Module 2 - Measuring Dependence

Pawel Chilinski

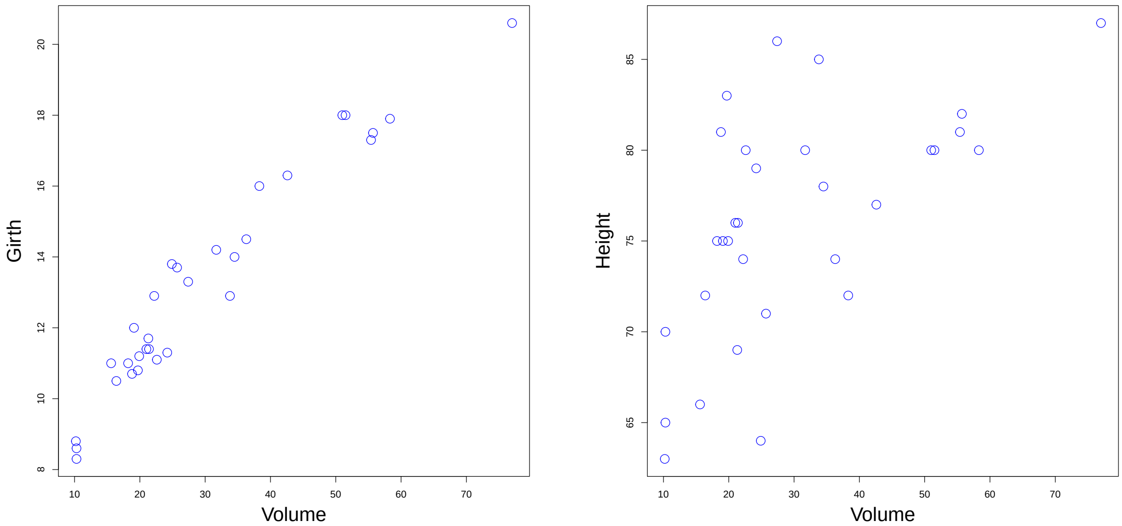
October 24, 2013

Exercise 2.

In the base library (loaded by default) of package R there is a data set trees available.

- Using function plot() obtain scatterplots of two pairs of variables: Volume and Girth, Volume and Height.

```
> par(mfrow=c(1,2),mar=c(5,5,5,5))
> plot(trees$Volume,trees$Girth, xlab="Volume",ylab="Girth",col="blue",cex.lab=2, cex=2)
> plot(trees$Volume,trees$Height, xlab="Volume",ylab="Height",col="blue",cex.lab=2, cex=2)
```



- Using function cor(), calculate an empirical correlation coefficient and an empirical Spearman rank correlation coefficient for the two pairs of variables.

```
> cor(trees$Volume,trees$Girth,method="pearson")

[1] 0.9671194

> cor(trees$Volume,trees$Height,method="pearson")

[1] 0.5982497

> cor(trees$Volume,trees$Girth,method="spearman")

[1] 0.9547151

> cor(trees$Volume,trees$Height,method="spearman")

[1] 0.5787101
```

- Using function cor.test(), perform correlation tests of independency based on Pearson's r coefficient and Spearman's ρ coefficient for two pairs of variables: Volume and Girth, Volume and Height, at a significance level 0.05

```
> cor.test(trees$Volume,trees$Girth,method = "pearson",conf.level = 0.95)

        Pearson's product-moment correlation

data:  trees$Volume and trees$Girth
t = 20.4783, df = 29, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
```

```
 0.9322519 0.9841887
sample estimates:
      cor
0.9671194
```

```
> cor.test(trees$Volume,trees$Height,method = "pearson",conf.level = 0.95)

        Pearson's product-moment correlation

data:  trees$Volume and trees$Height
t = 4.0205, df = 29, p-value = 0.0003784
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3095235 0.7859756
sample estimates:
      cor
0.5982497
```

```
> cor.test(trees$Volume,trees$Girth,method = "spearman",conf.level = 0.95)

        Spearman's rank correlation rho

data:  trees$Volume and trees$Girth
S = 224.613, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9547151
```

```
> cor.test(trees$Volume,trees$Height,method = "spearman",conf.level = 0.95)

        Spearman's rank correlation rho

data:  trees$Volume and trees$Height
S = 2089.598, p-value = 0.0006484
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.5787101
```

So all the tests allow us to reject the null hypothesis that correlation is 0 (using p cutoff value of 0.05).

Exercise 3.

File patients.txt contains 5 variables measured on 204 patients of Wroclaw outclinics. We are interested in examining dependence between education of patients and their marital status.

