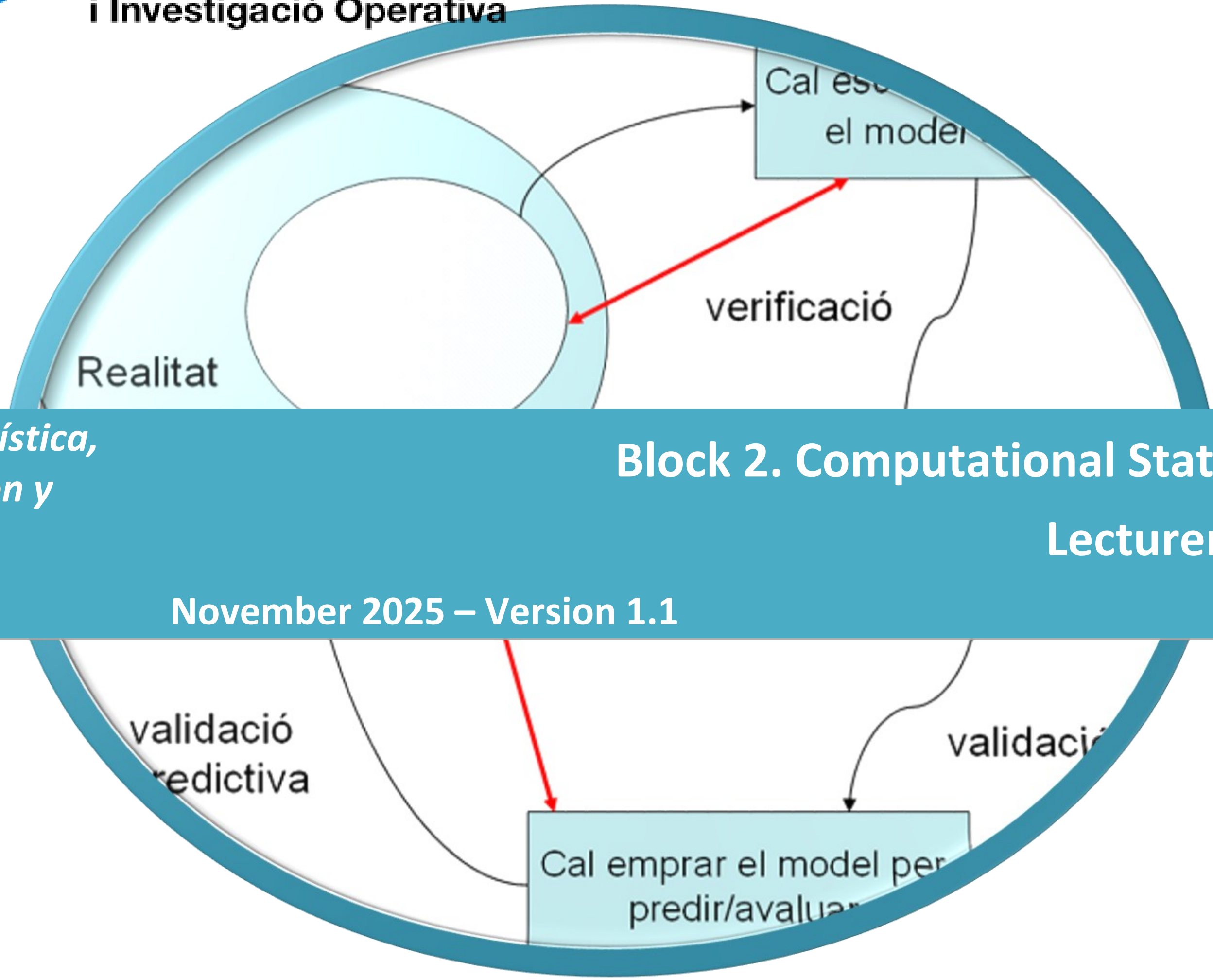
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Estadística
i Investigació Operativa

ACADEMIC YEAR 2025-26



P.01 - Estadística, Optimización y Sistemas – Estadística

Block 2. Computational Statistical Inference

Lecturer: Lídia Montero

November 2025 – Version 1.1

INDUSTRIA 4.0

TABLE OF CONTENTS

3-1. STATISTICAL INFERENCE AND COMPUTATIONAL EXPLORATION OF LARGE TABLES (DATABASE)

A single response variable is assumed, quantitative or qualitative, the paradigm is the same, the techniques differ:

- Sort systematically all explanatory variables according to their degree of association with the response variable.
- Tool description: contrast of hypothesis (statistical inference).
- The association can be defined:
  1. A global variable to variable: variables that best explain the response and to what degree.
  2. For groups (= between levels of categorical variables involved, whether response or explanatory).

It will be seen from the computational point of view R:

- Global association between quantitative resp. and a quantitative explanation (covariate).
- Global association between quantitative resp. and an explanatory resp. with I levels (factor).
- Global association between qualitative resp. with J levels and an explanatory resp. with I levels.
- Association between quantitative resp. and every level of an explanatory with I levels.
- Associating each J qualitative resp. level and each I explanatory resp. levels.

3-1 ... EXPLORATION OF LARGE TABLES (DATABASE)

3-1.1 Contrast of hypothesis

Working Paradigm: Formulation H0 (null hypothesis) and interpretation associated p value of contrast.

H0: "The X variable is not associated with Y"

1. Depending on the nature of H0 (type of the variables involved) will contrast with a statistical method or another.
2. All contrast provides a p value to be interpreted "Probability that the null hypothesis is true, if the conditions of applicability of contrast are given".
3. Calculating the p value according to the probability distribution of the statistical reference.

Interpretation of the p value:

- If p value is small (less than a threshold of 5 or 10%) then H0 is rejected, it is not credible and therefore evidence that its falsehood and it is claimed that there is evidence to believe that the variable Y is associated with the variable X.
- If p value is large (greater than a threshold of 5 or 10%) then H0 accepts and therefore there is no evidence that its falsity and hence the assertion that there is no evidence to believe that the variable Y is associated with the variable X.

3-1 ... INTERPRETATION OF TEST: LEVEL OF SIGNIFICANCE

Level of significance α

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$$

Frequently, α is set to 0.01 or 0.05

	(Hypothesis)	
	Actually True	Actually False
Accept	$1 - \alpha$	β (Type II error)
Reject	α (Type I error)	$1 - \beta$

3-1 ... EXPLORATION OF LARGE TABLES (DATABASE)

3-1.2 Global association between quantitative resp. and a quantitative explicative var.

- Measure the linear correlation between the 2 variables. You can test the null hypothesis of linear correlation coefficient equal to zero.
- The continuous explanatory variables would be sorted by their p-value in the contrast of hypothesis, low to high. Low p value indicates greater confidence in rejecting the null hypothesis and therefore greater association between the response and variable.

Parametric tests – Non parametric (method="spearman"): a universal example, weight for height.

```
# Continuous explanatory
attach(davis)
plot(weight ~ height, data=davis)
# Calculate the linear correlation coefficient between 2 quantitative variables involved
cor(davis$weight, davis$height, use="pairwise.complete.obs" )
```

```
# Calculate the linear correlation coefficient between all quantitative variables involved
cor(data.frame(weight, height,r_weight, r_height), use="pairwise.complete.obs"
)
```

```
#     cor(x, y, use = "all.obs", method = c("pearson", "kendall", "spearman"))
# cor.test()Methode contrast correlation equal to 0:
# cor.test(x, y,alternative = c("two.sided"
#      "spearman"),exact = NULL, conf.leve
cor.test(weight ~ height, data=d

# P.value attribute contains p v
cor.test(weight ~ height, data=d
```

```
> cor.test(davis$weight,davis$height)
        Pearson's product-moment correlation
data:  davis$weight and davis$height
t = 17.0397, df = 198, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to
0
95 percent confidence interval:
 0.7080838 0.8218898
```

3-1 ... EXPLORATION OF LARGE TABLES (DATABASE)

3-1.3 Global association between quantitative resp. and an explicative variate with I levels (factor)

- Acceptable normality assumption, essential independence and identical variance observations by groups (defined by levels, if not met R resetting G. L. test).
- Automated tools related to models analysis of variance with a factor (one-way). If I = 2 then it is a contrast of means from 2 subpopulations in unpaired samples. It is possible to relax the assumption of equal variance in the I subsamples reducing degrees of freedom of reference distributions.

Parametric tests with dichotomous factor: a universal example, weight by gender.

```
# Dichotomous explanatory
options(contrasts=c("contr.treatment","contr.treatment"))
attach(davis)
plot(weight ~ sex, data=davis)
plot.design(davis)
# Contrast equal variances: there are specific contrasts
var.test(weight ~ sex, data=davis)
# t.test() methode bilateral contrast with different variances (default)
t.test(weight ~ sex, data=davis)
```

3-1 ... EXPLORATION OF LARGE TABLES (DATABASE)

```
# t.test() methode contrast unilateral with different variances (default)
t.test(weight ~ sex, data=davis, alternative="greater")
```

```
t.test(weight ~ sex, data=davis, alternative="less")
# Equal variance contrast: generalizing of var.test() I > 2
bartlett.test(weight ~ sex, data=davis)
# One-way
oneway.test(weight ~ sex, data=davis)
```

```
pes <- oneway.test(weight ~ sex, data=davis, var.equal=FALSE)
```

```
# One-way: obtaining p_valor accessing attribute p.value
oneway.test(weight ~ sex, data=davis)$p.value
```

- **The explanatory variables available should be related to the p.valor: low to high. A lower p value, the greater degree of association.**
- **For enthusiasts of nonparametric mean contrasts: Kruskal-Wallis test, kruskal.test() in R.**
- **For non-parametric tests of equality of variances in subgroups: Fligner test, fligner.test() in R.**

- **FactoMineR package. Use condes(X, num.var) and check $quali output. Profiling tool.**

3-1 ... EXPLORATION OF LARGE TABLES (DATABASE)

Test for means (Y ~A)
Normal Y
A I =2 -> t.test(Y ~A)
A I > 2 -> oneway.test(Y ~A)
Non normally distributed Y
A I =2 -> wilcoxon.test(Y ~A)
A I > 2 -> kruskal.test(Y ~A)

Test for dispersion in groups defined by A (Y ~A)
Normal Y
A I =2 -> var.test(Y ~A)
A I > 2 -> bartlett.test(Y ~A)
Non normally distributed Y
A I >=2 -> fligner.test(Y ~A)

Assume H0 is rejected for a test on means -> There is at least one group with a mean that is not the common mean.

pairwise.t.test(Y, A) – Y Normally distributed
pairwise.wilcoxon.test(Y,A) – Y any

3-1 ... EXPLORATION OF LARGE TABLES (DATABASE)

3-1.4 Global association between qualitative resp. with J levels and an explicative var with I levels

- Statistically related to the analysis of contingency tables.
- Habitual Convention: response variable in columns and rows explanatory.
- You can make tables 1, 2 and more dimensions in R, but we focus on the current paragraph and therefore 2 dimensional tables.
- Total observations concept (n), univariate row total ($n_{i\bullet}$), univariate columns total ($n_{\bullet j}$).
- Marginal rows probability ($p_{i\bullet}$), marginal columns ($p_{\bullet j}$), bivariate marginal (p_{ij}).

$$\begin{array}{cccc} & 1 \quad \cdots \quad J & \\ \begin{matrix} 1 \\ \vdots \\ I \end{matrix} & \begin{bmatrix} & \vdots & \\ \cdots & n_{ij} & \cdots \\ & \vdots & \end{bmatrix} & \begin{matrix} n_{1\bullet} \\ n_{i\bullet} \\ n_{I\bullet} \end{matrix} \\ & n_{\bullet 1} \quad n_{\bullet j} \quad n_{\bullet J} & n \end{array}$$

$$p_{ij} = \frac{n_{ij}}{n} \qquad p_{i\bullet} = \frac{n_{i\bullet}}{n} \qquad p_{\bullet j} = \frac{n_{\bullet j}}{n}$$

B~A

3-1 ... EXPLORATION OF LARGE TABLES (DATABASE)

Absence of association contrast:

$$H_0: \quad n_{ij} = n \cdot p_{i\bullet} \cdot p_{\bullet j}$$

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - n \cdot p_{i\bullet} \cdot p_{\bullet j}\right)^2}{n \cdot p_{i\bullet} \cdot p_{\bullet j}} \approx \chi^2_{(I-1)(J-1)} \rightarrow P(\chi^2_{(I-1)(J-1)} > discrepancy) = pvalue$$

- If the number of observations in each cell is nonzero (and best practice as more than 5), the null hypothesis of no association between the explanatory variable and the response variable, which is equivalent at independence between rows and columns, can be contrasted by **Pearson statistic χ^2,** whose ***reference distribution*** is a Shi-squared: could ask for the calculation of **p value** of the null hypothesis, if the p value is less than $\alpha = 0.05$ then there is evidence to reject the null hypothesis and if the p-value is above the threshold $\alpha = 0.05$ then there is evidence to accept the null hypothesis.
- **FactoMineR package. Use catdes(X, num.var) and check $test.chi2 output**

```
> chisq.test(table(edad,residencia))
        Pearson's Chi-squared test data:  table(edad, residencia)
X-squared = 6.4, df = 6, p-value = 0.3799
> attributes(chisq.test(table(edad,residencia)))
$names
[1]  "statistic" "parameter" "p.value"    "method"     "data.name"  "observed"
"expected"  "residuals"
```

3-1 ... EXPLORATION OF LARGE TABLES (DATABASE)

- For each variable with qualitative treatment (well to be a factor initially or by discretization of the original numerical variable) can calculate the p value of the hypothesis of no association between levels of this and the J levels of the response variable.

