Part 2: Exercise 6

Dataset

Given the following dataset test whether there is a statistical difference between males and females on hair color.

```
hair_color <- read.table('hair_color.txt', header = TRUE, sep = ',', row.names = 1)
attach(hair_color)
head(hair_color)
```

```
##        fair red medium dark jetblack
## male    592 119    849  504       36
## female  544  97    677  451       14
```

```
# head(is.na.data.frame(hair_color))
```

The data is a contingency table containing 2 groups, male and female and the observed counts of hair color ranked categorically from light to dark, *fair, red, medium, dark, jet black.*

Hypothesis

We have two independent populations grouped by gender. We would like to check whether gender plays a role in hair color of individuals. We set the hypothesis as follow. In Null hypothesis we assume that there is no effect whatsoever, which suggests that no statistical relationship and significance exists in observed data.

H0 : Two groups(gender) are independent.
Ha : Two groups(gender) are not independent.

Analysis

The Chi-square test of independence works by comparing the observed counts to the expected counts. Two events are independent if their joint probability is equal to the product of marginal probabilities. In order to understand Chi-squared test we will manually calculate the expected counts of the given dataset.

We count the totals per rows and columns, which correspond to the sum of observations grouped by gender and the sum of observations grouped by each hair color type respectively. The following *RStudio* code snippet calculates the totals and shows them to table below.

```
totals <- cbind(hair_color, sum=rowSums(hair_color))
totals <- rbind(totals, colSums(hair_color))
totals
```

```
##        fair red medium dark jetblack  sum
## male    592 119    849  504       36 2100
## female  544  97    677  451       14 1783
## 3      1136 216   1526  955       50 1136
```

Chi-squared test calculations are based on *observed counts* and *expected counts*. For example, in order to calculate the *expected count* of Gender = male and Color = medium, we need to find the joint probability, which is defined as the product of the probability of being male P_male = (2100/(1783 + 2100)) and the probability of having medium hair color P_medium = (849/(1783 + 2100)). Therefore, the expected count is the previous joint probability multiplied with the sample size. Similarly, we can compute the expected counts for all cells of the table.

Additionally, we need to check how different *observed and expected* counts are. The Chi-square test formula is:

$$x^2 = \sum \frac{(observedcount - expectedcount)^2}{expectedcount} \tag{1}$$

Value of x-squared can be between [0-1]. 0 indicates that there is absolutely no difference between the observed and the expected counts. x-squared can never be negative which means Chi-square test is an one-tailed test.

RStudio: Chi-squared test

In RStudio, *chisq.test* command performs similar process as described above to calculate p-value.

```
tests.chi <- chisq.test(hair_color)
tests.chi
```

```
##
##  Pearson's Chi-squared test
##
## data:  hair_color
## X-squared = 10.467, df = 4, p-value = 0.03325
```

We observe that the *p-value = 0.0325* which is smaller than the significance level $a = 0.05$ thus, we reject the null hypothesis and adopt the alternative. Male/Female groups are not independent and there is a statistical significance relationship between them.

References

https://www.sciencedirect.com/topics/earth-and-planetary-sciences/null-hypothesis