- Represent the data as a contingency table (use function table()).

```
> patients <- read.table(file="patients.txt",header=T)
> ed_mar_table <- table(patients[,c("education","marital")])
> ed_mar_table

            marital
education    involved single
  elementary       29     61
  secondary        29     26
  university       14     20
  vocational       12     13
```

- Using function summary() or chisq.test(), perform chi-square independence test.

```
> summary(ed_mar_table)

Number of cases in table: 204
Number of factors: 2
Test for independence of all factors:
        Chisq = 6.489, df = 3, p-value = 0.09008
```

```
> test <- chisq.test(ed_mar_table)
> test

        Pearson's Chi-squared test

data:  ed_mar_table
X-squared = 6.4894, df = 3, p-value = 0.09008

> #Should we reject H0?
> test$p.value<0.05

[1] FALSE
```

We cannot reject H_0 that education and marital variables are independent using 0.05 significance level.

- Analyse Pearson residuals (chisq.test()$residuals) and relate them with the test result.

  If we compute standardized values of n_{ij} under H_0 then we see that all the values are not much bigger than values expected only by chance (sampling error):

```
> test$residuals/sqrt(1-test$expected/sum(ed_mar_table))

            marital
education      involved     single
  elementary -1.4633870  1.2871136
  secondary   1.4158646 -1.2176286
  university  0.0000000  0.0000000
  vocational  0.5456299 -0.4617959
```

so we shouldn't commit to reject H_0 which agrees with the test.

- Perform chi-square independence tests for subgroups of men and women seperately (whenever number of observations is satisfactory).

  We can only perform the test for the men data because it fulfils the rule of thumb that no cell of the expected table (under null hypothesis) should have values smaller than 5. We cannot reject the H_0 in this case either.

```
> patients_men <- patients[patients$gender=="male",]
> ed_mar_men_table <- table(patients_men[,c("education","marital")])
> ed_mar_men_table

            marital
education    involved single
  elementary       25     43
  secondary        22     17
  university       12     12
  vocational        8      7

> ed_mar_men_table_expected <- outer(apply(ed_mar_men_table,1,sum)/sum(ed_mar_men_table),
+                                    apply(ed_mar_men_table,2,sum)/sum(ed_mar_men_table))*
+   sum(ed_mar_men_table)
> ed_mar_men_table_expected

            involved    single
elementary 31.205479 36.794521
secondary  17.897260 21.102740
university 11.013699 12.986301
vocational  6.883562  8.116438

> #Rule violated?
> any(ed_mar_men_table_expected<5)

[1] FALSE

> chisq.test(ed_mar_men_table)

        Pearson's Chi-squared test

data:  ed_mar_men_table
X-squared = 4.5166, df = 3, p-value = 0.2108
```

```
> patients_women <- patients[patients$gender=="female",]
> ed_mar_women_table <- table(patients_women[,c("education","marital")])
> ed_mar_women_table

            marital
education    involved single
  elementary        4     18
  secondary         7      9
  university        2      8
  vocational        4      6

> ed_mar_wommen_table_expected <- outer(apply(ed_mar_women_table,1,sum)/sum(ed_mar_women_table),
+                                       apply(ed_mar_women_table,2,sum)/sum(ed_mar_women_table))*
+   sum(ed_mar_women_table)
> ed_mar_wommen_table_expected

           involved    single
elementary 6.448276 15.551724
secondary  4.689655 11.310345
university 2.931034  7.068966
vocational 2.931034  7.068966

> #Rule violated?
> any(ed_mar_wommen_table_expected<5)

[1] TRUE
```

- Using function fisher.test(), perform Fisher independence test.

```
> f_test <- fisher.test(ed_mar_table)
> f_test

        Fisher's Exact Test for Count Data

data:  ed_mar_table
p-value = 0.08696
alternative hypothesis: two.sided

> #Should we reject H0 that variables are independent at 5% significance level?
> f_test$p.value < 0.05

[1] FALSE
```

Result from the Fisher test doesn't allow us to reject H_0 either.

Using Fisher test we can test data from women:

```
> f_test_women <- fisher.test(ed_mar_women_table)
> f_test_women

        Fisher's Exact Test for Count Data

data:  ed_mar_women_table
p-value = 0.2769
alternative hypothesis: two.sided

> f_test_women$p.value < 0.05

[1] FALSE
```

So wo cannot reject H_0 at 0.05 significance level.

Exercise 4.

File kids.txt contains pupils' answers in a questionaire about the most important quantities for them.

```
> #load the data
> kids <- read.table(file="kids.txt",header=T)
```

- Do answers concerning importance of looks depend on the gender of the questioned? Perform an appropriate test and interpret its result.

```
> kids_table_gender_looks <- table(kids[,c("Gender","Looks")])
> kids_table_gender_looks

      Looks
Gender   1   2   3   4
  boy   44  74  59  50
  girl 141  52  42  16

> kids_table_gender_looks_expected <- outer(apply(kids_table_gender_looks,1,sum)/
+                                               sum(kids_table_gender_looks),
+                                         apply(kids_table_gender_looks,2,sum)/
+                                               sum(kids_table_gender_looks))*
+   sum(kids_table_gender_looks)
> kids_table_gender_looks_expected

            1        2        3       4
boy  87.85565 59.83682 47.96444 31.3431
girl 97.14435 66.16318 53.03556 34.6569

> #Can we perform Chi-square tests?
> all(kids_table_gender_looks_expected>=5)

[1] TRUE

> chisq_test_kids_gender_looks <- chisq.test(kids_table_gender_looks)
> chisq_test_kids_gender_looks

        Pearson's Chi-squared test

data:  kids_table_gender_looks
X-squared = 74.0589, df = 3, p-value = 5.765e-16

> #Can we reject H0?
> chisq_test_kids_gender_looks$p.value < 0.05

[1] TRUE
```

The H_0 can be rejected because Chi-square tests gives p-value much smaller than 0.05.