- Qualitative variables can be listed sorted by p value low to high.
- **FactoMineR package. Use catdes(X, num.var) and check $test.chi2 output**

3-1 ... EXPLORATION OF LARGE TABLES (DATABASE)

3-1.5 Association between quantitative resp. and every level of an explanatory with I levels

Given an explanatory with I levels, the issue will be detect the level with the highest association with the response, that is, those whose null hypothesis of no association has a probability (p value) smaller and thus points more significantly to the rejection of the null hypothesis.

- If for each qualitative explanatory variable are created virtually as many levels as auxiliary variables, so that for k = 1 to I, the null hypothesis is postulated "no difference in the population mean of response between the groups defined by the level k other levels ".

- It would make as many contrasts as detailed in paragraph 3-1.3**¡Error! No se encuentra el origen de la referencia.** as levels I, well now the variable that defines the group is dichotomous ***level k against all levels less k*** and therefore better to use a specialization of the contrast of the population mean of two subpopulations acceptably normal by t-test for unpaired subsamples. Be:
  - $\bar{y}_k$ the mean response in the subset defined by the category (level k) the treated explanatory variable.
  - $\bar{y}$ mean $\left(\bar{y} = \sum_{l=1}^{n} \frac{y_l}{n} \right)$ and s the sample standard deviation of the response $\left(s^2 = \sum_{l=1}^{n} \frac{(y_l - \bar{y})^2}{n-1} \right)$.
  - n_k number of observations belonging to group k and n number of observations.

3-1 ... EXPLORATION OF LARGE TABLES (DATABASE)

The contrast has described a statistical description:

$$H_0: \quad \mu_k = \mu \quad \text{for} \quad k=1 \ldots l$$

$$t = \frac{(\bar{y}_k - \bar{y})}{\sqrt{\left(1 - \frac{n_k}{n}\right) \cdot \frac{s^2}{n_k}}} \approx Student\ t_{n-1}$$

The computational procedure is:

- For each of the explanatory variables with qualitative treatment and for each level k of the explanatory calculate the p value of the null hypothesis of equal means in response according to the groups defined by level k and all levels.
- The result is a vector of lists, with many lists as explanatory variables within each list and many had p values as explanatory levels. They agree to give a matrix structure that allows sort lists low to high p value. Detects modalities (or levels) discriminant, that is, very informative to characterize the response.
- **FactoMineR package. Use condes(X, num.var) and check $category output**

3-1 ... EXPLORATION OF LARGE TABLES (DATABASE)

3-1.6 Associating each J qualitative response level and each I explanatory response levels

It can be statistically quantified using the normal approximation to the contrast of clusters resulting contingency table defining groups based on rows (k vs. rest) and groups of columns (the l vs. rest) of the response. Arguably following a similar reasoning to the previous paragraph but in the context of contingency tables.

For each level l=1..J of the response determine what level of explanatory k is more associated with k=1..I, by calculating the p value of:

$$\boxed{H_0: \quad p_{l/k} = p_{\bullet l} \quad \forall k, \forall l} \quad \text{where } p_{l/k} = \frac{n_{kl}}{n_{k\bullet}} \text{ and } X^2 = \sum_{i=k\,vs\,rest}\ \sum_{j=l\,rest} \frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}} \approx \chi^2_{(2-1)\cdot(2-1)} = \chi^2_1,$$

therefore its square root is a standard normal and specializing summation:

$$\frac{n_{kl}}{n_{k\bullet}} \approx N\left(\frac{n_{\bullet l}}{n}, \left(1 - p_{k\bullet}\right)\frac{p_{\bullet l}\left(1 - p_{\bullet l}\right)}{n_{k\bullet}}\right) \text{ and then, } z_{kl} = \frac{\dfrac{n_{kl}}{n_{k\bullet}} - \dfrac{n_{\bullet l}}{n}}{\sqrt{\left(1 - p_{k\bullet}\right)\dfrac{p_{\bullet l}\left(1 - p_{\bullet l}\right)}{n_{k\bullet}}}} \approx N(0,1)$$

3-1 ... EXPLORATION OF LARGE TABLES (DATABASE)

The result of the calculated p value results (a possibility) a matrix structure with many rows and columns as the explanatory levels (I) and the response (J) and containing the p value of level i of the explanatory ratio does not affect the response of the response level j. If they said matrix sorts by the first dimension discriminates the most explanatory categories for each category of the response appear.

- **FactoMineR package. Use catdes(X, num.var) and check $category output**

Now it will proceed to the previous paragraphs illustrate with *CBD Prizing* example.

3-2. CBD PRIZING EXAMPLE: ACCEPTANCE OF TRAFFIC AND MOVEMENT MEASURES AIMED AT REDUCING POLLUTANTS EMISSIONS

Determine a decision rule for accepting traffic and circulation measures aimed at reducing emissions (could come from a survey about the congestion charge applied to central London).

The hypothetical function score should classify individuals into three zones: green (light acceptance), orange (Near doubt) and red (light rejection).

Determine the percentage of acceptances, the percentage of actual rejections who have been labeled as clear acceptance and actual percentage of acceptances who have been labeled a clear rejection.

It features a sample of individuals that has collected information on socioeconomic characteristics and review. It has a data file and the following metadata:

- 99999999 value in the continuous variables indicating missing value.

3-2 *CBD PRIZING* EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

- *data dictionary* or features of the data matrix:

dictamen (Real review) (response variable)
- *1 positive / acceptance*
- *2 negative / rejection*

anys.feina (seniority) (years)

habitatge (Dwelling)
- *1 rental*
- *2 writing published*
- *3 private contract*
- *4 ignores contract*
- *5 parents*
- *6 others*

edat (Age)

estat.civil (marital status)
- *1 single*
- *2 married*
- *3 widower*
- *4 separate*
- *5 divorced*

tipus.feina (type of work)
- *1 empleado fijo*
- *2 empleado temporal*
- *3 autonomo*
- *4 otros*

despeses (Expenditure) (thousands €)

ingressos (Income) (thousands €)

patrimoni (Heritage) (thousands €)

carrega.patr (Heritage carryforwards) (thousands €)

import.assoc (Loans requested) (thousands €)

Plazo del préstamo más largo (Longer loan term) (months)

preu.final (Value of assets financed) (thousands €)

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

3-2.1 Global debugging more categorization of quantitative variables

```
opina <- read.table("kioto.txt",header=T,sep='\t',na.string='99999999')
```

```
summary(opina)
```

```
> summary(opina)
    dictamen       anys.feina       habitatge          plas            edat        estat.civil
 Min.   :0.000   Min.   : 0.000   Min.   :0.000   Min.   : 6.00   Min.   :18.00   Min.   :0.000
 1st Qu.:1.000   1st Qu.: 2.000   1st Qu.:2.000   1st Qu.:36.00   1st Qu.:28.00   1st Qu.:2.000
 Median :1.000   Median : 5.000   Median :2.000   Median :48.00   Median :36.00   Median :2.000
 Mean   :1.281   Mean   : 7.987   Mean   :2.657   Mean   :46.44   Mean   :37.08   Mean   :1.879
 3rd Qu.:2.000   3rd Qu.:12.000   3rd Qu.:4.000   3rd Qu.:60.00   3rd Qu.:45.00   3rd Qu.:2.000
 Max.   :2.000   Max.   :48.000   Max.   :6.000   Max.   :72.00   Max.   :68.00   Max.   :5.000

   registres      tipus.feina       despeses        ingressos       patrimoni       carrega.patr
 Min.   :1.000   Min.   :0.000   Min.   : 35.00   Min.   :  0.0   Min.   :     0   Min.   :    0.0
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 35.00   1st Qu.: 80.0   1st Qu.:     0   1st Qu.:    0.0
 Median :1.000   Median :1.000   Median : 51.00   Median :120.0   Median :  3000   Median :    0.0
 Mean   :1.174   Mean   :1.676   Mean   : 55.57   Mean   :130.6   Mean   :  5403   Mean   :  342.9
 3rd Qu.:1.000   3rd Qu.:3.000   3rd Qu.: 72.00   3rd Qu.:165.0   3rd Qu.:  6000   3rd Qu.:    0.0
 Max.   :2.000   Max.   :4.000   Max.   :180.00   Max.   :959.0   Max.   :300000   Max.   :30000.0
                                                  NA's   : 34.0   NA's   :    47   NA's   :   18.0
  import.assoc     preu.final
 Min.   : 100   Min.   :  105
 1st Qu.: 700   1st Qu.: 1118
 Median :1000   Median : 1400
 Mean   :1039   Mean   : 1463
 3rd Qu.:1300   3rd Qu.: 1692
 Max.   :5000   Max.   :11140
```

save.image("opina_raw.RData")

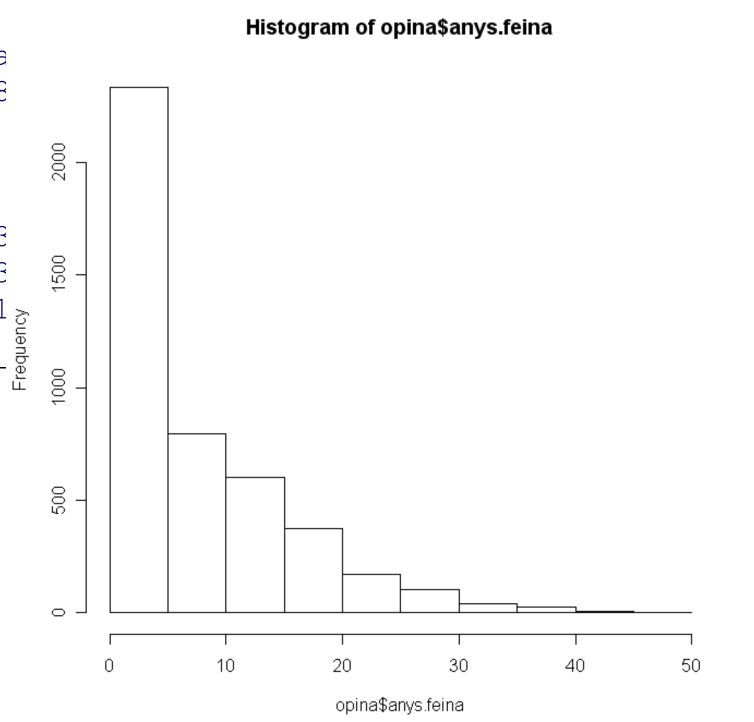
3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

DE between uni and bivariate categorical

```
> # dictamen: obs 3310 is 0, I delete it!
> opina <-opina[opina$dictamen!=0,]
> table(opina$dictamen)
   1    2
3200 1254
> opina$f.dictamen <- factor(opina$dictamen,levels=c("2","1"),labels=c("rebutja","accepta"))
> summary(opina$f.dictamen)
rebutja accepta
   1254    3200
```

Refined and discretization of "anys.feina"

```
> summary(opina$anys.feina)
   Min. 1st Qu.  Median    Mea
  0.000   2.000   5.000   7.98
> table(opina$anys.feina)

  0   1   2   3   4   5   6                                              17  18  19  20  21  22  23  24
535 510 454 336 233 266 181 13                                           56  65  39 151  23  41  26  19
 25  26  27  28  29  30  31  3                                           42  43  45  47  48
 62  14  14  14  11  49  10  1                                            1   2   3   1   1
> hist(opina$anys.feina)
```

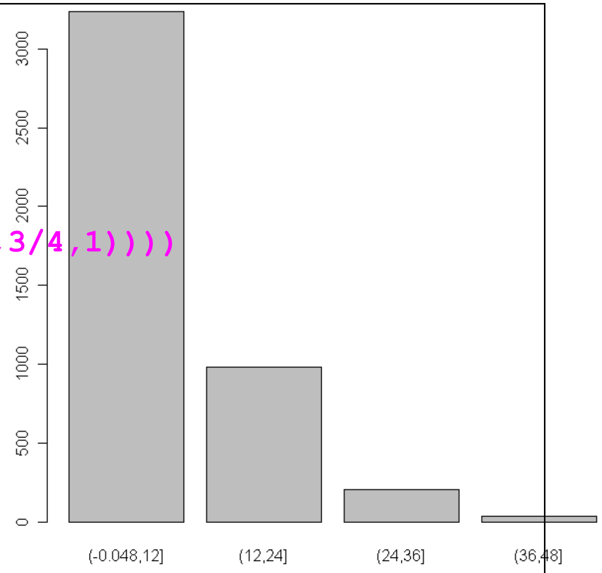


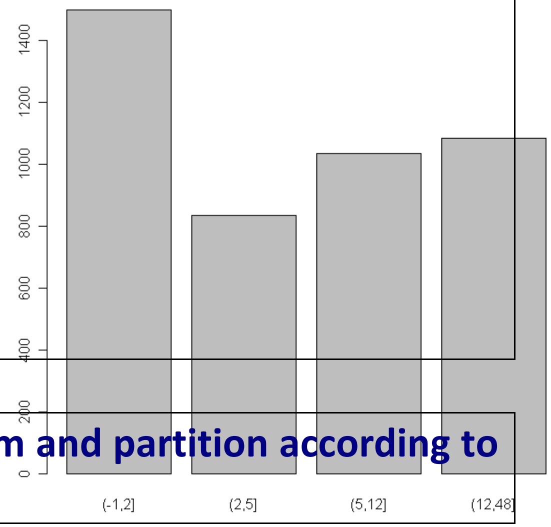
Exponential decay of difficult
Discretization i

