`R` instructions for the 1st seminar

download `R` : http://www.r-project.org/
user interface for R: http://www.rstudio.com/
An R Introduction to Statistics: http://www.r-tutor.com/

1)set working directory, 2)check files in working directory, 3)load the data, 4)find out the names of variables, 5)check variable classes

1) `setwd("C:/.../seminars/1_seminar")`
2) `dir()`
3) `load("Movies.RData")`
4) `names(Movies)`
5) `lapply(Movies,class)`

R Instructions for the problem 1:

•Frequency tables

absolute frequencies for the variable *fMovie*: `table(Movies$fMovie)`

absolute frequencies for all variables `lapply(Movies,table)`

relative frequencies for the variable *fMovie*: `prop.table(table(Movies$fMovie))`

relative frequencies for all variables: `lapply(Movies,function(my){prop.table(table(my))})`

absolute frequencies for the variable *fMovie* separately for man and woman:

`tapply(Movies$fMovie,Movies$fMan,table)`

relative frequencies for the variable *fMovie* separately for man and woman:

`tapply(Movies$fMovie,Movies$fMan,FUN=function(my)prop.table(table(my)))`

•Descriptive statistics

common descriptive statistics for the variable *Movie*:

```
mean(Movies$Movie, na.rm=TRUE)
median(Movies$Movie, na.rm=TRUE)
quantile(Movies$Movie, probs=c(0, 0.25, 0.5, 0.75, 1), na.rm=TRUE)
sd(Movies$Movie, na.rm=TRUE)
var(Movies$Movie, na.rm=TRUE)
```

the mean for variables *Movie* and *Man*:

`lapply(Movies[,1:2],mean, na.rm=TRUE)`

common descriptive statistics for variable *fMovie*:

`summary(Movies$fMovie)`

common descriptive statistics for all variables; notice the function `summary` gives different results for numeric and factor variables:

`summary(Movies)`

common descriptive statistics for the variable *Movie* categorized by variable *fMan*:

`tapply(Movies$Movie,Movies$fMan,summary,na.rm=T)`

A package "lattice" allows handy way to produce graphs like histograms, boxplots, scatterplots, etc. **Firstly**, this package should be installed; **secondly** it should be loaded every session user intends to use commands associated with this package.
ad1) `install.packages("lattice")` ad2) `library(lattice)`.

!`library(lattice)`

•Histograms

create a histogram for variable *Movie*:

`histogram(Movies$Movie,type="count",breaks=seq(0.5,5.5,1),col=24)`

create a histogram for variable *Movie* categorized by variable *fMan*:

`histogram(~Movies$Movie | Movies$fMan, type="percent",breaks=seq(0.5,5.5,1))`

(TODO rozmyslet oba grafy do jednoho obrazku jinak, nez pres "hist"; arg "'groups"' u histogram nefachci)

Assess the skewness - can the mean be used insted of the median?

•Box plots

create a boxplot for variable *Movie*:

`bwplot(Movies$Movie)`

create a boxplot for variable *Movie* categorized by variable *fMan*:

`bwplot(~Movies$Movie|Movies$fMan)`

create a boxplot for variable *Movie* categorized by variable *fMan*, both graphs in one picture:

`boxplot(Movies$Movie~Movies$fMan)`

`R` Instructions for the problem 2:
...analogously, e.g.:

```
summary(Household.marriage)
lapply(Household.marriage,function(my){m=mean(my);s=sd(my);v=var(my);return(c(m,s,v))})
```

```
histogram(Household.marriage$getmar ,type = "percent")
histogram(Household.marriage$ownhh ,type = "percent")
```

Notice that the shape of histogram of variable *ownhh* is skewed positively whereas *getmar* is "symetric". Thus using of the mean in case of the *ownhh* is unacceptable whereas in case of the *getmar* it can be accepted.

(TODO rozmyslet oba grafy do jednoho obrazku, par(mfrow(c(1,2)) nefachci

`R` Instructions for the problem 3:

•<u>*t*-test</u>
There are 1322 cases in a sample, so according to the large sample size, *t*-test is acceptable.
Have a look at categorized histograms, descriptive statistics,...

```
histogram(~Movies$Movie| Movies$fMan)
bwplot(~Movies$Movie| Movies$fMan)
tapply(Movies$Movie,Movies$fMan,function(my){m=mean(my);v=var(my);return(c(m,v))})
```

Before running t-test it is necessary to assess the assumption of equal sigmas which is done via F-test:

```
var.test(Movies$Movie[Movies$fMan=="man"],Movies$Movie[Movies$fMan=="woman"],ratio=1,
alternative="two.sided")
```

As the F-test did not rejected equality of sigmas (p-value = 0.4058) we can proceed with two-sample t-test:

```
t.test(Movies$Movie[Movies$fMan=="man"],Movies$Movie[Movies$fMan=="woman"],
mu=0,var.equal=T)
```

(Notice, p-value of the t-test (p-value = 0.0001746) can be supplemented by *Cohen's d* for effect size. Code for calculating *Cohen's d* is at the end of this file.)

•<u>Wilcoxon rank sum test</u>
A) Exact test

```
!library(exactRankTests)
wilcox.exact(Movies$Movie[Movies$fMan=="man"],Movies$Movie[Movies$fMan=="woman"],mu=0,ex
```

This exact test is feasiable only for small sample sizes, so considered sample "Movies" can not be processed by this test.

