Statistical Data Analysis
Assignment 6

Matt Kuodis & James Zoryk.
2640428 : 2663347
Group: 36
Due Date: 10.05.2022.

Question 6.1.

Part c. In this exercise we perform two correlation tests on the variables `crime` and `expend`, namely the Kendall's test, and the Spearman's rank correlation test. The variable `expend` gives the amount of money (in $1000) that a certain state spends to combat crime, and the `crime` variable gives the crime rate of a state (number of crimes per 100000 people). We create a plot of the data, where every coordinate has the crime rate as its x-coordinate, and the expenses on crime as its y-coordinate. This gives us the following plot.

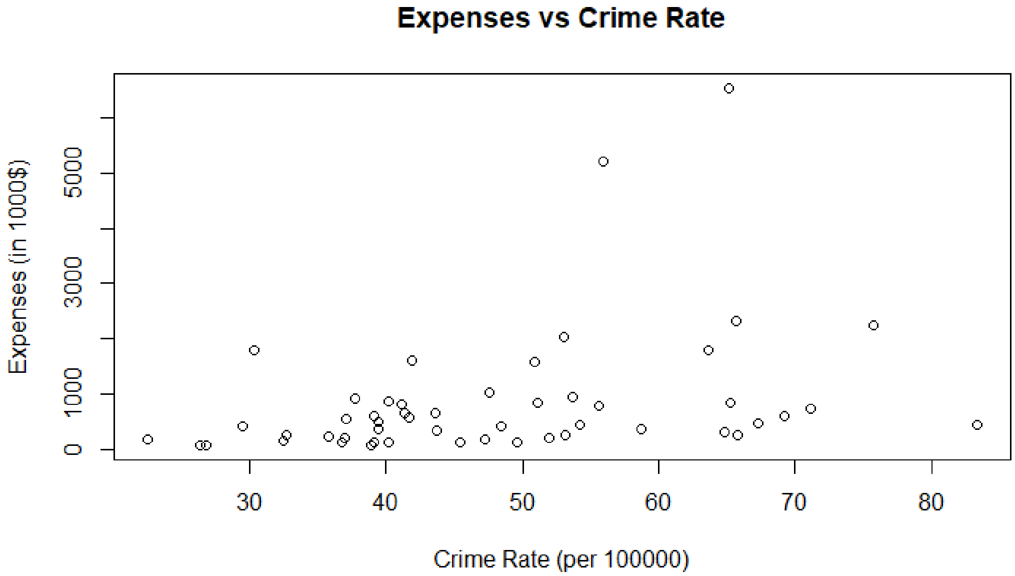

Figure 1: Expenses plotted against Crime Rate.

It is hard to estimate a correlation from the given data simply by viewing it, as some data points defy both expectation and common sense. We see that there are multiple states that spend around 1.5 million dollars to combat crime, yet these states display a crime rate (roughly equally) distributed between the 30 and 75. We also see some large outliers in the positive side of expenditure that still rank pretty highly in the crime rate, while other states spend less, and still retain a far lower crime rate.

We perform the two aforementioned tests to see whether the two data variables are correlated. In both tests we have the identical hypotheses H_0 : "Crime expenditure and crime rate are independent" against H_a : "Crime expenditure and crime rate are dependent", at significance level $\alpha = 5\%$. We use the built in `cor.test` function setting the parameter *method* to "*k*" for the Kendall test, and "*s*" for the Spearman test. The values we get are summarized in the following table.

Data	Kendall	Spearman
Test Statistic	z = 3.2733	S = 11938
p-value	0.001063	0.000781

We note that since neither of the tests has a p-value greater than 0.05, we cannot reject H_0, namely that the two data sets are independent.

Part d. In this exercise we perform a permutation test. As it is impossible (within our lifetime) to perform the computation for all 52! permutations, we perform thew test on a simulation of 2000 sample permutations. We use the same hypotheses as in the previous question, as the topic of investigation remains the correlation between crimes and crime expenditure.

We use the p-value from the built in test as our test statistic, and then we compute the left and right p-values, and compute the p-value of this test as 2× `min`(p_{left}, p_{right}) which gives us p-value = 0.002, and thus we do not reject the null hypothesis.

