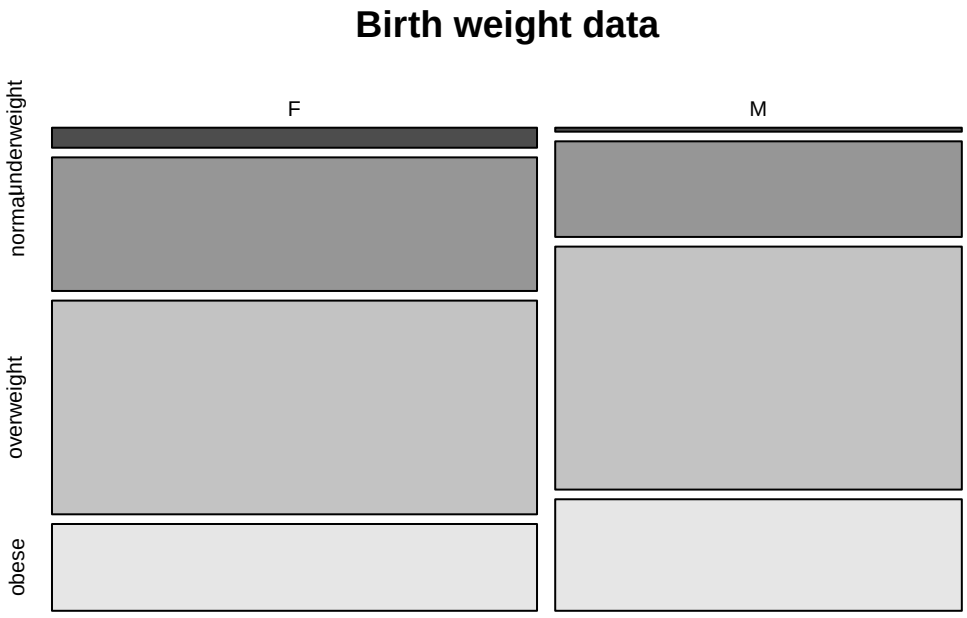
(c) Contingency table:

```
(tab1 <- table(human$Gender, human$bmi.fac))
```

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

(d) We produce the graph using `mosaicplot`

```
mosaicplot(tab1, col = T, main = 'Birth weight data')
```



In the graph we observe that the proportion of underweight and normal males is lower than for women while for overweight and obese it is the other way round: the proportion for males is higher.

(e) We can do this with the function `addmargins`:

```
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
```

(f) We can use the function `prop.table` for this

```
prop.table(tab1,1)
```

```
##
```

```
##     underweight     normal overweight      obese
##   F  0.04411765 0.29411765 0.47058824 0.19117647
##   M  0.00877193 0.21052632 0.53508772 0.24561404
```

We confirm the comment we made with the mosaic plot: the proportion of underweight and normal males is lower than for females while for overweight and obese, the proportion is higher for males.

(g) One possibility is to use the Chi-square test, which compares the expected frequencies assuming equal distributions for the genders with the observed values. The test requires that the expected values for all entries in the table be greater than or equal to five. We will verify this after running the test.

```
(csq.out <- chisq.test(tab1))
```

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

The p value in the test is small, and we have evidence in the data to reject the null hypothesis of equal distributions among the genders.

Observe that we saved the output of the test in `csq.out`. We did this to retrieve from the output the matrix of expected values in order to verify the condition that these values should all be greater than or equal to five:

```
csq.out$expected
```

```
##
##     underweight normal overweight  obese
##   F       7.616 69.632        136 58.752
##   M       6.384 58.368        114 49.248
```

We see that the condition is satisfied. It is also possible to use Fisher's exact test:

```
fisher.test(tab1)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tab1
## p-value = 0.007634
## alternative hypothesis: two.sided
```

The conclusion is the same as with the Chi-square test.

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.

(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.

Solution

(a) If p denotes the proportion of boys born, we are testing

$$H_0 : p = 0.5 \quad \text{vs.} \quad H_A : p > 0.5$$

We can either use the proportions test, which uses a normal approximation, or the exact binomial test. The first test assumes that the sample is large enough for the approximation to be valid. In this case, it is reasonable to assume that the approximation is valid, but since we can also carry out the exact test, we can compare both results.

```
prop.test(2705,5235,.5,'greater')
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  2705 out of 5235, null probability 0.5
## X-squared = 5.7834, df = 1, p-value = 0.008089
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.5052527 1.0000000
## sample estimates:
##         p
## 0.5167144
```

```
binom.test(2705,5235,.5,'greater')
```

```
##
##  Exact binomial test
##
## data:  2705 and 5235
## number of successes = 2705, number of trials = 5235, p-value = 0.008085
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.5052536 1.0000000
## sample estimates:
## probability of success
##              0.5167144
```

The p-values are practically the same, because the sample size is very large. We reject the hypothesis at the 1% level.

(b) In this case we carry out a two-sample test for proportions, which again uses the normal approximation. The sample sizes are big enough for the normal approximation to be valid.

```
prop.test(c(2705,1905), c(5235,3765))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(2705, 1905) out of c(5235, 3765)
## X-squared = 0.9682, df = 1, p-value = 0.3251
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.01042527  0.03190193
## sample estimates:
##    prop 1    prop 2
## 0.5167144 0.5059761
```

The p-value big, so we do not reject the null hypothesis of equal proportions.