SDA- Assignment 6

Jose Chacon (2699643)

6.1: Correlation between Crime and Expenses

In the first exercise, we are given a table with information about state expenditure on fighting criminality. From all the variables in the table, we are only going to focus on the crime rate and the state expenditures. Furthermore, we only consider the states with a population of at least 3,000,000. We have performed the following tests/studies to understand the relationship between these two factors.

Note: for every statistical test that we realized, we will state the null and alternative hypothesis, the test statistic and the distribution under the null hypothesis.

c) Rank Correlation Tests: Kendall's and Spearman's

The first test that we performed on the data is the Kendall's rank correlation test. We tested at significance level $\alpha = \%5$ if the two variables: expenditures and crime rate are dependent or independent. We obtained a p-value of 0.037 (and tau=0.28), which is less than our alpha. Therefore, we reject the null hypothesis and we conclude that the columns/variables are dependent according to the Kendall's correlation test.

Information about performed statistical test: Kendall's Rank Correlation Test

1. Null Hypothesis: The crime rate and expenditure rate are independent.
2. Alternative Hypothesis: The crime rate and expenditure are dependent.
3. Description of Test Statistic: Kendall's Tau, which is a measure of association between two variables (in this case: crime and expend with a value of 0.28) that takes values from -1 to 1.
4. Formula for the Test Statistic: $\tau = \frac{\sum\sum_{i\neq j} sgn(R_i{=}R_j)sgn(S_i{,}S_j)}{n(n-1)} = \frac{4N_\tau}{n(n-1)} - 1$
5. Distribution Under the Null: for this sample size it is usually a t-distribution with n-2 degrees (for larger samples, it is normally distributed).

Second, we performed the same test but now with the Spearman's method. The p-value was 0.032 (and rho =0.41), which again is less than the significance level. We again reject the null hypothesis. This is in line with the first test.

Information about performed statistical test: Spearmans's Rank Correlation Test

1. Null Hypothesis: The crime rate and expenditure rate are independent.
2. Alternative Hypothesis: The crime rate and expenditure are dependent.
3. Test Statistic: Rho, which is the measure of association between two variables. In this case, rho = 0.41.
4. Formula of the test Statistic: $r_s = \frac{(\sum_{i=1}^{n}(R_i-\overline{R})(S_i-\overline{S}))}{(\sum_{i=1}^{n}(R_i-\overline{R})^2\sum_{i=1}^{n}(S_i-\overline{S})^2)^{\frac{1}{2}}}$
5. Distribution Under the Null: for this sample size it is usually a t-distribution with n-2 degrees (for larger samples, it is normally distributed).

d) Permutation Test based on Spearman's rank correlation coefficient

We again test for dependency between the crime and the expenditures; however, now we use a permutation test based on Spearman's rank correlation coefficient (also at level $\alpha = \%5$).

Information about performed statistical test: Permutation Test

1. Null Hypothesis: There is no correlation between crime and expenditure.
2. Alternative Hypothesis: There is correlation between crime and expenditure.
3. Test Statistic: the observed Spearman's rank correlation coefficient between crime and expend. (In this case, t equals 2164)
4. Distribution Under the Null: it follows a normal distribution

Instead of computing all 51! possible permutations, we use the bootstrap approximation method to perform the test. The permutations test goes as following:

1. The two samples/columns are combined.
2. Two equally sized samples are re-sampled from the combined sample.
3. The Spearman's rank correlation coefficient of these new samples is computed and added it to the bootstrap sample.
4. After iterating these first steps B times, we compute the left p-value and the right p-value by checking how many values in the bootstrap sample are larger (smaller) than the Spearman's correlation coefficient of the original two variables.
5. The p-value is computed with the following formula: p-value = 2*min(pl,pr)

After conducting this procedure, we have computed a p-value of 0.034. Therefore, just as in the previous two tests, we rejected the null hypothesis.

e) Simulate the asymptotic relative efficiency

Lastly, we would like to approximate the value of the asymptotic relative efficiency (A.R.E.) of Kendall's rank correlation test with respect to Spearman's rank correlation test, when the data come from a bi-variate t-distribution.