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

Discretization options

```
> opina$f.afei <- factor(cut(anys.feina, 4))   # In 4 intervals
> table(opina$f.afei)

(-0.048,12]     (12,24]     (24,36]     (36,48]
       3235         979         207          33
> # Assistant to find the best grouping (in obs.nb.)
> opina$aux <- factor(cut(opina$anys.feina, quantile(anys.feina,c(0,1/4,2/4,3/4,1))))
> tapply(opina$anys.feina,opina$aux,median)
  (0,2]   (2,5]  (5,12] (12,48]
      1       4       9      18
> table(opina$aux)

  (0,2]   (2,5]  (5,12] (12,48]
    964     835    1034    1086
> sum(table(opina$aux))
[1] 3919
> opina$aux <- factor(cut(opina$anys.feina, breaks=c(-1,2,5,12,48)))
> tapply(opina$anys.feina,opina$aux,median)
 (-1,2]   (2,5]  (5,12] (12,48]
      1       4       9      18
> table(opina$aux)

 (-1,2]   (2,5]  (5,12] (12,48]
   1499     835    1034    1086
> sum(table(opina$aux))
[1] 4454
>
```

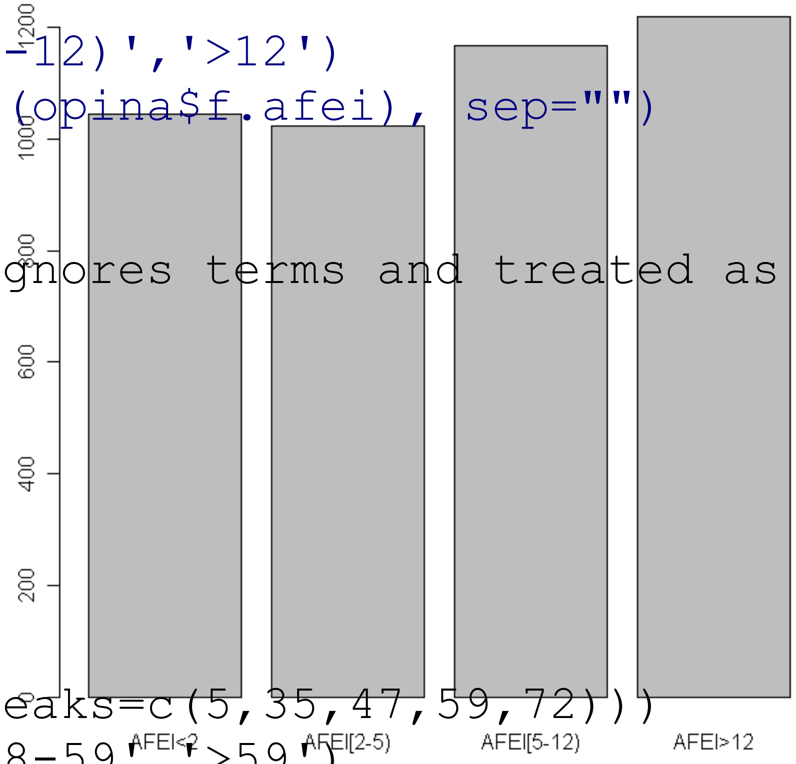




Tip: Depending on user specified intervals (with meaning), after observing histogram and partition according to quantiles.

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

```
opina$f.afei <- factor(cut(opina$anys.feina, breaks=c(-1,1.99,4.99,11.99,48)))
levels(opina$f.afei)<- c('<2','[2-5)','[5-12)','>12')
levels(opina$f.afei)<-paste("AFEI",levels(opina$f.afei), sep="")
summary(opina$f.afei)

# Housing: There are 6 missing. Group 4 ignores terms and treated as missing: 20 people
summary(opina$habitatge)
plot(opina$habitatge)

# Term
hist(opina$plas)
summary(opina$plas)
opina$f.plas <- factor(cut(opina$plas, breaks=c(5,35,47,59,72)))
levels(opina$f.plas)<- c('<36','36-47','48-59','>59')
levels(opina$f.plas) <- paste("PLAÇ",levels(opina$f.plas), sep="")

# Age
hist(opina$edat)
summary(opina$edat)
opina$f.edat<- factor(cut(opina$edat, breaks=c(17,28,36,45,68)))
levels(opina$f.edat) <- paste("EDAT",levels(opina$f.edat), sep="")
```



3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

```
# Marital Status: There is a missing (3320). One separated and divorced
in a group

summary(opina$estat.civil)
table(opina$estat.civil)
opina[opina$estat.civil==0,]
opina$f.eciv <- opina$estat.civil
opina$f.eciv[opina$estat.civil==5]<- 4
opina$f.eciv                                                         <-
factor(opina$f.eciv,levels=c("1","2","3","4"),labels=c("ECIV.solter"
,"ECIV.casat","ECIV.vidu","ECIV.sepdiv"))
table(opina$f.eciv)

# Records: what is it? Delete a variable undocumented
opina$registres <- NULL

# Type of work
summary(opina$tipus.feina)
table(opina$tipus.feina)
opina$f.tfei                                                         <-
factor(opina$tipus.feina,levels=c("1","2","3","4"),labels=c("TFEI.fi
x","TFEI.tmp","TFEI.auto","TFEI.altr"))
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

```
# Now come the quintessential covariates
# Expenditure: without missing. There are more high cost outliers in the group
of rejection
summary(opina$despeses)
table(opina$despeses)
hist(opina$despeses)
boxplot(opina$despeses)
tapply(opina$despeses,opina$f.dictamen,mean)
tapply(opina$despeses,opina$f.dictamen,median)
plot(opina$despeses~opina$f.dictamen)

# To find the best clustering (in obs.nb.)
opina$f.desp<- factor(cut(opina$despeses, breaks=c(34,35,45,60,75,90,180)))
levels(opina$f.desp) <- paste("DESP",levels(opina$f.desp), sep="")
summary(opina$f.desp)

# Income: without missing. There is 1 outlier (807) high income group rejection
(has many assets but little income)

summary(opina$ingressos)
table(opina$ingressos)
par(mfrow=c(1,2))
hist(opina$ingressos)
```

```
boxplot(opina$ingressos)
tapply(opina$ingressos,opina$f.dictamen,mean)
tapply(opina$ingressos,opina$f.dictamen,median)
plot(opina$dictamen,opina$ingressos)
text(opina$dictamen,opina$ingressos,labels=row.names(opina))
opina$f.ingr<- factor(cut(opina$ingressos, breaks=c(-1,50,100,150,250,960)))
levels(opina$f.ingr) <- paste("INGR",levels(opina$f.ingr), sep="")
summary(opina$f.ingr)

# Heritage: 47 missing. There is 1 outlier (807) high income group rejection
(has many assets but little income)

summary(opina$patrimoni)
table(opina$patrimoni)
par(mfrow=c(1,2))
hist(opina$patrimoni)
boxplot(opina$patrimoni)
tapply(opina$patrimoni,opina$f.dictamen,mean)
tapply(opina$patrimoni,opina$f.dictamen,median)
plot(opina$patrimoni~opina$f.dictamen)
opina$f.patr<-            factor(cut(opina$patrimoni,            breaks=c(-
1,500,4000,8000,15000,300000)))
levels(opina$f.patr) <- c('<0.5','0.5-4','4-8','8-15','>15')
levels(opina$f.patr) <- paste("PATR",levels(opina$f.patr), sep="")
summary(opina$f.patr)
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

```
# carrega.patr: 18 missing. There is 1 outlier (807) high income group rejection
(has many assets but little income)

summary(opina$carrega.patr)
table(opina$carrega.patr)
par(mfrow=c(1,2))
hist(opina$carrega.patr)
boxplot(opina$carrega.patr)
opina$f.carrpatr<- factor(cut(opina$carrega.patr, breaks=c(-1,0,5000,30000)))
levels(opina$f.carrpatr) <- c('Nul','Mig','Alt')
levels(opina$f.carrpatr) <- paste("CARRP",levels(opina$f.carrpatr), sep="")
summary(opina$f.carrpatr)

# import.assoc: 0 missing.
summary(opina$import.assoc)
table(opina$import.assoc)
par(mfrow=c(1,2))
hist(opina$import.assoc)
boxplot(opina$import.assoc)
opina$f.import<-                              factor(cut(opina$import.assoc,
breaks=c(0,700,1000,1300,5000)))
levels(opina$f.import) <- c('<.7','.7-1','1-1.3','1.3-5')
levels(opina$f.import) <- paste("IMP",levels(opina$f.import), sep="")
summary(opina$f.import)
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

```
# preu.final: 0 missing.
```

```
summary(opina$preu.final)
table(opina$preu.final)
par(mfrow=c(1,2))
hist(opina$preu.final)
boxplot(opina$preu.final)
opina$f.preu <-
    factor(cut(opina$preu.final,breaks=c(0,1100,1400,1700,12000)))
levels(opina$f.preu) <- c('<1.1','1.1-1.4','1.4-1.7','1.7+')
levels(opina$f.preu) <- paste("PREU",levels(opina$f.preu), sep="")
summary(opina$f.preu)
> opina$registres<-NULL
> summary(opina)
    dictamen        anys.feina        habitatge          plas             edat         estat.civil
 Min.   :1.000   Min.   : 0.000   Min.   :0.000   Min.   : 6.00   Min.   :18.00   Min.   :0.000
 1st Qu.:1.000   1st Qu.: 2.000   1st Qu.:2.000   1st Qu.:36.00   1st Qu.:28.00   1st Qu.:2.000
 Median :1.000   Median : 5.000   Median :2.000   Median :48.00   Median :36.00   Median :2.000
 Mean   :1.282   Mean   : 7.987   Mean   :2.657   Mean   :46.44   Mean   :37.08   Mean   :1.879
 3rd Qu.:2.000   3rd Qu.:12.000   3rd Qu.:4.000   3rd Qu.:60.00   3rd Qu.:45.00   3rd Qu.:2.000
 Max.   :2.000   Max.   :48.000   Max.   :6.000   Max.   :72.00   Max.   :68.00   Max.   :5.000
 tipus.feina        despeses        ingressos        patrimoni       carrega.patr    import.assoc
 Min.   :0.000   Min.   : 35.00   Min.   :  0.0   Min.   :     0   Min.   :    0   Min.   : 100
 1st Qu.:1.000   1st Qu.: 35.00   1st Qu.: 80.0   1st Qu.:     0   1st Qu.:    0   1st Qu.: 700
 Median :1.000   Median : 51.00   Median :120.0   Median :  3000   Median :    0   Median :1000
 Mean   :1.676   Mean   : 55.57   Mean   :130.6   Mean   :  5404   Mean   :  343   Mean   :1039
 3rd Qu.:3.000   3rd Qu.: 72.00   3rd Qu.:165.0   3rd Qu.:  6000   3rd Qu.:    0   3rd Qu.:1300
 Max.   :4.000   Max.   :180.00   Max.   :959.0   Max.   :300000   Max.   :30000   Max.   :5000
                                  NA's   : 34.0   NA's   :    47   NA's   :   18
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

```
  preu.final       f.dictamen          f.afei             f.habi             f.plas
 Min.    :  105   rebutja:1254   AFEI<2     :1045   HAB.lloguer: 973   PLAÇ<36  : 673
 1st Qu.: 1117   accepta:3200   AFEI[2-5) :1023   HAB.scrpu  :2107   PLAÇ36-47: 971
 Median : 1400                   AFEI[5-12):1167   HAB.contpri: 246   PLAÇ48-59: 877
 Mean   : 1463                   AFEI>12    :1219   HAB.pares  : 783   PLAÇ>59  :1933
 3rd Qu.: 1692                                      HAB.altres : 319
 Max.   :11140                                      NA's       :  26

         f.edat              f.eciv             f.tfei              f.desp               f.ingr
 EDAT(17,28]:1196   ECIV.solter: 977   TFEI.fix :2805   DESP(34,35] :1211   INGR(-1,50]  : 496
 EDAT(28,36]:1160   ECIV.casat :3241   TFEI.tmp : 452   DESP(35,45] : 915   INGR(50,100] :1222
 EDAT(36,45]:1030   ECIV.vidu  :  67   TFEI.auto:1024   DESP(45,60] :1079   INGR(100,150]:1364
 EDAT(45,68]:1068   ECIV.sepdiv: 168   TFEI.altr: 171   DESP(60,75] : 743   INGR(150,250]:1051
                    NA's       :   1   NA's     :   2   DESP(75,90] : 349   INGR(250,960]: 287
                                                        DESP(90,180]: 157   NA's         :  34

       f.patr          f.carrpatr       f.import          f.preu
 PATR<0.5 :1636   CARRPNul:3669   IMP<.7   :1165   PREU<1.1    :1065
 PATR0.5-4:1126   CARRPMig: 739   IMP.7-1  :1294   PREU1.1-1.4:1165
 PATR4-8  : 925   CARRPAlt:  28   IMP1-1.3: 948   PREU1.4-1.7:1172
 PATR8-15 : 443   NA's    :  18   IMP1.3-5:1047   PREU1.7+    :1052
 PATR>15  : 277
 NA's     :  47
```