Part e. For this exercise we estimate the asymptotic relative efficiency of the two tests. We perform 5000 simulations in which we create a bivariate sample from the t-distribution with 6 degrees of freedom. We do this for three sample sizes, namely 48, 50, and 52. We then compute the power of the two tests for each of these sample sizes, and inspect their a.r.e, which gives us a.r.e(Kendall, Spearman) = 0.9790382 for the sample of size 48, 0.9756569 for the sample of size 50, and 0.9796226 for the sample of size 52. We see that in all cases, this value is the closest to 0.96 out of the three given values (the others being 1, and 1.4), and thus 0.96 is the closest to the true asymptotic relative efficiency. This means that Spearman's rank correlation test is more powerful for the data samples we have.

Question 6.2.

Part a. Here we have a sample sample of size $n = 2234$. The underlying distribution is multinomial, namely 4-nominal with the following parameters $2234, p_{1,1}, p_{1,2}, p_{2,1}, p_{2,2}$ such that

$$\sum_{i=1}^{2}\sum_{j=1}^{2} p_{i,j} = 1.$$

Here the hypothesis of the independence of the row and column variables is to be tested, in other words whether there is a relationship between gender and the fatalities among the infected. For this test we can state the hypotheses as

$$H_0 : p_{i,j} = p_{i,\cdot}p_{\cdot,j} \quad \forall i = 1,2,\ j = 1,2$$
$$H_1 : p_{i,j} \neq p_{i,\cdot}p_{\cdot,j} \quad \exists i = 1,2,\ j = 1,2$$

or in other words

H_0 : Gender and fatalities are independent.

H_1 : There exists some dependence between gender and fatalities.

With

$$p_{i,\cdot} = \sum_{j=1}^{2} p_{i,j}\ i = 1,2, \quad p_{\cdot,j} = \sum_{i=1}^{2}\ j = 1,2.$$

The Fisher's exact test uses the test statistic N_{11} which has the given H_0 for a hypergeometric distribution

$$P(X = k) = \frac{\binom{m}{k}\binom{n-m}{l-k}}{\binom{n}{l}}$$

with the parameters $n = 2234, l = 1097, m = 47$. Note that alternatively we could have used $n = 2234, l = 47, m = 1097$. We then conduct the Fisher's test using RStudio with the significance level $\alpha = 0.05$. From which we obtain the one-sided p-value of 0.05421. Since the p-value is larger than our significance level, namely $0.05421 > 0.05$, we can conclude that there is not enough evidence to reject the null hypothesis H_0.

Part b. Similar to part [a] but here we use the hypothesis as

H_0 : No dependence between fatalities are more common among men than women.

H_1 : fatalities are more common among men than women.

the Fisher's test in RStudio as 'fisher.test' with the parameter alternative = 'greater'. The p-value is computed as 0.02874. In this case with the significance level $\alpha = 0.05$ we have enough evidence to reject the null hypothesis h_0.

Part c. We compute the lower phyper as 0.9860 with the associated p-value as 0.02873, which confirms the evidence found in part [b], namely that we can reject the null hypothesis.

Question 6.3.

Part a. Here we have five samples (of a person given either of the drugs: placebo, chlorpromazine, Dimenhydrinate, pentobarbital-100mg and pentobarbital-150mg), each of a different size ([165, 152, 85,67, 85]). Each patient is categorized under either 'Nausea' or 'No-nausea'. Since in this analysis we are studying the class of drugs against nausea, or simply the rows against the columns, we opt to use the 'II B' model for this data set. Moreover, we do not choose the model 'II A' since we are not analysing if there is a dependence between the row and column variables. Similarly, we can rule out 'II C' since that is 'II B' but with the rows and columns swapped.

Drug	Nausea	Non - Nausea	Total
Placebo	95	70	165
Chlorpromazine	52	100	152
Dimenhydrinate	52	33	85
Pentobarbital(100mg)	35	32	67
Pentobarbital(150mg)	37	48	85
Total	271	283	554

Table 1: Incidence of nausea associated to the drug prescribed to the patient.