We created a function that ultimately computes the A.R.E of Kendall's rank correlation test in respect to Spearman's rank correlation test. The functions accepts as input two integers n (number of observations in the random sample) and B (number of simulations). The functions is composed mainly of the following steps:

1. we create a random sample with observations that come from a bi-variate t-distribution with six degrees of freedom and a scale matrix:
$$\begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}$$
.
2. Second, we use the correlation test with the Kendall's and Spearman's method to compute the p-value and save them in corresponding vectors
3. Steps 1 and 2 were repeated B (B¿=1000) times.
4. The average of the observation with a value lower than the significance level $\alpha = 0.05$ is computed. This values are saved in variables called powerK and powerSp, respectively.
5. We used the formula of the asymptotic relative efficiency:
$$\text{ARE} = \left(\frac{\text{powerK}}{\text{powerSp}}\right)^2$$
6. return the A.R.E.

After we successfully simulated the A.R.E of Kendall's test with respect to Spearman's test, we obtained the value 1.27. This means that the Kendall's test is 1.27 more efficient. Equivalently, it means that Kendall's needs 27% less sample size to achieve the same power as the test based on Spearman's.

6.2: Virus Infection

In this exercise, we are given a table with number of deaths and recoveries of men and women from a virus infection. We would like to understand if there is a relationship between the gender and the deaths among infected people. The contingency table is as follows:

	Infected	Deaths	Recoveries
Men	30	1067	1097
Women	17	1120	1137
Total	47	2187	2234

Table 1: Contingency Table: Virus Infection

a) Fisher's exact test

First, we shortly describe the null and alternative hypothesis when testing with the Fisher's exact test.

- Null Hypothesis: There is no association (dependence) between gender and deaths.
- Alternative Hypothesis: There is a significant association (dependence) between gender and deaths.

Then we perform the Fisher's exact test. We obtained a p-value 0f 0.054, which is larger than the significance level. Therefore, we failed to reject the null hypothesis that there is no relationship between gender and deaths.

b) Fisher's exact test 2

A variation of the previous test will be presented to gain more insights about the relationship between gender and deaths.

- Null Hypothesis: There is no association (dependence) between gender and deaths.
- Alternative Hypothesis: The proportion of male deaths is significantly higher than the proportion of women's deaths.

After performing the test, we obtained a p-value of 0.03, which is lower than the significance level. Therefore, we reject the hypothesis that there is no association between gender and fatalities among infected. We believe that there is a relationship between gender and fatalities; namely that the males are more likely to die than females when infected.

c) Find the p-value

The last part of this exercise was to find the p-value of b) through a suitable application of *phyper*. We realized that under H_0 $N_1 1$ is hypergeom(2234, 1097,47). That is equivalenty in R to phyper(30, 1074, 23, 47) for the left p-value and 1-phyper(30-1, 1074, 23, 47) for the right p-value. After analyzing these p-values, we again reject the null hypothesis. We can conclude that deaths among men are more common than among women.

6.3: Effectiveness Against Nausea

We are given a table that contains data about post-operative nausea after medication against nausea. The rows are composed of the different medicines/methods given to the patients. The first column contains the total number of patients and the second contains the incidence of nausea.

Treatment	Number of Patients	Incidence of Nausea
Placebo	165	95
Chlorpromazine	152	52
Dimenhydrinate	85	52
Pentobarbital (100mg)	67	35
Pentobarbital (150mg)	85	37

Table 2: Incidence of Nausea in Different Treatment Groups

a) Most suitable model

The most suitable model for these data is the model II B. This model assumes in the null hypothesis that the r samples are homogeneous (every row is independent). In this particular case, it assumes that the effectiveness of the drug/method is the same for all drugs/methods (the rows).

The other models are not a good fit. Model II A assumes that rows and columns are independent, while Model II C assumes Independence in the columns. However, under this specific circumstances, only the rows are independent that is why model II B is the best fit.

b) Hypothesis Test

In this section, we use the chi-square test independence of categorical variables. First, we ensure that the rule of thumb is satisfied. Because the sample size is not that large, we might need to confirm these results by simulating the p-value (later on). The null and alternative hypothesis are the following:

- Null Hypothesis: Drugs/methods are homogeneous in the sense that they have the same effectiveness against nausea.
- Alternative Hypothesis: It is not homogeneous; and there are some drugs that are significantly more efficient than others.
- Under H_O, the distribution is approx. X^2-distributed with (r-1)(c-1) dfs.

With the alternative hypothesis, we are testing if the nausea depends on the drug used or not. After conducting the test, we get a p-value of 0.07, which is higher than the significance level. Therefore, we cannot reject the null hypothesis.

c) Simulate p-value

Next, we use the same R method *chisq.test*, but now with the parameter simulate.p.value set as True. This will simulate a p-value based on the variable given. In this case, we obtained, just as before, a p-value of 0.07. Based on this result, we failed to reject the null.