- **FactoMineR package. Use catdes(X, num.var) and condes(X, num.var) after data imputation (no NAs).**
- **Imputation tools: missMDA package. For numeric variables imputePCA(X_num) method and imputeMCA(X_factors) for qualitative variables. Validation of imputation results is mandatory.**

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

3-2.2 Global association: quantitative resp. and explicative var with I levels (factor)

Change of roles: quantitative response variables with continuous basis and as explanatory variable the actual review. The addressed question is whether there is a difference according to the actual opinion (opinion of acceptance or rejection) in seniority, age, expenses, income, assets, equity load, amount of credit, longer term loan and the value of items loans.

```
attach(opina)
> attributes(opina)$names
 [1] "dictamen"     "anys.feina"   "habitatge"    "plas"         "edat"         "estat.civil"
 [7] "tipus.feina"  "despeses"     "ingressos"    "patrimoni"    "carrega.patr" "import.assoc"
[13] "preu.final"   "f.dictamen"   "f.afei"       "f.habi"       "f.plas"       "f.edat"
[19] "f.eciv"       "f.tfei"       "f.desp"       "f.ingr"       "f.patr"       "f.carrpatr"
[25] "f.import"     "f.preu"
# Variables with digital processing: creating list of variables
attach(opina)
vacs <-
list(opina[,2],opina[,4],opina[,5],opina[,8],opina[,9],opina[,10],opina[,11],
opina[,12],opina[,13])
nomsvacs<-names(opina)[c(2,4,5,8:13)]);nomsvacs
pvalvacs <-NULL
# Object initialization loop pvalvacs list: p values contrast null hypothesis
for (i in 1:9 ) {pvalvacs[ i ] <- oneway.test(vacs[[i]]~f.dictamen,var.equal
= FALSE)$p.value }
pvalvacs
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

```
# Change list formatting to 1-dimensional matrix in columns
pvalvacs <- matrix( pvalvacs )
pvalvacs
# Give name to the rows of the matrix pvalvacs, variables relating to
row.names(pvalvacs) <-
c("anys.feina","plas","edat","despeses","ingressos","patrimoni","carrega.patr","import.assoc","
preu.final")
pvalvacs
# Sort first matrix dimension: rows. Output: variables p-value less than major.
sort( pvalvacs[,1] )
> vacs <-
list(opina[,2],opina[,4],opina[,5],opina[,8],opina[,9],opina[,10],opina[,11],opina[,12],opina[,13])
> pvalvacs <-NULL
> for (i in 1:9 ) { pvalvacs[ i ] <- oneway.test(vacs[[i]]~f.dictamen,var.equal = FALSE)$p.value }
> pvalvacs
[1] 3.378876e-89 5.213283e-13 6.032429e-11 7.113231e-02 2.264256e-46 5.429938e-14 5.068654e-01
[8] 2.677710e-21 5.012962e-01
> pvalvacs <- matrix( pvalvacs )
> pvalvacs
              [,1]
 [1,] 3.378876e-89
 [2,] 5.213283e-13
 [3,] 6.032429e-11
 [4,] 7.113231e-02
 [5,] 2.264256e-46
 [6,] 5.429938e-14
 [7,] 5.068654e-01
 [8,] 2.677710e-21
 [9,] 5.012962e-01
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

```
> row.names( pvalvacs ) <-
c("anys.feina","plas","edat","despeses","ingressos","patrimoni","carrega.patr","import.assoc","
preu.final")
> pvalvacs
                     [,1]
anys.feina   3.378876e-89
plas         5.213283e-13
edat         6.032429e-11
despeses     7.113231e-02
ingressos    2.264256e-46
patrimoni    5.429938e-14
carrega.patr 5.068654e-01
import.assoc 2.677710e-21
preu.final   5.012962e-01

> sort( pvalvacs[,1] )
  anys.feina    ingressos import.assoc    patrimoni         plas         edat     despeses
3.378876e-89 2.264256e-46 2.677710e-21 5.429938e-14 5.213283e-13 6.032429e-11 7.113231e-02
  preu.final carrega.patr
5.012962e-01 5.068654e-01

>nomsvacs<-names(opina)[c(2,4,5,8:13)]);nomsvacs

>library(FactoMineR)

> catdes(opina[,c("f.dictamen",nomsvacs)],num.var=1)

$quanti.var

                   Eta2      P-value

anys.feina   0.067804443 6.071892e-70

ingressos    0.043837874 5.628929e-45

import.assoc 0.023885476 3.261359e-25

plas         0.010125845 1.686207e-11
```

```
patrimoni    0.009420326 1.070367e-10
edat         0.009082235 1.854456e-10


$quanti
$quanti$rebutja
                 v.test Mean in category Overall mean sd in category
import.assoc  10.313197      1156.062998 1038.918276     535.454076
plas           6.714938        48.794258   46.438707      12.859222
edat          -6.359496        35.408293   37.080377      10.421689
patrimoni     -6.501447      3602.786062 5403.979351    8653.748988
ingressos    -14.053672       101.508104  130.564253      80.814576
anys.feina   -17.376225         4.586922    7.986753       6.115582
               Overall sd      p.value
import.assoc   474.492724 6.142358e-25
plas            14.653817 1.881465e-11
edat            10.983365 2.024164e-10
patrimoni    11573.118739 7.955097e-11
ingressos       86.367110 7.314150e-45
anys.feina       8.173388 1.249107e-67


$quanti$accepta
                 v.test Mean in category Overall mean sd in category
anys.feina    17.376225         9.319062    7.986753       8.486593
ingressos     13.890300       141.818267  130.564253      85.820705
patrimoni      6.452196      6104.474945 5403.979351   12455.914562
edat           6.359496        37.735625   37.080377      11.127476
```

```
plas          -6.714938        45.515625    46.438707       15.200547
import.assoc -10.313197       993.012187  1038.918276      439.922117
               Overall sd      p.value
anys.feina       8.173388 1.249107e-67
ingressos       86.367110 7.253072e-44
patrimoni    11573.118739 1.102407e-10
edat            10.983365 2.024164e-10
plas            14.653817 1.881465e-11
import.assoc   474.492724 6.142358e-25


attr(,"class")
[1] "catdes" "list "
>
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

- The final review (*f.dictamen*) is associated with all of said quantitative variables except the value of objects loan (*preu.final*) and the financial burden. The costs show the contrast p value of association with the final review of 7% (slightly higher confidence level of 5%), there is to be strict and in my opinion there is evidence to reject the null hypothesis of no association and therefore to say that there are differences in costs between the 2 groups of opinion.
- The **null hypothesis** asserts that the two focus groups (acceptance or rejection) mean values of the variables at work seniority, age, expenses, income, assets, credit amount and term of the loan longer have p values **are equal** lower than the typical threshold of 5% (0.05 in per unit) and therefore **the null hypotheses are rejected** and taken for **valid alternatives differences between the 2 groups**.
- The **null hypothesis** asserts that the two focus groups (acceptance or rejection) mean values of the financial burden and value objects **are equal** loans are given for valid by showing p values above the threshold of 5%.
- Attention: the file is purged and the specific numbers may differ if *R script* execution with other similar criterion debugging variable is repeated, the more the findings should not differ if debugging is robust and sensible. You can use the FactoMineR R package.

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

3-2.3 Global association between qualitative resp. with J levels and an explicative var with I levels

- The role of response variable in this case is set by f.dictamen (dichotomous, J = 2). Categorical explanatory variables are the original.
- Over those created by the process of discretization of numerical variables originally included.

```
> load("E:/LTLIDIA/LIDIA/MLTM-MCAID/raw material/CURS06-07/Tema 2/opina.RData")
> attach(opina)
> names(opina)
 [1] "dictamen"     "anys.feina"   "habitatge"    "plas"         "edat"         "estat.civil"
 [7] "tipus.feina"  "despeses"     "ingressos"    "patrimoni"    "carrega.patr" "import.assoc"
[13] "preu.final"   "f.dictamen"   "f.afei"       "f.habi"       "f.plas"       "f.edat"
[19] "f.eciv"       "f.tfei"       "f.desp"       "f.ingr"       "f.patr"       "f.carrpatr"
[25] "f.import"     "f.preu"
```

Manual procedure, untapped computational advantages of R

```
# seniority
summary(opina$f.afei)
# Hypothesis independence dictamen vs. f.afei rejected X-squared= 42.8832, df
= 3, p-value = 2.606e-09
chisq.test(opina$f.afei,opina$f.dictamen)
table(opina$f.afei,opina$f.dictamen)
```

```
# Housing: There are 6 missing. Group 4 ignores terms and treated as missing:
20 people
# Hypothesis independence dictamen vs. f.habi rejected (X-squared = 217.8476,
df = 4, p-value < 2.2e-16)
summary(opina$ f.habi)
chisq.test(opina$f.habi,opina$f.dictamen)
table(opina$f.habi,opina$f.dictamen)
```

… They continue with all factors, analyzing contingency tables

```
> table(opina$f.afei,opina$f.dictamen)

             rebutja accepta
  AFEI<2         512     533
  AFEI[2-5)      330     693
  AFEI[5-12)     263     904
  AFEI>12        149    1070

> table(opina$f.ingr,opina$f.dictamen)

                rebutja accepta
  INGR(-1,50]       279     217
  INGR(50,100]      441     781
  INGR(100,150]     289    1075
  INGR(150,250]     172     879
  INGR(250,960]      53     234
```

```
> table(opina$f.tfei,opina$f.dictamen)

            rebutja accepta
  TFEI.fix      580    2225
  TFEI.tmp      271     181
  TFEI.auto     333     691
  TFEI.altr      68     103

> table(opina$f.habi,opina$f.dictamen)

              rebutja accepta
  HAB.lloguer     388     585
  HAB.scrpu       390    1717
  HAB.contpri      84     162
  HAB.pares       233     550
  HAB.altres      146     173
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

Automated procedure using the computational advantages of R

```
> names(opina)
# [1] "dictamen"     "anys.feina"   "habitatge"    "plas"         "edat"         "estat.civil"
# [7] "tipus.feina"  "despeses"     "ingressos"    "patrimoni"    "carrega.patr" "import.assoc"
#[13] "preu.final"   "f.dictamen"   "f.afei"       "f.habi"       "f.plas"       "f.edat"
#[19] "f.eciv"       "f.tfei"       "f.desp"       "f.ingr"       "f.patr"       "f.carrpatr"
#[25] "f.import"     "f.preu"
```

```
vads <-
list(f.afei,f.habi,f.plas,f.edat,f.eciv,f.tfei,f.desp,f.ingr,f.patr,f.carrpatr,f.import,f.preu)
pvalvads <-NULL
for (i in 1:12 ) { pvalvads[ i ] <- chisq.test(vads[[i]],f.dictamen)$p.value }
pvalvads <- matrix( pvalvads )
row.names( pvalvads ) <-
c("f.afei","f.habi","f.plas","f.edat","f.eciv","f.tfei","f.desp","f.ingr","f.patr","f.carrpatr"
,"f.import","f.preu")
sort( pvalvads[,1] )
```

```
> sort( pvalvads[,1] )
      f.afei       f.ingr       f.tfei       f.habi       f.patr     f.import       f.plas
2.913760e-87 8.229399e-75 4.303683e-70 5.446068e-46 9.349481e-43 1.291256e-23 1.912465e-12
      f.eciv       f.edat       f.desp       f.preu   f.carrpatr
2.300075e-11 2.605645e-09 7.466957e-08 7.595702e-08 2.534421e-01
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

- All variables are associated with the review, unless the discretization of the financial burden, which is consistent with the analysis as a covariate in the financial burden.

- The final price as a covariate is not associated with the review: the null hypothesis of equal final price in the 2 opinion groups is accepted, however, when considering the discretization of the final price (f.preu) there are differences. There is an issue of outliers.