B) Asymptotical test without continuity correction

```
wilcox.test(Movies$Movie[Movies$fMan=="man"],Movies$Movie[Movies$fMan=="woman"],mu=0,
exact=FALSE, correct=FALSE)
```

C) Asymptotical test with continuity correction.

```
wilcox.test(Movies$Movie[Movies$fMan=="man"],Movies$Movie[Movies$fMan=="woman"],mu=0,
exact=FALSE, correct=TRUE)
```

•<u>χ^2 test</u>
Firstly we have to create contingency tables of categorial variables fMovie and FMan:
`t<-table(Movies$fMovie,Movies$fMan); t` table of absolute frequences
`addmargins(t)` absolute frequences with margins
`prop.table(t)` relative frequences (cell percentages)

`addmargins(prop.table(t))` relative frequences with margins
`prop.table(t,1)` row percentages
`prop.table(t,2)` column percentages
`margin.table(t,1)` row margins
`margin.table(t,2)` column margins

Performing χ^2-test:
`chisq.test(t)` provides p-value and the test-statistic
`chisq.test(t)$expected`
`chisq.test(t)$observed`
`chisq.test(t)$residuals`

R Instructions for the problem 4:
•ANOVA
As there are 1322 values in a data set ANOVA is acceptable.

a) **Categorised boxplots in one picture:**
`boxplot(Household.education$ownhh~Household.education$fdegree4)`

b) **Descriptive statistics:**
```
tapply(Household.education$ownhh,Household.education$fdegree4,mean)
tapply(Household.education$ownhh,Household.education$fdegree4,sd)
tapply(Household.education$ownhh,Household.education$fdegree4,summary)
```

c) **Categorized histograms in separate pictures:**
`histogram(~Household.education$ownhh|Household.education$fdegree4)`
(Categorized QQ-plots can be supplemented. A code for it is at the end of this ?le.)

d) **Assumption of equality of variances:**
`!library(DescTools)`
`LeveneTest(Household.education$ownhh~Household.education$fdegree4,center=mean)`
(For our data equality of variances was rejected and ANOVA should not be performed. Following steps are just to demonstrate the R functions related to ANOVA method. However, R offers also function for performing ANOVA with unequal variances, this test is only asymptotical and not included in basic textbooks.
`oneway.test(Household.education$ownhh~Household.education$fdegree4,var.equal = FALSE)`)

e) **ANOVA test:**
`model<-lm(Household.education$ownhh~Household.education$fdegree4)`
Attention! The factor variable in a model definition must be of class "factor". Otherwise, e.g class "numeric" for variable "degree" will lead to false results.
`anova(model)`
or `aov(Household.education$ownhh~Household.education$fdegree4)`
`summary(aov(Household.education$ownhh~Household.education$fdegree4))`

f) **Post-hoc tests:**
Firstly package "agricolae" has to be installed and downloaded.
`!library(agricolae)`
`scheffe.test(aov(ownhh~fdegree4,data=Household.education),"fdegree4", group=TRUE,`

`console=TRUE,alpha=0.104)`
(This function does not provide p-values, only shows whether or not means are significantly different at the α level. When $\alpha = 0.05$, no pair is for our data significantly different; here contrast can be significant.)

TukeyHSD test is appropriate for balanced design (which is not our case), thus following `R` function is just to demonstrate a syntax.
`TukeyHSD(aov(Household.education$ownhh~Household.education$fdegree4))`
(Performed p-values are smaller then true p-values, as the test assumes balanced groups.)

Another option in base package is Bonferroni Multipl comparisons method:
```
pairwise.t.test(Household.education$ownhh,Household.education$fdegree4,
p.adjust.method="bonferroni")
```

f) **Analyzing residuals:**
`shapiro.test(residuals(lm(Household.education$ownhh~Household.education$fdegree4)))`
(As the sample size is large it is no surprise that the normality was rejected.

```
qqnorm(residuals(lm(Household.education$ownhh~Household.education$fdegree4)))
qqline(residuals(lm(Household.education$ownhh~Household.education$fdegree4)))
```

•Kruskal-Walis test
```
histogram(~Household.education$ownhh|Household.education$fdegree4, type="percent")
kruskal.test(Household.education$ownhh~Household.education$fdegree4)
```

•χ^2 test
```
t<-table(Household.education$ownhh , Household.education$fdegree4); t
chisq.test(t)
```

script for Cohen's d:

```
...........................................................................
cohens_d <- function(x, y) {
    lx <- length(x)- 1
    ly <- length(y)- 1
    md  <- abs(mean(x) - mean(y))        ## mean difference (numerator)
    csd <- lx * var(x) + ly * var(y)
    csd <- csd/(lx + ly)
    csd <- sqrt(csd)                     ## common sd computation

    cd  <- md/csd                        ## cohen's d
}
> res <- cohens_d(Movies$Movie[Movies$fMan=="man"],Movies$Movie[Movies$fMan=="woman"])
> res
...........................................................................
```

script for normal Q-Q plot:

```
...........................................................................
cat.qq<-function(x,y){
  l<-length(levels(y))
  par(mfrow=c(l/2,2))
  for(i in levels(y) ){qqnorm(x[y==i],main=(paste(i," Normal Q-Q Plot")) )
                     qqline(x[y==i])}
  }
...........................................................................
```

useful:

!library(DescTools)

```
Freq(Household.marriage$ownhh )
PercTable(Household.marriage$ownhh , Household.marriage$getmar)
```