SDA - Assignment 6

Leon Lušić (2670440) - no group

Exercise 6.1

(b) Plotting the initial lawyers rate versus `crime`, we find that there is a noticeable outlier to the data (in the upper-right corner).

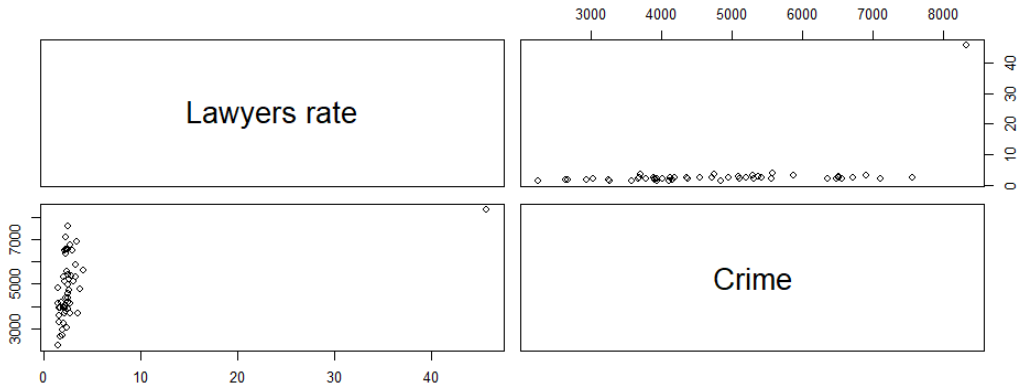

Figure 1: Lawyers rate vs. `crime`

In order to handle this, we may simply set the value of lawyers rate for this data point to the mean of the sample to get a better idea of the shape of the data. Doing so, we obtain the following:

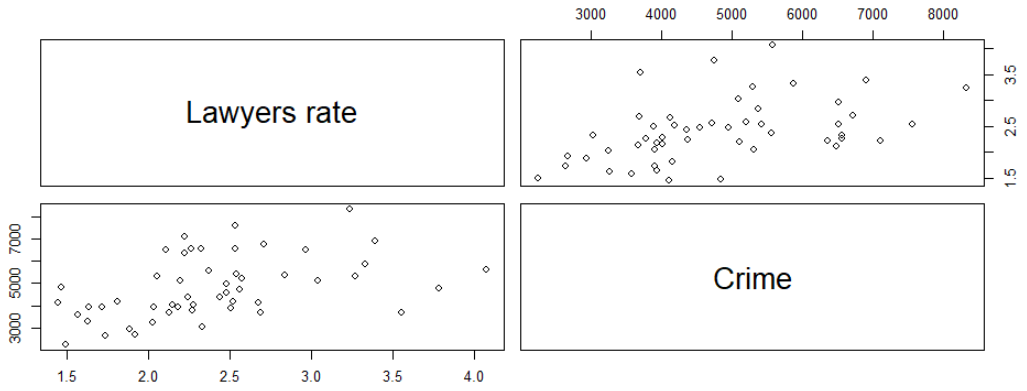

Figure 2: Lawyers rate adjusted vs. `crime`

Based on the adjusted plot, there seems to be a positive correlation between the lawyers rate and `crime`.

(c) Performing the Kendall's and Spearman's correlation tests, we find the following p-values:

Test	Kendall's corr. test	Spearman's corr. test
p-value	0.00019	0.00012

Table 1: Table of p-values

Based on the results of these tests, we may confidently reject the null hypothesis that the lawyers rate and `crime` are independent.

(d) Performing a permutation test based on Kendall's rank correlation coefficient with 10000 bootstrapped samples, we find the p-values of 0.0002 which is very close to the p-value found in part **(c)**.

(e) Given that both p-values from part **(c)** and part **(d)** are significantly less than 0.05, we may safely reject the null hypothesis and conclude that the lawyers rate and `crime` are dependent.

(f) In order to calculate asymptotic relative efficiency of Kendall's rank correlation test to Spearman's rank correlation test, we calculate the power of both tests for samples sizes $n = 45$, $n = 50$, and $n = 55$. Calculating the ARE from these simulations, we obtain the following values:

Kendall \ Spearman	n = 45	n = 50	n = 55
n = 45	0.999	0.967	0.950
n = 50	1.031	0.998	0.981
n = 55	1.050	1.017	0.999

Table 2: ARE for different sample sizes

As can be seen, the values closest to 1 lie along the diagonal of the Table (2). Hence, the true asymptotic relative efficiency of Kendall's test with respect to Spearman's test appears to be 1.

Exercise 6.2

The dataset for Exercise 6.2 is given by the following table:

infected	deaths	recovered	total
men	24	1020	1044
women	14	1167	1182
total	39	2187	2226

Table 3: Outcomes of infected patients based on gender and outcome

(a) First, we test whether the row and column variables are independent. Hence, the null and alternative hypotheses are:

H_0 : the row and column variables are independent,

H_1 : the row and column variables are not independent.