```
# Graphs of the four variables associated and discrepant preu.final
par(mfrow=c(1,2))
plot(opina$f.dictamen,opina$f.afei)
plot(opina$f.dictamen,opina$anys.feina)
par(mfrow=c(1,2))
plot(opina$f.dictamen,opina$f.ingr)
plot(opina$f.dictamen,opina$ingressos)
par(mfrow=c(1,2))
plot(opina$f.dictamen,opina$f.tfei)
plot(opina$f.dictamen,opina$f.habi)

par(mfrow=c(1,2))
plot(opina$f.dictamen,opina$f.preu)
plot(opina$f.dictamen,opina$preu.final)
```

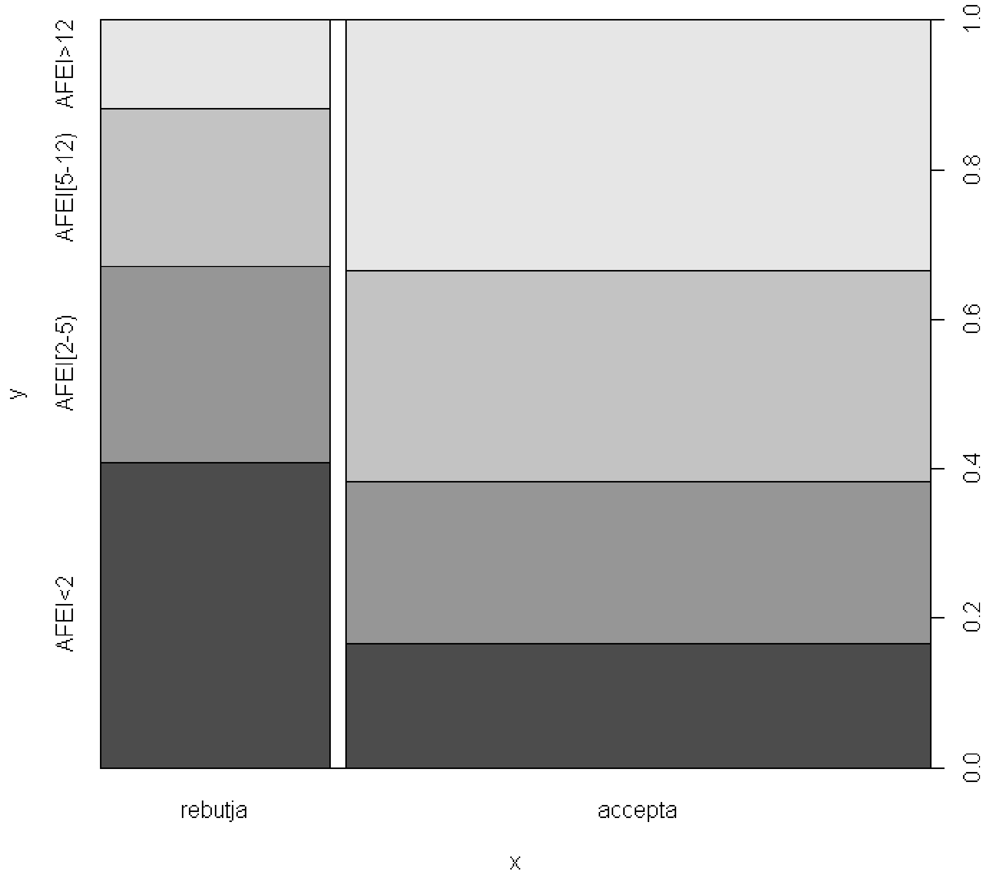
3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

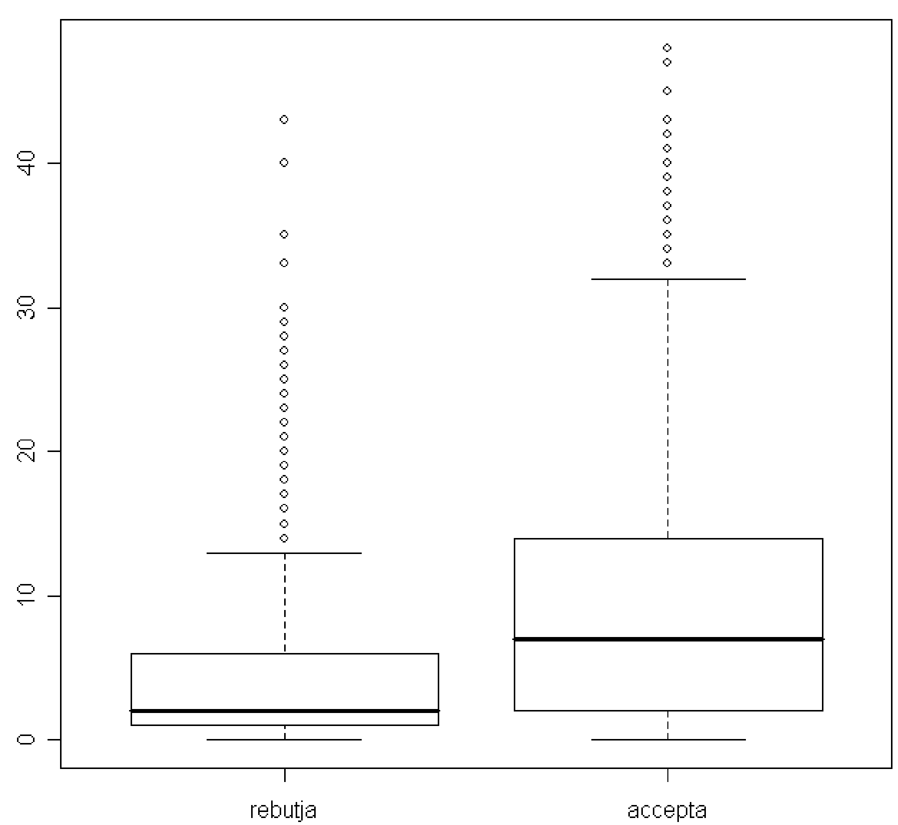
Numerical univariate descriptive analysis by focus opinion groups in numeric variables seniority, income and final price of items subject to loans:

```
> tapply(opina$anys.feina,opina$f.dictamen,summary)
$rebutja
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   2.000   4.587   6.000  43.000

$accepta
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   2.000   7.000   9.319  14.000  48.000

> tapply(opina$ingressos,opina$f.dictamen,summary)
$rebutja
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    0.0    58.0    90.0   101.5   136.0   959.0    20.0

$accepta
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    0.0    92.0   127.5   141.8   175.0   905.0    14.0

> tapply(opina$preu.final,opina$f.dictamen,summary)
$rebutja
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    105    1063    1423    1474    1728    6802

$accepta
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    125    1134    1400    1459    1678   11140
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

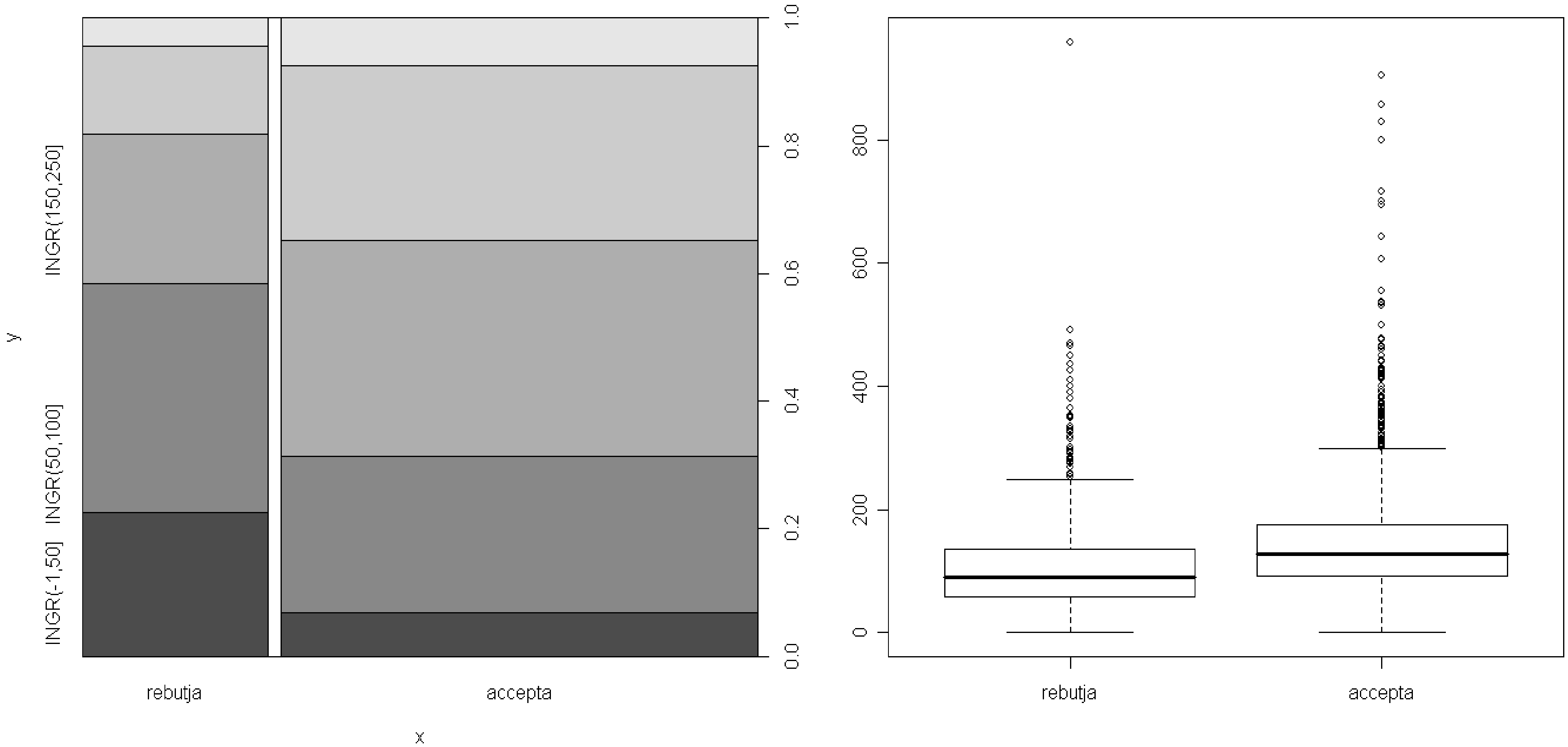
Graphic univariate descriptive analysis by focus opinion groups of varying seniority, income and final price of items subject to loans:





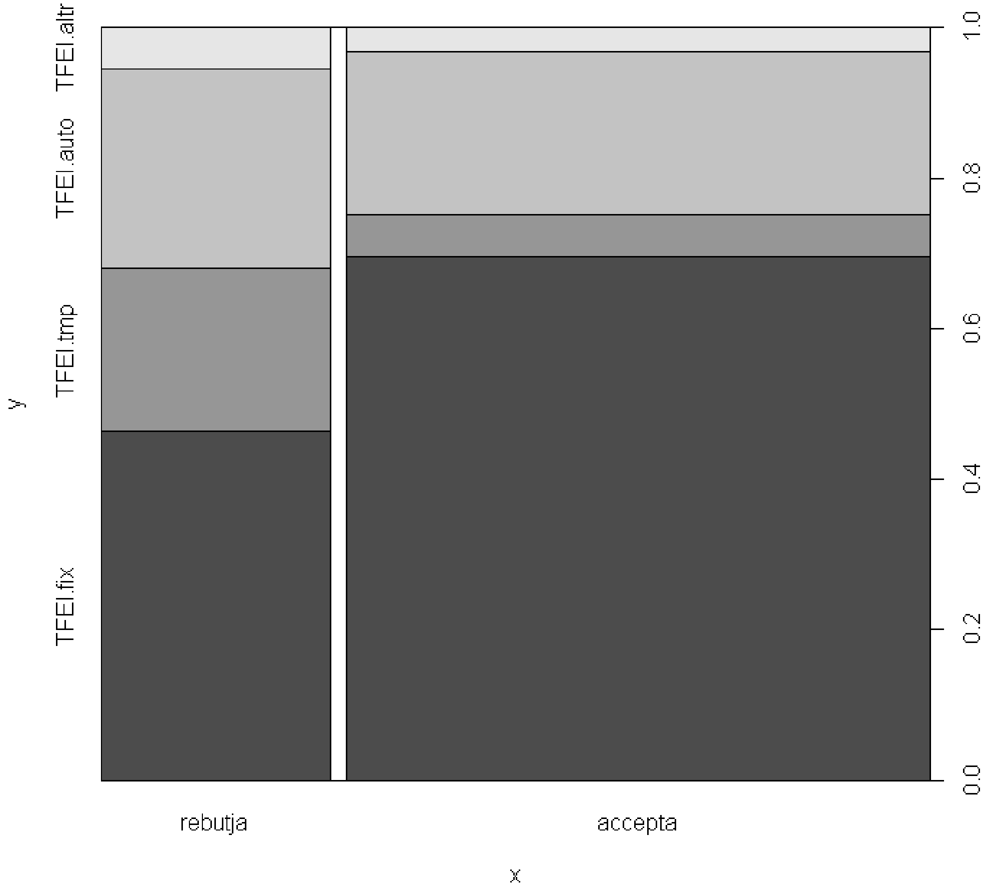
3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