Note: this method is very useful when the rule of thumb is not satisfied.

d) Contributions and Standardized Residuals for Tests

The next step is to find the contributions and the standardized residuals for the chi square test performed before. From these two table, we obtained several insights. We found out that the categories that appear more often than expected are: While the categories that appear less than expected are:

e) Bootstrap Method

In this section, we performed another hypothesis test. However, this time we conducted the test through the bootstrap method. We use the test statistic = "The largest of the absolute value of the contributions".

f) One-sided Fisher's Exact Test

The last part of the assignment was to perform a one-sided Fisher's exact test to find out whether Chlorpromazine works better than the placebo. The Null and Alternative hypothesis for this test are the following:

- Null Hypothesis: The proportion of patients with incidences of nausea is independent of the different medication/methods used to treat them.
- Alternative Hypothesis: There is a significant difference between the proportion of patients with incidence of nausea among different methods. Namely, chlorpronazine works better than the placebo.

However, we must re formulate the alternative so it is in terms of N11. We then have as alternative hypothesis: "The placebo method is less effective than clorpronazine."

We obtained a a p-value of $2.3x10^{-}5$. Therefore we reject the null hypothesis and accept that chlorpronazine works better than the placebo.

Appendix

6.1: Code

```
expenses_crime = read.table('expensescrime.txt', header=TRUE)
sample61 = subset(expenses_crime, pop*1000 > 3000000)

numSample = subset(sample61, select=c('expend', 'bad', 'crime', 'lawyers', 'employ', 'pop'))
pairs(numSample)

plot(sample61$expend, sample61$crime)
cor(sample61$expend, sample61$crime) #0.3626

#Kendalls
cor.test(sample61$expend, sample61$crime, method='k')

#Spearmans
cor.test(sample61$expend, sample61$crime, method='s')

B=1000
t2 = cor.test(sample61$expend, sample61$crime, method='s')[[1]]
t_perm2 = numeric(B)
twoSamples2 = c(sample61$expend, sample61$crime)
for (i in 1:B){
  sample0 = sample(twoSamples2, size=length(twoSamples2))
  sample12 = sample(sample0, length(sample61$crime))
  sample22 = sample(setdiff(sample0, sample12), size=length(sample61$expend))
  t_perm2[i] =  cor.test(sample12, sample22, method='s')[[1]]
}
pl2 = mean(t_perm2 < t2)
pr2 = mean(t_perm2 > t2)
p2 = 2 *min(pl2,pr2)

areSimulation = function (B, n){
  are=0
  pvalK = pvalSp = numeric(B)
  for (i in 1:B){
    x = rmvt(n, sigma=matrix(c(1,0.2,0.2,1), 2, 2), df=6)
    pvalK[i] = cor(x, method='k')[[3]]
    pvalSp[i] = cor(x, method='s')[[3]]
  }
  powerK = mean(pvalK < 0.05)
  powerSp = mean(pvalSp < 0.05)
  rbind(c("Kendals", 'Spearmans'), c(powerK, powerSp) )
  are = (powerK/powerSp)^2
  return (are)
}

areSimulation(1000, length(sample61$crime)) #aprox = 1.27
```

6.2: Code

```
table = matrix(c(30, 17, 1067, 1120),2,2)

fisher.test(table)
# p-value = 0.054 > 0.05 => we do not reject hypothesis
```

```
fisher.test(table, alternative='g')
# p-value=0.03 < 5 => we reject the null hypothesis

pl=phyper(30, 1074,23,47)
pr=1-phyper(30-1, 1074, 23, 47)
c(pl, pr)
```

6.3: Code

```
source('functions_Ch7.txt')

table = read.table('nausea.txt')
cont_table = cbind(table, table[,1]-table[,2])
colnames(cont_table)[3] = 'No Incidence'
cont_table = cont_table[,c(2,3)]

chisq.test(cont_table)$expected # rule of thumb is fine
chisq.test(cont_table) # p-value=6.32e-5

chisq.test(cont_table, simulate.p.value=T) #p-value=0.001

residuals = cont_table - chisq.test(cont_table)$expected
residuals
contribution = chisq.test(cont_table)$residuals
contribution

round(chisq.test(cont_table)$stdres,2)

t = max(abs(chisq.test(cont_table)$residuals)) #2.59
maxcontributionscat(cont_table) #2.59
BS = bootstrapcat(cont_table ,1000, maxcontributionscat)
mean(BS>=t2)

cont_table[c(1,2), ]
temp = matrix(c(95,52,70,100),2,2)
fisher.test(temp)
fisher.test(temp, alternative='g')
```