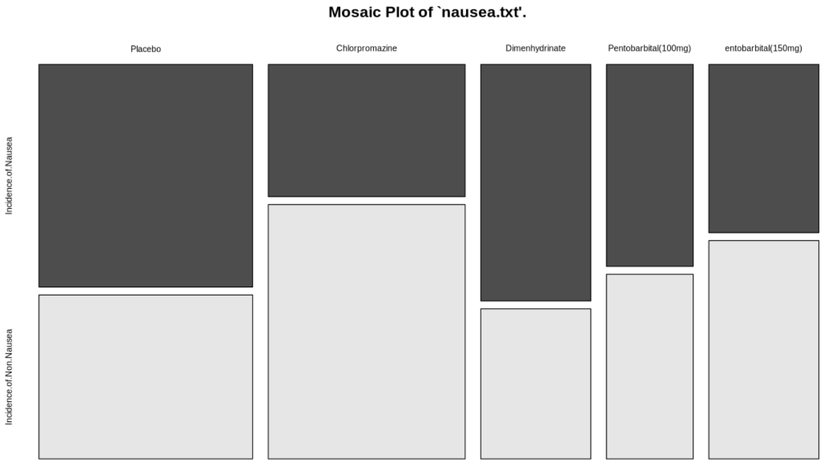

Figure 2: Mosaic plot of the Table 1

Part b. If we assume the model 'II B' is to be used at the significance level $\alpha = 0.05\%$ then the null hypothesis of homogeneity amongst the 5 drug samples can be formulated as

$$H_0 : p_{i,j} \equiv p_j \ \forall i. \quad \text{(The 5 drugs are homogeneous.)}$$
$$H_1 : p_{i,j} \not\equiv p_i \ \exists i \quad \text{(The drugs are non-homogeneous).}$$

In other words H_0 states that the incidence of categories ('nausea' and 'non-nausea') is the same, no matter which drug has been administered to the patients. The test statistic used for this test is

$$X_B^2 = \sum_{i=1}^{r}\sum_{i=1}^{c} \frac{(N_{i,j} - N_i \cdot p_{i,j})^2}{N_i \cdot p_{i,j}},$$

with r being the total number of rows (drugs) and c being the total number of columns (categories).

The 'rule of thumb' for the chi-square test: $\mathbf{E}(N_{i,j} \geq 1 \forall i, j$ and $\mathbf{E}(N_{i,j} \geq 5$ for at least 80% of the cells. We compute the values as $\mathbf{E}(N_{i,1} = n_t \cdot \frac{r}{n_t} \frac{c}{n_t}$, which can be done in RStudio. The yields are seen in Table 2. Here, we see that all the values in the Table 2 are greater than 1 and all are greater than 5. Hence it follows that more than 80% are greater than 5. Thus the 'Rule of thumb' holds here.

Conducting the Chi-squared test we get the following values, for the test statistic $X^2 = 24.502$ with the p-value of $6.334 \cdot 10^{-5}$, noting that degrees of freedom equal $df = 4$. Since the p-value is less than our significance level α, there is evidence to reject the null hypothesis. Thus, there is evidence that the samples are non-homogeneous. In other words, there is a relationship between the drug given and the incidence of nausea.

Drug	Nausea	Non - Nausea
Placebo	80.713	84.287
Chlorpromazine	74.353	77.646
Dimenhydrinate	41.579	43.420
Pentobarbital(100mg)	32.774	34.225
Pentobarbital(150mg)	41.579	43.420

Table 2: Chi-squared expected values.

Part c. We conduct the Chi-squared tested with simulated p-values based on 2000 replicates by using the command 'chisq.test' in RStudio with the option 'simulate.p.value' to obtain the p-value as 0.0004998.

Again, in this test the p-value is less than our significance level meaning that there is evidence to reject the null hypothesis. Since we conducted two tests on the data which both returned the same conclusion we can assume that the p-value computed in part [b] is reliable.

Part d. We use RStudio to compute the table of residuals which is given in Table 3 and also the standardised residuals (Table 4) for the Chi-squared test conducted in part [b].