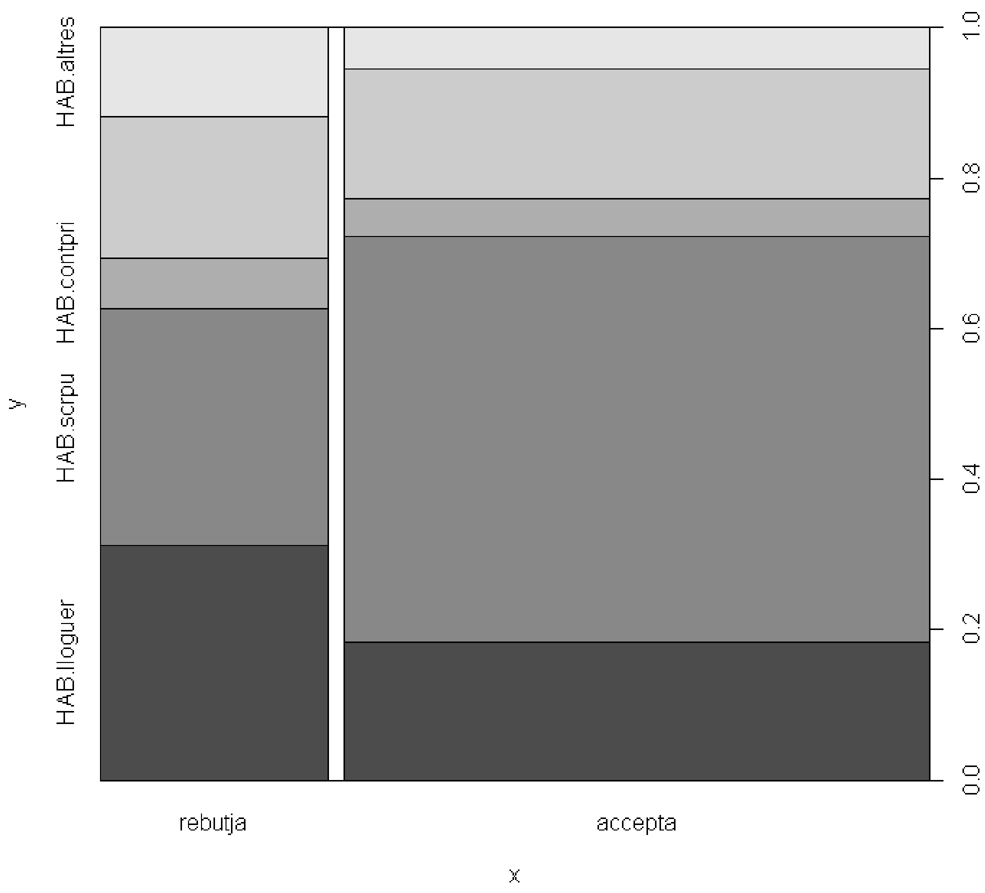
Graphic univariate descriptive analysis by focus groups of variable income:



3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

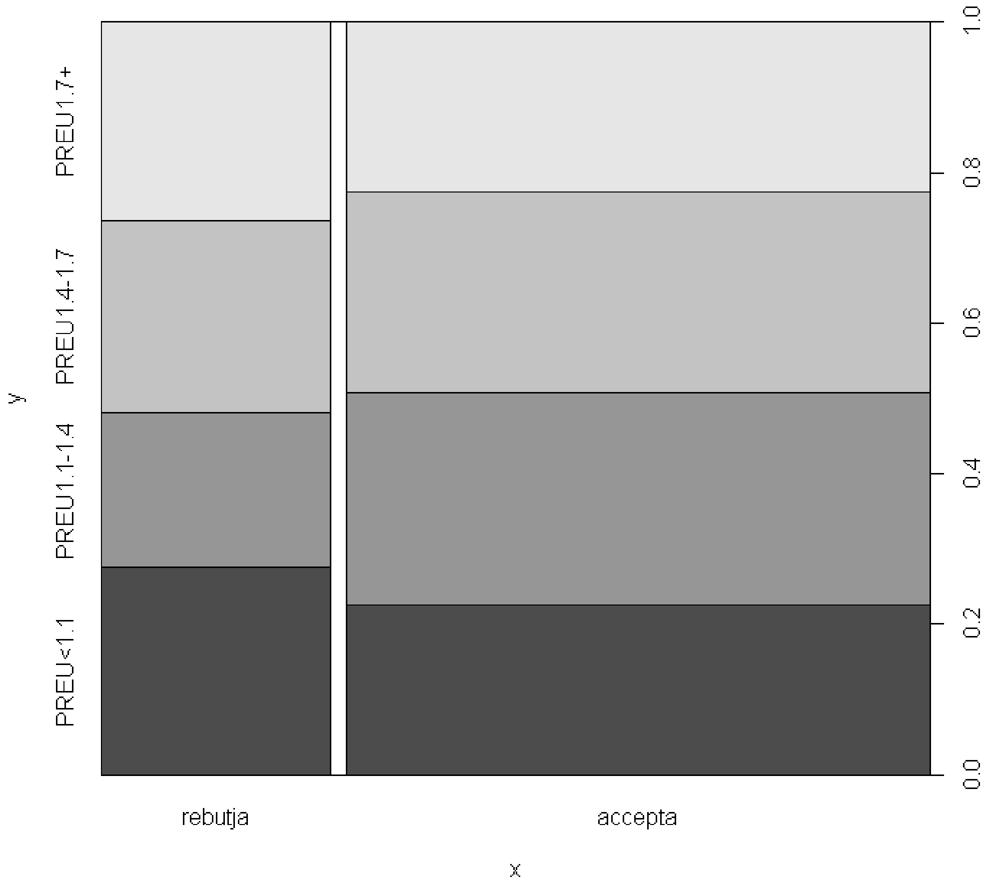
Graphic univariate descriptive analysis by focus opinion groups of the variable type of work and ownership regime residence:

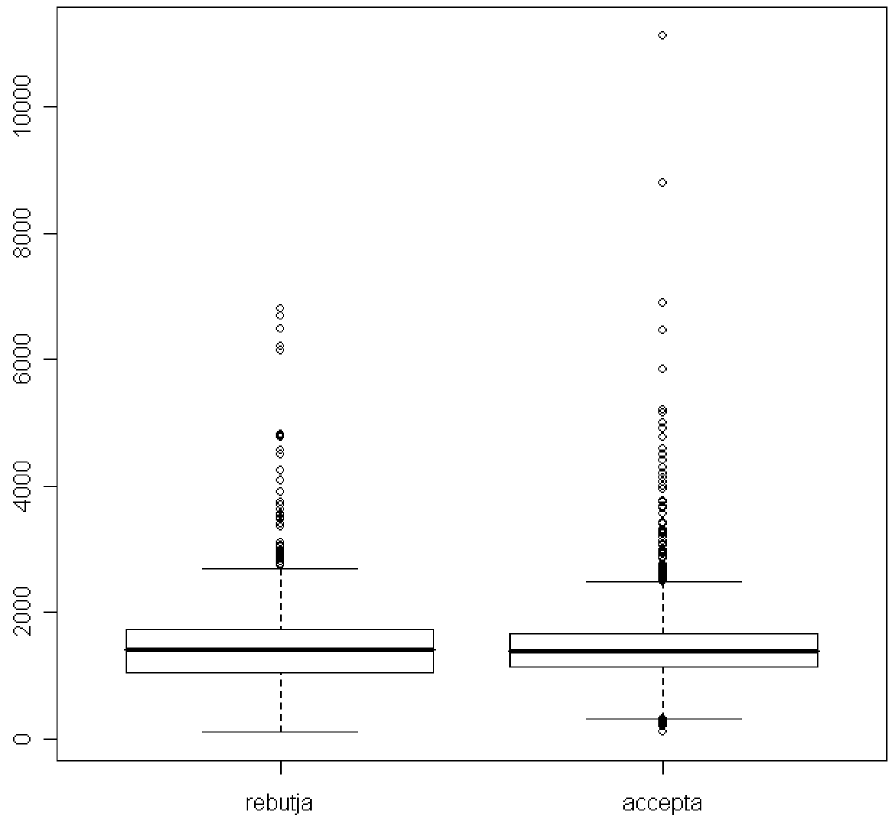




3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

Graphic univariate descriptive analysis by focus opinion groups final price of items subject to loans.:





3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

```
>library(FactoMineR)
> names(opina)
 [1] "dictamen"     "anys.feina"   "habitatge"    "plas"         "edat"         "estat.civil"
 [7] "tipus.feina"  "despeses"     "ingressos"    "patrimoni"    "carrega.patr" "import.assoc"
[13] "preu.final"   "f.dictamen"   "f.afei"       "f.habi"       "f.plas"       "f.edat"
[19] "f.eciv"       "f.tfei"       "f.desp"       "f.ingr"       "f.patr"       "f.carrpatr"
[25] "f.import"     "f.preu"
> nomsvads<-names(opina)[c(14:25)];nomsvads
 [1] "f.dictamen" "f.afei" "f.habi" "f.plas"     "f.edat"    "f.eciv"    "f.tfei"      "f.desp"
 [9] "f.ingr"     "f.patr"     "f.carrpatr" "f.import"
> catdes(opina[,c(nomsvads)],num.var=1)
$test.chi2
                p.value df
f.afei     2.913760e-87  3
f.ingr     7.456113e-77  5
f.tfei     4.314139e-70  4
f.habi     2.760016e-46  5
f.patr     9.013802e-43  5
f.import   1.291256e-23  3
f.plas     1.912465e-12  3
f.eciv     8.783901e-11  4
f.edat     2.605645e-09  3
f.desp     7.466957e-08  5
f.carrpatr 1.629441e-04  3

$category  .... and much more
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

Library FactoMineR – profiling and feature selection – CSI tools
Y numeric target

condes(df, num.var)

- **$quanti – Numeric explanatory vars that are globally associated to Y**
- **$quali - Qualitative explanatory vars that are globally associated to Y**
- **$category – Those categories belonging to explanatory factors whose mean is not equal to grand mean**

B – Factor target

catdes(df, num.var)

- **$ test.chi2 – Qualitative vars that are globally associated to B target**
- **$quanti.var - Quantitative vars that are globally associated to B target**
- **$quanti - Means on numèric vars that differ from global mean at each level of B factor**
- **$category – Specific relation between I factor from a qualitative factor and each level of B factor**

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

3-2.4 Association for levels resp. with J levels and qualitative factors

Helper function to calculate the p value of the contrast statistic for each level of an explanatory factor in files (I level) and each of the response levels in columns:

```
# Create a function to calculate z for each cell of the table
# original (for var. Categorically a table): response columns
zij <- function( x, y ){
taula <- table( x, y);
# added table rows which is the first
dimension
n__ <- sum( taula );
ni_ <- apply(taula, 1, sum ) ;
pi_ <- apply(taula, 1, sum ) / n__;
# sum table columns is the second dimension
n_j <- apply( taula, 2, sum ) ;
p_j <- apply( taula, 2, sum ) / n__;
# Calculate the row profiles: nij/ni_  (annotate with p(j/i))
pf <- taula/(ni_);
pc <- taula/(n_j);
# Replicating the marginal univ column many times as rows: n_j/n__
pcI <- matrix( data=p_j, nrow=dim(pf)[1], ncol=dim(pf)[2], byrow=T );
# Numerator
dpf <- pf - pcI;
# Denominator
denom <- sqrt( ( (1-pi_)/(ni_) ) %*% t( p_j*(1-p_j) ) );
zij <- dpf / denom;
pzij <- 1-pnorm(zij);
list(perfilfila=pf,perfilcolumna=pc,p.fila=p_j,p.col=pi_, vtest=zij, pval=pzij )
}
```

```
# Output:
# One list with zij (vtest) and pvalues (pval) plus,
# p.col is marginal row: pi_=sum_j (nij)/n
# p.fila is marginal column: p_j=sum_i (nij)/n
# perfilcolumna are column percentages: nij/n_j
# perfilfila are row percentages: nij/ni_
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

```
> pvalvadij1 <- zij( f.afei, f.dictamen )$pval
> pvalvadij1
           y
x                rebutja      accepta
  AFEI<2     0.000000e+00 1.000000e+00
  AFEI[2-5)  4.420680e-04 9.995579e-01
  AFEI[5-12) 9.999997e-01 3.394110e-07
  AFEI>12    1.000000e+00 0.000000e+00
> pvalvadij2 <- zij( f.habi, f.dictamen )$pval
> pvalvadij2
             y
x                  rebutja      accepta
  HAB.lloguer 0.000000e+00 1.000000e+00
  HAB.scrpu   1.000000e+00 0.000000e+00
  HAB.contpri 1.393059e-02 9.860694e-01
  HAB.pares   1.172637e-01 8.827363e-01
  HAB.altres  1.201261e-13 1.000000e+00
…
> pvalvadij12 <- zij( f.preu, f.dictamen )$pval

> pvadij <-
rbind(pvalvadij1,pvalvadij2,pvalvadij3,pvalvadij4,pvalvadij5,pvalvadij6,pvalvadij7,pvalvadij8,pvalvadij
9,pvalvadij10,pvalvadij11,pvalvadij12)
> pvadij
                   rebutja      accepta
AFEI<2        0.000000e+00 1.000000e+00
AFEI[2-5)     4.420680e-04 9.995579e-01
AFEI[5-12)    9.999997e-01 3.394110e-07
AFEI>12       1.000000e+00 0.000000e+00
HAB.lloguer   0.000000e+00 1.000000e+00
…
PREU1.7+      3.996315e-03 9.960037e-01
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

You could automate more generically using iterative loop in R.