The p-value in this case is 0.0745 which is not less than $\alpha = 0.05$, so we do not reject the null hypothesis.

(b) Next, we test whether men are more often among the fatalities than women. Hence, the null and alternative hypotheses are:

H_0 : men are less or equally often among the fatalities as women,

H_1 : men are more often among the fatalities than women.

The p-value in this case is 0.0458 which is less than 0.05, so we do reject the null hypothesis and accept the hypothesis that men are more likely among the fatalities than women.

(c) Performing the same directed test, but using the theoretical expected hypergeometric distribution, we find that twice the minimum of the right and left-sided p-values is equal to 0.0917 which is greater than 0.05. Therefore, via this approach, we should not reject the null hypothesis.

Exercise 6.3

The data can be equivalently presented in the following form:

Type of medication	Nausea	No nausea	Number of patients
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 4: Table of 95%-conf. intervals - excluding outliers

We will use the 'Nausea' and 'No nausea' data for the following exercises.

(a) The most suitable model to the data is model II B as it assumes, in this case, five 2-nomial distributions. Within each sample, the patient either experienced nausea or did not. Hence, the sum of probabilities for either event withing each sample (represented by rows) should be 1. This setup corresponds to model II B as explained in the handout slides.

(b) Under the null hypothesis, the five samples should be homogeneous. Naturally, under the alternative hypothesis, they shouldn't. Applying the χ^2-test, we find the p-value of 0.00006 which is a lot smaller than $\alpha = 0.05$. After checking the rule of thumb, we find that all expected values are non-negative and that at least 80% of them (in fact, all of them) are greater than 5. Therefore, we may confidently reject the null hypothesis.

(c) Using bootstrap simulations via `chisq.test(..., simulate.p.value=TRUE)`, we find a p-value of 0.0005. Although almost 10 times larger than the p-value found in part **(b)**, it is still under 0.05. This further verifies our previous result and improves reliability.

(d) Computing the contributions and standardized residuals for the χ^2-test, we obtain the following values:

Type of medication	Nausea	No nausea
Placebo	1.59	-1.56
Chlorpromazine	-2.59	2.54
Dimenhydrinate	1.62	-1.58
Pentobarbital (100mg)	0.39	-0.38
Pentobarbital (150mg)	-0.71	0.69

Table 5: Contributions for the χ^2-test

Type of medication	Nausea	No nausea
Placebo	2.66	-2.66
Chlorpromazine	-4.26	4.26
Dimenhydrinate	2.46	-2.46
Pentobarbital (100mg)	0.58	-0.58
Pentobarbital (150mg)	-1.08	1.08

Table 6: Standardized residuals for the χ^2-test

Given that we chose the significance level of 0.05 and since $\Phi^{-1}(\alpha/2) \approx -1.96$ and $\Phi^{-1}(1-\alpha/2) \approx 1.96$, we notice that the 'Placebo', 'Chlorpromazine', and 'Dimenhydrinate' samples stand out.

(e) Performing the bootstrap test using the statistic T corresponding to the largest absolute value of the contributions, we find a p-value of 0.0004.

(f) Although the p-value found in part **(e)** is an order of magnitude larger than the value found in part **(b)**, it is still a few orders of magnitude below 0.05. Therefore, we still safely reject the null hypothesis. Note, however, that it is very similar to the p-value found in part **(c)**.

(g) As can be seen in the contributions table in part **(d)**, "Chlorpromazine" sample appears to be the most responsible for the value of the test statistic T (as described in part **(e)** since the values coming from that sample have the largest absolute value. Hence, we should conduct a one-sided Fisher's exact test to check whether Chlorpromazine works better than the placebo. The null and alternative hypotheses are as follows:

H_0 : the number of people who experience nausea is independent among the patients that took a placebo and that took Chlorpromazine.

H_1 : the number of people who experienced nausea is higher among patients that took a placebo.

Performing the test, we find the p-value of 0.000023 which is vastly smaller than 0.05. Therefore, we may confidently reject the null hypothesis and conclude that Chlorpromazine works better than the placebo.