Residuals

Drug	Nausea	Non - Nausea
Placebo	1.590	-1.556
Chlorpromazine	-2.592	2.536
Dimenhydrinate	1.616	-1.581
Pentobarbital(100mg)	0.388	-0.380
Pentobarbital(150mg)	-0.710	0.694

Table 3: Residuals of the Chi-squared test preformed in part [b].

From Table 3 we observe that the largest contribute occurs in cells ('Placebo','Nausea'), ('Chlorpromazine', Non-nausea') and ('Dimenhydrinate','Nausea'), which contribute more often than expected under the null hypothesis. The cells ('Placebo','Non-nausea'), ('Chlorpromazine', 'Nausea'), and ('Dimenhydrinate','Non-nausea') contribute less than expected under the null hypothesis.

Stanardised Residuals

Drug	Nausea	Non - Nausea
Placebo	2.665	-2.665
Chlorpromazine	-4.258	4.258
Dimenhydrinate	2.457	-2.457
Pentobarbital(100mg)	0.580	-0.580
Pentobarbital(150mg)	-1.080	1.080

Table 4: Stanardised Residuals of the Chi-squared test.

Now consider the standardised residuals in Table 4 compared to $\Phi^{-1}(\alpha/2)$ and $\Phi^{-1}(1-\alpha/2)$ which for $\alpha = 0.05$ is ± 1.96. The same cell mentioned above contribute more or less than expected under the null hypothesis. This leads us to reject the null hypothesis, as previously stated.

Part e. Here we consider the same test in part [b], however this time we shall conduct it using a bootstrap procedure. In this case our test statistic is given as $T =$ *(The largest of the absolute values of the contributions).* We ran the bootstrap with $B = 4000$ and with the function 'maxcontributionscat' to obtained $t \approx 2.592$. Furthermore we carry of the bootstrap test to get the p-value of $2 \cdot 10^{-4}$. Since this is less than our significance level $\alpha = 0.05$, there is evidence to reject the null hypothesis.

Part f. Here we conduct a one sided Fisher's exact test in order to see if the drug 'Chlorpromazine' has more of an effect than the 'Placebo'. We set our significance level for the test as $\alpha = 0.05\%$ and set our null and alternative hypotheses as:

H_0 :There is no dependencies between the prescribed drug and the incidence of nausea.

H_1 :Incidence of nausea is more common to those prescribed 'Placebo' than 'Chlorpromazine'.

Our test statistic for these cases is the cell $N_{1,1}$ which under H_0 is distributed according to the hypergeometric distribution with the parameter $(n, N_{1,\cdot}, N_{\cdot,1})$. So it follows that in this case $N_{1,1} = 95$. Using this we compute the p-value as $2.312 \cdot 10^{-5}$, which is smaller than our significance level, hence there is evidence to reject the null hypothesis and consider the alternative hypothesis. In other words, there is a larger incidence of nausea among patients given the Placebo than those given Chlorpromazine.