```
names(opina)
# [1] "dictamen"     "anys.feina"   "habitatge"    "plas"         "edat"         "estat.civil"
# [7] "tipus.feina"  "despeses"     "ingressos"    "patrimoni"    "carrega.patr" "import.assoc"
#[13] "preu.final"   "f.dictamen"   "f.afei"       "f.habi"       "f.plas"       "f.edat"
#[19] "f.eciv"       "f.tfei"       "f.desp"       "f.ingr"       "f.patr"       "f.carrpatr"
#[25] "f.import"     "f.preu"

vads <-
list(f.afei,f.habi,f.plas,f.edat,f.eciv,f.tfei,f.desp,f.ingr,f.patr,f.carrpatr,f.import,f.preu)
pvadkl <-NULL
for (i in 1:12 ) { pvadkl <- rbind(pvadkl, zij(vads[[i]],f.dictamen)$pval) }
pvadkl

#pvadkl <-
rbind(pvalvads[1],pvalvads[2],pvalvads[3],pvalvads[4],pvalvads[5],pvalvads[6],pvalvads[7],pvalv
ads[8],pvalvads[9],pvalvads[10],pvalvads[11],pvalvads[12])
pvadkl

sort(pvadkl[,1])
sort(pvadkl[,2])
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

```
# Acceptance opinion: factor levels more to less explanatory
> sort(pvadij[,2])
     AFEI>12    HAB.scrpu     TFEI.fix INGR(150,250]      PLAÇ<36 INGR(100,150]      PATR4-8
0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 3.541611e-14 1.329648e-11 2.179690e-11
  ECIV.casat   EDAT(45,68]      IMP<.7  PREU1.1-1.4     PATR8-15   AFEI[5-12)     PATR>15
1.865655e-10 1.086958e-09 1.680927e-09 5.589849e-08 1.394307e-07 3.394110e-07 8.868735e-07
     IMP.7-1 INGR(250,960]   PATR0.5-4   DESP(35,45]  DESP(45,60]  DESP(60,75]     CARRPMig
4.792517e-06 1.115941e-04 2.623134e-03 4.568517e-03 1.027362e-02 5.204700e-02 7.883559e-02
 PREU1.4-1.7   EDAT(28,36]   ECIV.vidu    PLAÇ36-47     CARRPAlt    HAB.pares     CARRPNul
2.486637e-01 4.220112e-01 5.144387e-01 7.033886e-01 8.199414e-01 8.827363e-01 8.850363e-01
    IMP1-1.3    PLAÇ48-59  EDAT(36,45]  DESP(34,35]  HAB.contpri     PREU1.7+    AFEI[2-5)
9.296074e-01 9.351384e-01 9.654742e-01 9.821484e-01 9.860694e-01 9.960037e-01 9.995579e-01
   TFEI.altr     PREU<1.1    TFEI.auto DESP(90,180]      PLAÇ>59  DESP(75,90]  EDAT(17,28]
9.997235e-01 9.997898e-01 9.998193e-01 9.998260e-01 9.999113e-01 9.999584e-01 9.999883e-01
 ECIV.solter  ECIV.sepdiv   HAB.altres  INGR(50,100]      AFEI<2  HAB.lloguer     TFEI.tmp
9.999896e-01 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
 INGR(-1,50]     PATR<0.5     IMP1.3-5
1.000000e+00 1.000000e+00 1.000000e+00
# Rejection opinion: factor levels unless more explanatory
> sort(pvadij[,1])
      AFEI<2  HAB.lloguer     TFEI.tmp  INGR(-1,50]     PATR<0.5     IMP1.3-5 INGR(50,100]
0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 3.586020e-14
  HAB.altres  ECIV.sepdiv  ECIV.solter  EDAT(17,28]  DESP(75,90]      PLAÇ>59 DESP(90,180]
1.201261e-13 4.014773e-08 1.038939e-05 1.167609e-05 4.158758e-05 8.871978e-05 1.739983e-04
   TFEI.auto     PREU<1.1    TFEI.altr    AFEI[2-5)     PREU1.7+  HAB.contpri  DESP(34,35]
1.807043e-04 2.102092e-04 2.765210e-04 4.420680e-04 3.996315e-03 1.393059e-02 1.785162e-02
 EDAT(36,45]    PLAÇ48-59     IMP1-1.3     CARRPNul    HAB.pares     CARRPAlt    PLAÇ36-47
3.452581e-02 6.486157e-02 7.039256e-02 1.149637e-01 1.172637e-01 1.800586e-01 2.966114e-01
   ECIV.vidu  EDAT(28,36]  PREU1.4-1.7     CARRPMig  DESP(60,75]  DESP(45,60]  DESP(35,45]
4.855613e-01 5.779888e-01 7.513363e-01 9.211644e-01 9.479530e-01 9.897264e-01 9.954315e-01
   PATR0.5-4 INGR(250,960]     IMP.7-1      PATR>15   AFEI[5-12)     PATR8-15  PREU1.1-1.4
9.973769e-01 9.998884e-01 9.999952e-01 9.999991e-01 9.999997e-01 9.999999e-01 9.999999e-01
      IMP<.7  EDAT(45,68]   ECIV.casat      PATR4-8 INGR(100,150]      PLAÇ<36      AFEI>12
1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
   HAB.scrpu     TFEI.fix INGR(150,250]
1.000000e+00 1.000000e+00 1.000000e+00
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

```
# Use FactoMineR package (forget the functions)
> names(opina)
 [1] "dictamen"     "anys.feina"   "habitatge"    "plas"         "edat"         "estat.civil"
 [7] "tipus.feina"  "despeses"     "ingressos"    "patrimoni"    "carrega.patr" "import.assoc"
[13] "preu.final"   "f.dictamen"   "f.afei"       "f.habi"       "f.plas"       "f.edat"
[19] "f.eciv"       "f.tfei"       "f.desp"       "f.ingr"       "f.patr"       "f.carrpatr"
[25] "f.import"     "f.preu"
> nomsvads<-names(opina)[c(14:25)];nomsvads
 [1] "f.dictamen" "f.afei"     "f.habi"     "f.plas"     "f.edat"     "f.eciv"     "f.tfei"     "f.desp"
 [9] "f.ingr"     "f.patr"     "f.carrpatr" "f.import"
> catdes(opina[,c(nomsvads)],num.var=1)
$test.chi2
                p.value df
f.afei     2.913760e-87  3
f.ingr     7.456113e-77  5
f.tfei     4.314139e-70  4
f.habi     2.760016e-46  5
f.patr     9.013802e-43  5
f.import   1.291256e-23  3
f.plas     1.912465e-12  3
f.eciv     8.783901e-11  4
f.edat     2.605645e-09  3
f.desp     7.466957e-08  5
f.carrpatr 1.629441e-04  3

$category
$category$rebutja
                      Cla/Mod   Mod/Cla     Global      p.value    v.test
f.afei=AFEI<2        48.99522 40.829346 23.4620566 2.066308e-61 16.534634
f.tfei=TFEI.tmp      59.95575 21.610845 10.1481814 2.077713e-50 14.930796
f.ingr=INGR(-1,50]   56.25000 22.248804 11.1360575 3.185727e-44 13.949106
f.patr=PATR<0.5      40.03667 52.232855 36.7310283 3.948997e-40 13.259999
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

```
f.import=IMP1.3-5     39.63706 33.094099 23.5069600 3.490432e-20   9.202707
f.habi=HAB.lloguer    39.87667 30.940989 21.8455321 3.720151e-19   8.944981
f.ingr=INGR(50,100]   36.08838 35.167464 27.4360126 1.247073e-12   7.100055
f.habi=HAB.altres     45.76803 11.642743  7.1621015 5.145047e-12   6.901516
f.eciv=ECIV.sepdiv    46.42857  6.220096  3.7718904 4.282184e-07   5.055966
f.eciv=ECIV.solter    33.57216 26.156300 21.9353390 3.015307e-05   4.172307
f.edat=EDAT(17,28]    32.85953 31.339713 26.8522676 3.248560e-05   4.155303
f.desp=DESP(75,90]    37.24928 10.366826  7.8356533 1.543626e-04   3.783944
f.plas=PLAÇ>59        31.03983 47.846890 43.3991917 2.080250e-04   3.709066
f.carrpatr=NA         72.22222  1.036683  0.4041311 2.562996e-04   3.655882
f.ingr=NA             58.82353  1.594896  0.7633588 3.356363e-04   3.586119
f.tfei=TFEI.auto      32.51953 26.555024 22.9905703 5.245459e-04   3.467898
f.desp=DESP(90,180]   40.76433  5.103668  3.5249214 7.273158e-04   3.379070
f.tfei=TFEI.altr      39.76608  5.422648  3.8392456 1.112552e-03   3.260400
f.afei=AFEI[2-5)      32.25806 26.315789 22.9681185 1.123301e-03   3.257672
f.habi=NA             50.00000  1.036683  0.5837449 2.945121e-02   2.177394
f.desp=DESP(34,35]    30.47069 29.425837 27.1890436 3.985629e-02   2.055235
f.habi=HAB.contpri    34.14634  6.698565  5.5231253 4.044316e-02   2.049194
f.patr=NA             42.55319  1.594896  1.0552313 4.688543e-02   1.987334
f.desp=DESP(45,60]    25.39388 21.850080 24.2254154 2.199695e-02  -2.290421
f.desp=DESP(35,45]    24.69945 18.022329 20.5433318 9.678926e-03  -2.587094
f.patr=PATR0.5-4      24.77798 22.248804 25.2806466 3.756515e-03  -2.897916
f.ingr=INGR(250,960]  18.46690  4.226475  6.4436462 1.224361e-04  -3.841196
f.import=IMP.7-1      23.49304 24.242424 29.0525370 9.155198e-06  -4.436221
f.afei=AFEI[5-12)     22.53642 20.972887 26.2011675 5.723537e-07  -5.000320
f.patr=PATR>15        15.52347  3.429027  6.2191289 5.347060e-07  -5.013421
f.patr=PATR8-15       17.60722  6.220096  9.9461159 8.470896e-08  -5.356800
f.import=IMP<.7       21.45923 19.936204 26.1562640 2.228576e-09  -5.980205
f.edat=EDAT(45,68]    20.97378 17.862839 23.9784463 1.302426e-09  -6.067097
f.eciv=ECIV.casat     25.57853 66.108453 72.7660530 8.651867e-10  -6.132482
f.patr=PATR4-8        19.35135 14.274322 20.7678491 8.247049e-12  -6.834189
f.ingr=INGR(100,150]  21.18768 23.046252 30.6241581 3.735941e-12  -6.946827
f.plas=PLAÇ<36        16.19614  8.692185 15.1100135 7.084890e-15  -7.782957
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

```
f.ingr=INGR(150,250] 16.36537 13.716108 23.5967670 7.929587e-24 -10.064492
f.habi=HAB.scrpu     18.50973 31.100478 47.3057925 1.491928e-42 -13.672021
f.tfei=TFEI.fix      20.67736 46.251994 62.9770992 2.619632e-46 -14.287468
f.afei=AFEI>12       12.22313 11.881978 27.3686574 8.107785e-53 -15.296180

$category$accepta
                      Cla/Mod  Mod/Cla     Global      p.value    v.test
f.afei=AFEI>12       87.77687 33.43750 27.3686574 8.107785e-53 15.296180
f.tfei=TFEI.fix      79.32264 69.53125 62.9770992 2.619632e-46 14.287468
f.habi=HAB.scrpu     81.49027 53.65625 47.3057925 1.491928e-42 13.672021
f.ingr=INGR(150,250] 83.63463 27.46875 23.5967670 7.929587e-24 10.064492
f.plas=PLAÇ<36       83.80386 17.62500 15.1100135 7.084890e-15  7.782957
f.ingr=INGR(100,150] 78.81232 33.59375 30.6241581 3.735941e-12  6.946827
f.patr=PATR4-8       80.64865 23.31250 20.7678491 8.247049e-12  6.834189
f.eciv=ECIV.casat    74.42147 75.37500 72.7660530 8.651867e-10  6.132482
f.edat=EDAT(45,68]   79.02622 26.37500 23.9784463 1.302426e-09  6.067097
f.import=IMP<.7      78.54077 28.59375 26.1562640 2.228576e-09  5.980205
f.patr=PATR8-15      82.39278 11.40625  9.9461159 8.470896e-08  5.356800
f.patr=PATR>15       84.47653  7.31250  6.2191289 5.347060e-07  5.013421
f.afei=AFEI[5-12)    77.46358 28.25000 26.2011675 5.723537e-07  5.000320
f.import=IMP.7-1     76.50696 30.93750 29.0525370 9.155198e-06  4.436221
f.ingr=INGR(250,960] 81.53310  7.31250  6.4436462 1.224361e-04  3.841196
f.patr=PATR0.5-4     75.22202 26.46875 25.2806466 3.756515e-03  2.897916
f.desp=DESP(35,45]   75.30055 21.53125 20.5433318 9.678926e-03  2.587094
f.desp=DESP(45,60]   74.60612 25.15625 24.2254154 2.199695e-02  2.290421
f.patr=NA            57.44681  0.84375  1.0552313 4.688543e-02 -1.987334
f.habi=HAB.contpri   65.85366  5.06250  5.5231253 4.044316e-02 -2.049194
f.desp=DESP(34,35]   69.52931 26.31250 27.1890436 3.985629e-02 -2.055235
f.habi=NA            50.00000  0.40625  0.5837449 2.945121e-02 -2.177394
f.afei=AFEI[2-5)     67.74194 21.65625 22.9681185 1.123301e-03 -3.257672
f.tfei=TFEI.altr     60.23392  3.21875  3.8392456 1.112552e-03 -3.260400
f.desp=DESP(90,180]  59.23567  2.90625  3.5249214 7.273158e-04 -3.379070
f.tfei=TFEI.auto     67.48047 21.59375 22.9905703 5.245459e-04 -3.467898
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

```
f.ingr=NA             41.17647  0.43750  0.7633588 3.356363e-04  -3.586119
f.carrpatr=NA         27.77778  0.15625  0.4041311 2.562996e-04  -3.655882
f.plas=PLAÇ>59        68.96017 41.65625 43.3991917 2.080250e-04  -3.709066
f.desp=DESP(75,90]    62.75072  6.84375  7.8356533 1.543626e-04  -3.783944
f.edat=EDAT(17,28]    67.14047 25.09375 26.8522676 3.248560e-05  -4.155303
f.eciv=ECIV.solter    66.42784 20.28125 21.9353390 3.015307e-05  -4.172307
f.eciv=ECIV.sepdiv    53.57143  2.81250  3.7718904 4.282184e-07  -5.055966
f.habi=HAB.altres     54.23197  5.40625  7.1621015 5.145047e-12  -6.901516
f.ingr=INGR(50,100]   63.91162 24.40625 27.4360126 1.247073e-12  -7.100055
f.habi=HAB.lloguer    60.12333 18.28125 21.8455321 3.720151e-19  -8.944981
f.import=IMP1.3-5     60.36294 19.75000 23.5069600 3.490432e-20  -9.202707
f.patr=PATR<0.5       59.96333 30.65625 36.7310283 3.948997e-40 -13.259999
f.ingr=INGR(-1,50]    43.75000  6.78125 11.1360575 3.185727e-44 -13.949106
f.tfei=TFEI.tmp       40.04425  5.65625 10.1481814 2.077713e-50 -14.930796
f.afei=AFEI<2         51.00478 16.65625 23.4620566 2.066308e-61 -16.534634

attr(,"class")
[1] "catdes" "list "
>
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

The conclusions drawn from the results of inference are the previous page:

- The most discriminant rules by accepting CBD Prizing measures are AFEI> 12 (years of working more than 12 years), HAB.scrpu (available housing property with formalized writing), TFEI.fix (have a permanent contract) and INGR (150,250] (annual income between 150000 to 250000 €.

- The more discriminating arrangements with rejection of CBD Prizing measures are AFEI<2 (years of working less than 2 years), HAB.lloguer (usual rental housing), TFEI.tmp (have a temporary contract) , INGR (-1.50] (annual income below € 50000), PATR <0.5 (less than 5000 € worth), etc..

- The less discriminating modalities with rejection of CBD Prizing measures are AFEI>12 (years of working more than 12 years), HAB.scrpu (available housing property with formalized writing), TFEI.fix (have a contract permanent employment) and INGR (150,250] (annual income between € 150000 to 250000. Of course, are most associated with the form of acceptance of our response.

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

3-2.5 Association of quantitative target and I level factors

Changing roles: quantitative response variables with continuous character and income as explanatory variables all factors. The addressed question is whether there is a difference in income target depending on job, age, expenses, income, assets, equity load, amount of credit, term loans and longer loan impairment objects.