- What measures of dependence may be used in this situation?

  We can use dependence measures suitable for nominal data only because gender is nominal and Looks is ordinal: Goodman-Kruskal dependence index (the average decrease of uncertainty about Looks if we know gender) , Conditional Gini index(where V(Looks|gender=female)=0 means that Looks is completely determined when gender is female), average value of conditional Gini index(average uncertainty about Looks when we know gender).

- Calculate Gini index for variable Looks.

```
> options(width=60)
> #computes gini index for variable with name y
> gini <- function(y){
+   #compute gini index by summing up the products of estimates of the probabilty of giving given
+   #category and its complementary probabilty estimate.
+   sum(table(kids[[y]])/nrow(kids)*(1-table(kids[[y]])/nrow(kids)))
+ }
> gini("Looks")

[1] 0.717013
```

- Calculate Goodman-Kruskal dependence index for Gender and Looks.

```
> #computes V(Y|x), where y,x are names of variables and x_val is specific value for x
> gini_conditional <- function(y,x,x_val){
+   cont_table <- table(kids[,c(x,y)])
+   1-sum(sapply(unique(kids[[y]]),function(y_val){
+     cont_table[x_val,y_val]/sum(cont_table[x_val,])
+   })^2)
+ }
> #computes E(V(Y|X)), where y,x are names of variables
> expected_gini_conditional <- function(y,x){
+   cont_table <- table(kids[,c(x,y)])
```

```
+   sum(sapply(unique(kids[[x]]),function(x_val){
+      sum(cont_table[x_val,])/sum(sum(cont_table)) * gini_conditional(y,x,x_val)
+   }))
+ }
> #computes Goodman-Kruskal tau index, where y,x are names of variables
> goodman_kruskal_tau <- function(y,x){
+   (gini(y)- expected_gini_conditional(y,x))/gini(y)
+ }
> #relative decrease of variability of Looks when we know Gender
> goodman_kruskal_tau("Looks","Gender")
```

```
[1] 0.06349011
```

```
> #relative decrease of variability of Gender when we know Looks
> goodman_kruskal_tau("Gender","Looks")
```

```
[1] 0.154935
```

- Calculate Kendall's τ coefficient (function cor() or function Kendall() in package Kendall) for all pairs of ordered variables which are present in the data set.

```
> options(width=60)
> ordered_variables <- c("Grade","Age","Grades","Looks","Sports","Money")
> variable_combinations <- combn(ordered_variables,2)
> variable_combinations
```

```
     [,1]    [,2]     [,3]    [,4]     [,5]    [,6]
[1,] "Grade" "Grade"  "Grade" "Grade"  "Grade" "Age"
[2,] "Age"   "Grades" "Looks" "Sports" "Money" "Grades"
     [,7]    [,8]     [,9]    [,10]    [,11]    [,12]
[1,] "Age"   "Age"    "Age"   "Grades" "Grades" "Grades"
[2,] "Looks" "Sports" "Money" "Looks"  "Sports" "Money"
     [,13]    [,14]   [,15]
[1,] "Looks"  "Looks" "Sports"
[2,] "Sports" "Money" "Money"
```

```
> for(col in 1:ncol(variable_combinations)){
+   cat("variables: ",variable_combinations[1,col]," and ",variable_combinations[2,col],
+       ", Kendall's tau = ",
+       cor(kids[[variable_combinations[1,col]]],kids[[variable_combinations[2,col]]],
+           method="kendall"),"\n")
+ }
```

```
variables:  Grade  and  Age , Kendall's tau =  0.8121702
variables:  Grade  and  Grades , Kendall's tau =  0.2212102
variables:  Grade  and  Looks , Kendall's tau =  -0.133247
variables:  Grade  and  Sports , Kendall's tau =  -0.0920032
variables:  Grade  and  Money , Kendall's tau =  -0.02577502
variables:  Age  and  Grades , Kendall's tau =  0.1829944
variables:  Age  and  Looks , Kendall's tau =  -0.09011102
variables:  Age  and  Sports , Kendall's tau =  -0.1061929
variables:  Age  and  Money , Kendall's tau =  -0.02113476
variables:  Grades  and  Looks , Kendall's tau =  -0.3994896
variables:  Grades  and  Sports , Kendall's tau =  -0.111494
variables:  Grades  and  Money , Kendall's tau =  -0.3694347
variables:  Looks  and  Sports , Kendall's tau =  -0.380216
variables:  Looks  and  Money , Kendall's tau =  -0.04469412
variables:  Sports  and  Money , Kendall's tau =  -0.2601036
```