R code

```
#Assignment 6

# set working dir
setwd("~/Maths/SDA/Assignment_6")

# exercise 6.1
data_crimes <- read.table("expensescrime.txt", header = TRUE)

# part a
pairs(matrix(as.numeric(unlist(data_crimes[1:51,2:7])),nrow=51))
# part b
rate_expend = t(data_crimes['expend'])
rate_crime = t(data_crimes['crime']) / 100
plot(rate_crime,rate_expend, main = "Expenditure vs Crime", ylab = "
    Expenditure (in $1000)",
     xlab = "Crime rate (per 100000)")
plot(t(data_crimes['crime']), t(data_crimes['expend']))

#part c
cor.test(rate_crime,rate_expend, method='k')
cor.test(rate_crime,rate_expend, method='s')

# part d
names(rate_crime)<- NULL

test_stat = cor.test(rate_crime,rate_expend, method='s')[[4]]
n = 2000
permutation_value = numeric(n)

for(i in 1:n) {

  permutation_value[i] = cor.test(sample(rate_crime, replace=TRUE), sample(
      rate_expend, replace=TRUE), method="s")[[4]]

}

p_left = sum(permutation_value <= test_stat)/n
p_right = sum(permutation_value > test_stat)/n
p_final = 2*min(p_left, p_right)

# part e
install.packages("mvtnorm")
library("mvtnorm")

#function follows setup from slides in lecture 8
aresimulation = function(B, n, deg_free){

  pval_spear = pval_ken = numeric(B)

  for(i in 1:B){
    x = rmvt(n, sigma=matrix(c(1, 0.3, 0.3,1), 2,2), df=deg_free)
    pval_spear[i] = cor.test(x[,1], x[,2], method = "s")[[3]]
    pval_ken[i] = cor.test(x[,1], x[,2], method = "k")[[3]]
  }

  power_spear = sum(pval_spear<0.05)/B
  power_ken = sum(pval_ken<0.05)/B
  rbind(c("Spearman", "Kendall"), c(power_spear, power_ken))

}

sample_size_48 = aresimulation(5000, 48, 6)
```

```
sample_size_50 = aresimulation(5000, 50, 6)
sample_size_52 = aresimulation(5000, 52, 6)

#a.r.e. for different sample sizes
are1 = as.numeric(sample_size_48[2])/as.numeric(sample_size_48[4])
are2 = as.numeric(sample_size_50[2])/as.numeric(sample_size_50[4])
are3 = as.numeric(sample_size_52[2])/as.numeric(sample_size_52[4])

################################################################################
# exercise 6.2
# create the table
men_deaths = 30;men_recover = 1067; men_total = men_deaths + men_recover
women_deaths = 17;women_recover = 1120;women_total = women_deaths + women_
    recover
deaths_total = men_deaths + women_deaths; recover_total = men_recover + women_
    recover
people_total = men_total + women_total
table = matrix(c(men_deaths,women_deaths,men_recover,women_recover),nrow=2,
    ncol=2)


#part a
#H_0: No dependencies between gender and fatalities
#H_1: Dependency between gender and fatalities
fisher.test(table)

#part b

# based on stat men_deaths, and hypothetical unknown men_deaths2
# H_0: No dependency that fatalities are more common among men than women
# H_1: Dependency between gender and fatalities
# H0: men die leq // men_deaths <= men_deaths2
# H1: men die nore // men_deaths > men_deaths2
fisher.test(table, alternative = "greater")
# can reject H0

#part c
pl = phyper(men_deaths, men_total, women_total, deaths_total)
pr = 1 - phyper(men_deaths - 1, men_total, women_total, deaths_total)
c(pl,pr)
# it looks like we can reject H0 from both sides but i cant find the pvalue
    from that (it is in fact the same as before)

################################################################################
# exercise 6.3
nausea <- read.table("nausea.txt")

nausea_table<- matrix(NA, nrow=6, ncol=3)
rownames(nausea_table) = c('Placebo', 'Chlorpromazine', 'Dimenhydrinate', '
    Pentobarbital(100mg)', 'entobarbital(150mg)','Total')
colnames(nausea_table) = c('Incidence.of.Nausea','Incidence.of.Non.Nausea','
    Number.of.Patients')


for( i in 1:5){ nausea_table[i,] <- c(nausea[i,2], nausea[i,1]-nausea[i,2],
    nausea[i,1])}
nausea_table[6,] <- c(sum(nausea_table[1:5,1]),sum(nausea_table[1:5,2]),sum(
    nausea_table[1:5,3]) )

nausea_table
mosaicplot(nausea_table[1:5,1:2], main='Mosaic Plot of ‘nausea.txt\'. ', color
    =TRUE)
# part b
```

```r
chisq.test(nausea_table[1:5,1:2])$expected
chisq.test(nausea_table[1:5,1:2])

#part c
chisq.test(nausea_table[1:5,1:2], simulate.p.value = TRUE)

# part d
chisq.test(nausea_table[1:5,1:2])$residuals
round(chisq.test(nausea_table[1:5,1:2])$stdres,3)

# part e
source('functions_Ch7.txt')
B = 5000
t = maxcontributionscat(nausea_table[1:5,1:2])
bs_t = bootstrapcat(nausea_table[1:5,1:2], B, maxcontributionscat)
t_star = mean(bs_t >=t)

err = (chisq.test(nausea_table[1:5,1:2])$p.value - t_star )/t_star

#part f

fisher.test(nausea_table[1:2,1:2], alternative = 'g')
```

code/sda_hw6_main_code.R