```
> attributes(opina)$names
 [1] "dictamen"     "anys.feina"   "habitatge"    "plas"         "edat"         "estat.civil"
 [7] "tipus.feina"  "despeses"     "ingressos"    "patrimoni"    "carrega.patr" "import.assoc"
[13] "preu.final"   "f.dictamen"   "f.afei"       "f.habi"       "f.plas"       "f.edat"
[19] "f.eciv"       "f.tfei"       "f.desp"       "f.ingr"       "f.patr"       "f.carrpatr"
[25] "f.import"     "f.preu"

# Variables con tratamiento categórico: creación lista de variables

attach(opina)
vads <-
list(f.afei,f.habi,f.plas,f.edat,f.eciv,f.tfei,f.desp,f.ingr,f.patr,f.carrpatr,f.import,f.preu)
pvackl <-NULL
for (i in 1:12 ) { pvackl <- rbind(pvackl, tijs.test(vads[[i]],ingressos)$pval) }
pvackl
sort(pvackl[,1])

> pvackl
                      [,1]
AFEI<2        1.000000e+00
AFEI[2-5)     9.799195e-01
AFEI[5-12)    3.130949e-04
AFEI>12       9.008350e-12
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

```
HAB.lloguer    7.257076e-01
...
PREU<1.1       1.000000e+00
PREU1.1-1.4    9.591006e-01
PREU1.4-1.7    7.225851e-01
PREU1.7+       0.000000e+00

> sort(pvackl[,1])
    HAB.scrpu    ECIV.casat      TFEI.fix  DESP(90,180] INGR(150,250] INGR(250,960]      PREU1.7+
 0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
     PATR>15       IMP1.3-5      AFEI>12   DESP(60,75]   DESP(75,90]      PATR4-8      CARRPMig
 1.065814e-14  8.471002e-14  9.008350e-12  3.800549e-10  4.311581e-10  6.956680e-09  1.069969e-07
    PATR8-15   EDAT(45,68]      CARRPAlt   AFEI[5-12)      PLAÇ<36   EDAT(36,45]   EDAT(28,36]
 1.836212e-07  3.321193e-07  4.011699e-05  3.130949e-04  7.755021e-03  2.455788e-02  1.646013e-01
  DESP(45,60]     PLAÇ48-59     PLAÇ>59      IMP1-1.3   HAB.contpri   PREU1.4-1.7   HAB.lloguer
 2.786199e-01  3.241068e-01  4.018773e-01  5.366704e-01  6.000310e-01  7.225851e-01  7.257076e-01
    ECIV.vidu      IMP.7-1   ECIV.sepdiv   PREU1.1-1.4    AFEI[2-5) INGR(100,150]     PATR0.5-4
 7.745403e-01  8.207252e-01  9.010088e-01  9.591006e-01  9.799195e-01  9.930373e-01  9.950800e-01
   PLAÇ36-47   DESP(35,45]     TFEI.auto     TFEI.altr   HAB.altres      CARRPNul       IMP<.7
 9.976938e-01  9.997583e-01  9.999918e-01  9.999995e-01  9.999998e-01  1.000000e+00  1.000000e+00
    HAB.pares      TFEI.tmp      PREU<1.1   EDAT(17,28]       AFEI<2   ECIV.solter   DESP(34,35]
 1.000000e+00  1.000000e+00  1.000000e+00  1.000000e+00  1.000000e+00  1.000000e+00  1.000000e+00
  INGR(-1,50]  INGR(50,100]     PATR<0.5
 1.000000e+00  1.000000e+00  1.000000e+00
>
> library(FactoMineR)
> nomsvars<-names(opina)[c(14:26)];nomsvars
 [1] "f.dictamen" "f.afei" "f.habi"   "f.plas"   "f.edat"   "f.eciv" "f.tfei"
"f.desp"
 [9] "f.ingr"     "f.patr"     "f.carrpatr" "f.import"   "f.preu"
> # The answer can not be NAs
> df<-opina[!is.na(opina$ingressos),c("ingressos",nomsvars)]
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

```
> dim(df)
[1] 4420   14
> condes(df,num.var=1)
$quali
                  R2      p.value
f.ingr     0.848002754 0.000000e+00
f.desp     0.056348645 2.709087e-53
f.dictamen 0.043837874 5.628929e-45
f.patr     0.039470569 1.566999e-36
f.tfei     0.027772478 6.197342e-26
f.preu     0.025942786 5.366424e-25
f.afei     0.023337468 1.855788e-22
f.habi     0.022446905 4.695900e-20
f.eciv     0.017957949 1.722417e-16
f.import   0.016140145 1.701961e-15
f.edat     0.014824500 3.122086e-14
f.carrpatr 0.012758064 2.965749e-12
f.plas     0.002604839 9.225141e-03

$category
                Estimate       p.value
INGR(250,960]  199.563277  0.000000e+00
INGR(150,250]   38.786475 1.011900e-280
accepta         20.155082  5.628929e-45
PREU1.7+        22.581156  1.772866e-23
DESP(90,180]    52.310244  8.904500e-23
IMP1.3-5        17.050889  6.756882e-14
PATR>15         36.128442  1.936061e-13
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

```
AFEI>12          15.211624  4.866698e-12
CARRPAlt         68.932323  4.190370e-07
EDAT(45,68]      11.089009  8.987020e-07
HAB.scrpu        16.931709  3.202381e-06
AFEI[5-12)        8.353976  1.396559e-04
PATR8-15         15.218194  2.471397e-04
DESP(75,90]      14.307945  3.751772e-04
PATR4-8           9.603444  4.781514e-03
TFEI.fix         48.227245  4.989573e-03
PLAÇ<36           6.974560  1.122255e-02
NA              -29.467112  2.663771e-03
PLAÇ36-47        -7.506832  1.883317e-03
HAB.pares       -13.459898  1.149712e-03
PATR0.5-4       -10.696171  1.026747e-03
HAB.altres      -18.567615  2.827650e-04
NA              -63.763973  1.744878e-05
DESP(45,60]     -12.376591  3.185092e-06
IMP<.7          -13.930197  2.819359e-10
PATR<0.5        -20.786798  1.000923e-11
PREU<1.1        -16.983877  5.773270e-14
EDAT(17,28]     -16.887951  1.246510e-14
DESP(35,45]     -22.681156  6.839499e-16
AFEI<2          -19.557552  1.262618e-17
DESP(34,35]     -35.929223  9.766366e-44
rebutja         -20.155082  5.628929e-45
INGR(100,150]   -26.347475 9.291599e-163
INGR(50,100]    -71.953740  0.000000e+00
INGR(-1,50]    -140.048538  0.000000e+00
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

Helper function to calculate the p value of the contrast statistic for each level of an explanatory factor in rows (I levels) and response columns:

```
#  Create a function to calculate t for each contrast level of a variable

tijs.test <- function( x, y ){
pc <- tapply( y, x,mean,na.rm = TRUE);
# taula amb mitjana pels nivells de x
mit <- mean(y,na.rm = TRUE);
std2 <- var(y,use="complete.obs");

# Replicated as many times as global average levels of x.
pc <- matrix( data=pc, nrow=1, ncol= length( levels(x) ), byrow=T );
pcI <- matrix( data=mit, nrow=1, ncol= length( levels(x) ), byrow=T );
taula <- table( x,y );
tpx <- t(matrix(apply(taula,1,sum)));
tt <- sum( taula );

# Numerator
dpf <- pc - pcI;

# Denominator
denom <- sqrt( (1-(tpx/tt) )* (std2/tpx) );
tij <- dpf /denom;
ptij <- 1-pt(tij,tpx-1);
tij <- t( tij );
ptij <- t( ptij );
row.names(tij) <- levels(x);
row.names(ptij) <- levels(x);

list(vtest=tij, pval=ptij )}
```

```
# Output:
# One list with tij (vtest) and pvalues (pval)
```

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

3-2.6 Work Sample and Test

A mandatory standard in any statistical model is its validation and hence the possibility of application to other data and the ability to use their predictions.

A universal technique for analysis and data mining is to divide the data randomly split into two sets: the subset of the test work and the test. The work sample will be used to build and estimate the statistical model (models if different techniques are used), while the test sample will be used to validate the proposed model so that it is possible to construct a validation table for a binary response style:

Test	Y=1	Y=0	Total
$\hat{y}_i = 1$	True positives (a)	False positives (b)	a+b
$\hat{y}_i = 0$	False negatives (c)	True negatives (d)	c+d
	a+c	b+d	n

- *Sensitivity* is the proportion of true values 1 that are estimated 1: Sn =a/(a+c).
- *Specificity* is the proportion of true values 0 that are estimated 0: Sp = d/(b+d).
- The *positive and negative predictive values* of the model are defined as.: P+=a/(a+b) y P- =d/(c+d).

Predictive capacity : (a+d)/n

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

The working subsample generally contain between 50 and 95% of the original data.

If the amount of data is not high then can be applied cross-validation techniques. In this technique data is randomly divided into n groups (equal size, that is n = 10.) Reserves as a group subsample of Test and n-1 remaining as Working subsample, and this process is repeated n times, leaving a distinct group as Test at each iteration, for each iteration error rate is obtained with the test sample for . Finally, a model is built with all the data and accuracy and error rate is obtained by averaging the n error rates available.

Another technique for estimating the error of a model, when few data are available or when the inference is complicated, lies in the bootstrapping: A first model with the entire sample is constructed and then point estimates are derived, then numerous sample datasets ASCR (over 1000 samples) and (from the point estimates) with these samples confidence intervals or error rates for the whole model is built.

3-2 CBD PRIZING EXAMPLE: EXPLORATION OF LARGE TABLES (DATABASE)

In the R statistical work environment, could be from the starting sample in subsamples of Working and Test (each in a data.frame) with type instructions which are used below:

```
     sample(x, size, replace = FALSE, prob = NULL)

     x: Either a (numeric, complex, character or logical) vector of more than one element from which to choose,
or a positive integer, in this case 1:x range is choosen.

    size: non-negative integer giving the number of items to choose.

 replace: Should sampling be with replacement?

    prob: A vector of probability weights for obtaining the elements of the vector being sampled.
```

```
# Create the WORK and TEST sample: for example 67% and 33% (o 75-25)

learn <- sample(1:(dim(opina)[1]),round(0.67*dim(opina)[1]))
learn <- sort( learn )
opina$filtre <- rep(TRUE,dim(opina)[1])
attach(opina)
opina[-learn,]$filtre <- FALSE
opina$filtre

# Create data.frame of Work and for Test, and firstly insightful columns are selected according
to the specification.
df.treball <- data.frame(… )
df.test <- df.treball[ -learn, ]
df.treball <- df.treball[ learn, ]
np <- dim(df.treball)[2]
dim(df.treball)[1] # For validation sample size LEARN
```