Appendix

```
### Exercise 6.1 ###

data = read.csv("expensescrime.txt", sep=" ")
nausea = read.csv("nausea.txt", sep=" ")

library(mvtnorm)
```

```
source("functions_Ch7.txt")

# a)

pairs(cbind(data$expend, data$bad))
pairs(cbind(data$expend, data$crime))
pairs(cbind(data$expend, data$lawyers))
pairs(cbind(data$expend, data$employ))
pairs(cbind(data$expend, data$pop))

pairs(cbind(data$bad, data$crime))
pairs(cbind(data$bad, data$lawyers))
pairs(cbind(data$bad, data$employ))
pairs(cbind(data$bad, data$pop))

pairs(cbind(data$crime, data$lawyers))
pairs(cbind(data$crime, data$employ))
pairs(cbind(data$crime, data$pop))

pairs(cbind(data$lawyers, data$employ))
pairs(cbind(data$lawyers, data$pop))

pairs(cbind(data$employ, data$pop))

# b)

lawyers_rate = data$lawyers/data$pop

pairs(cbind(lawyers_rate, data$crime), labels=c('Lawyers rate', '
   Crime'))

lawyers_rate[8] = mean(lawyers_rate)

pairs(cbind(lawyers_rate, data$crime), labels=c('Lawyers rate', '
   Crime'))

# c)

cor.test(lawyers_rate, data$crime, method='k')
cor.test(lawyers_rate, data$crime, method='s')

# d)

B = 10000
t = cor.test(lawyers_rate, data$crime, method='k')$p.value
permutationval = numeric(B)
for(i in 1:B) {
  sample_crime = sample(data$crime)
  pl = cor.test(lawyers_rate,
                sample_crime,
```

```
                method='k',
                alternative='g')$p.value
  pr = cor.test(lawyers_rate,
                sample_crime,
                method='k',
                alternative='l')$p.value
  p = 2*min(pl, pr)
  permutationval[i] = p
}

length(permutationval[permutationval<=t])/B

# f)

# Asymptotic Relative Efficiency function
aresimulation = function(B, n) {
  pvalkendalltest=numeric(B)
  pvalspearmantest=numeric(B)
  for(i in 1:B)
  {
    v = rmvnorm(n, mean=c(0,0), sigma=matrix(c(1,0.5,0.5,1), 2,2))
    pvalkendalltest[i] = cor.test(v[,1], v[,2], method='k')$p.value
    pvalspearmantest[i] = cor.test(v[,1], v[,2], method='s')$p.value
  }
  powerkendall=sum(pvalkendalltest<0.05)/B
  powerspearman=sum(pvalspearmantest<0.05)/B
  sim = rbind(c("Kendall", "Spearman"), c(powerkendall, powerspearman
   ))
  sim[2,] = as.numeric(sim[2,])
  sim
}

B = 10000

# test 1: n = 45
n = 45
sim1 = aresimulation(B, n)

# test 2: n = 50
n = 50
sim2 = aresimulation(B, n)

# test 3: n = 55
n = 55
sim3 = aresimulation(B, n)

col1 = c(as.numeric(sim1[2,1])/as.numeric(sim1[2,2]),
         as.numeric(sim2[2,1])/as.numeric(sim1[2,2]),
         as.numeric(sim3[2,1])/as.numeric(sim1[2,2]))

col2 = c(as.numeric(sim1[2,1])/as.numeric(sim2[2,2]),
```

```
         as.numeric(sim2[2,1])/as.numeric(sim2[2,2]),
         as.numeric(sim3[2,1])/as.numeric(sim2[2,2]))

col3 = c(as.numeric(sim1[2,1])/as.numeric(sim3[2,2]),
         as.numeric(sim2[2,1])/as.numeric(sim3[2,2]),
         as.numeric(sim3[2,1])/as.numeric(sim3[2,2]))

round(matrix(c(col1, col2, col3), nrow=3, ncol=3),3)

### Exercise 6.2 ###

alpha = 0.05

infected = matrix(c(24, 15, 1020, 1167),
                  nrow=2, ncol=2,
                  dimnames=list(c('men', 'women'),
                                c('deaths', 'recoveries')))

# a)

# null hypothesis: the row and column variables are independent
# alternative hypothesis: the row and column variables are dependent

fisher.test(infected)

# b)

# null hypothesis: men are less often among the fatalities than women
# alternative hypothesis: men are more often among the fatalities
    than women

fisher.test(infected, alt='g')

# c)

pl = phyper(24, 1044, 1182, 39)
pr = 1 - phyper(24-1, 1044, 1182, 39)
2*min(c(pl, pr))


### Exercise 6.3 ###

data = data.frame(nausea$Incidence.of.Nausea,
                  nausea$Number.of.Patients-nausea$Incidence.of.
   Nausea,
                  row.names = rownames(nausea))
colnames(data) = c('Nausea', 'No nausea')

alpha = 0.05

# b)
```

```
chisq.test(data)

chisq.test(data)$expected

# c)

chisq.test(data, simulate.p.value=TRUE)

# d)

round(chisq.test(data)$residuals, 2)

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

# e)

B=10000

t = maxcontributionscat(data)
boot = bootstrapcat(data, B, maxcontributionscat)

mean(boot>=t)

# g)

fisher.test(data[1:2,], alternative='greater')
```