Profiling

K. Gibert[(1)]

[(1)]Department of Statistics and Operation Research

Knowledge Engineering and Machine Learning group at Intelligent Data Science and Artificial Intelligence Research Center

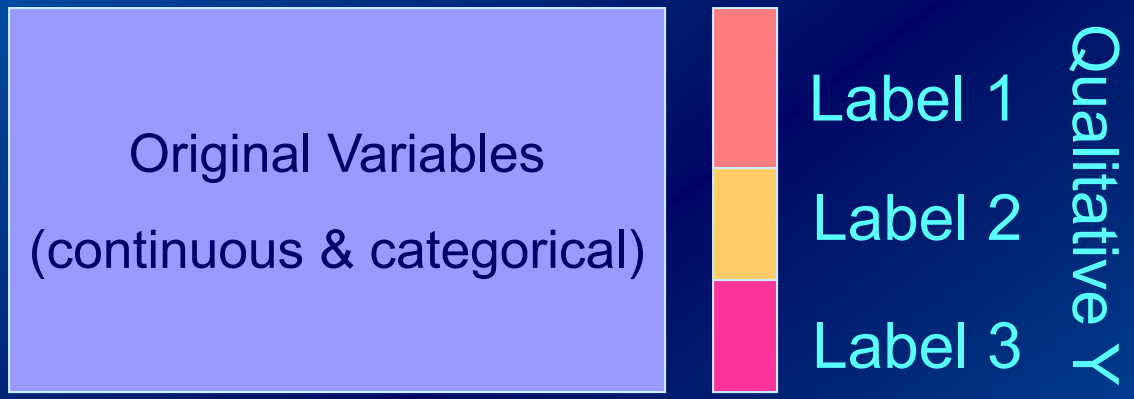
Universitat Politècnica de Catalunya, Barcelona

karina.gibert@upc.edu

https://www.eio.upc.edu/homepages/karina

Profiling

- Diferential characterisation among different groups



- Statistical characterization:
  - Testing
  - Profiling tools
  - Factorial graphs
  - Class panel graph
  - Traffic lights panel

Conceptualize the class

Characterizing a Qualitative Variable

1. Find significant variables wrt a qualitative variable Y

   *(feature extraction, r*elevant characteristics for Y*)*

   1. X numerical: ANOVA (Ftest) or Kruskall-Walis test

      Multiple boxplot or multiple histogram

   2. X quali: chisquare independence test

      cross table and barplots

2. For significant variables: find sense of differences

   (characterize significant differences) Test-values (Lebart)

   1. X num: Extension of t-test for the means comparison

      means profiling graph (barchart of local and global means)

   2. X quali: Extension of proportions comparison

      snake graphs for local and global proportions

statistical tests + visualization

1 test x modality

Y=Class variable

Characterizing a Qualitative Variable

1. Find significant variables wrt a qualitative variable Y

   *(feature extraction, r*elevant characteristics for Y*)*

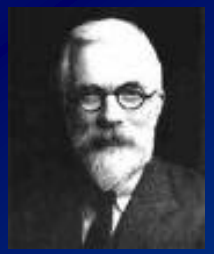
   1. X numerical: ANOVA (Ftest) or Kruskall-Walis test

      Multiple boxplot or multiple histogram

   2. X quali: chisquare independence test

      cross table and barplots

Association between one categorical variable and one numerical



Sir Ronald A. Fisher
English, 1890-1962

The ANOVA or F-Test

Requires Normality

To be used for feature selection

Test: H_0: $\mu_{Y|X=x_1} = \mu_{Y|X=x_2} = \ldots = \mu_{Y|X=x_s} = \mu$ (X, Y independent)

H_1: $\exists\, x \in \{x_1 \ldots x_s\}$: $\mu_{Y|X=x} \neq \mu$ (X,Y associated)

Statistics:

$$F = \frac{S_B^2/(q-1)}{S_W^2/(n-q)} \sim F_{q-1,n-q}$$

levels	*means*	*counts*
1	$\bar{x}_1$	n_1
$\vdots$	$\vdots$	$\vdots$
q	$\bar{x}_q$	n_q

overlapping

Equivalent:

$$\eta^2 = \frac{S_B^2}{S_W^2 + S_B^2}$$

$$S_W^2 = \sum_{k=1}^{q}\sum_{i=1}^{n_k}(x_{ki} - \bar{x}_k)^2$$

$$S_B^2 = \sum_{k=1}^{q} n_k(\bar{x}_k - \bar{x})^2$$

$\bar{x}, s^2$ global mean and

Alternative Kruskal-Wallis

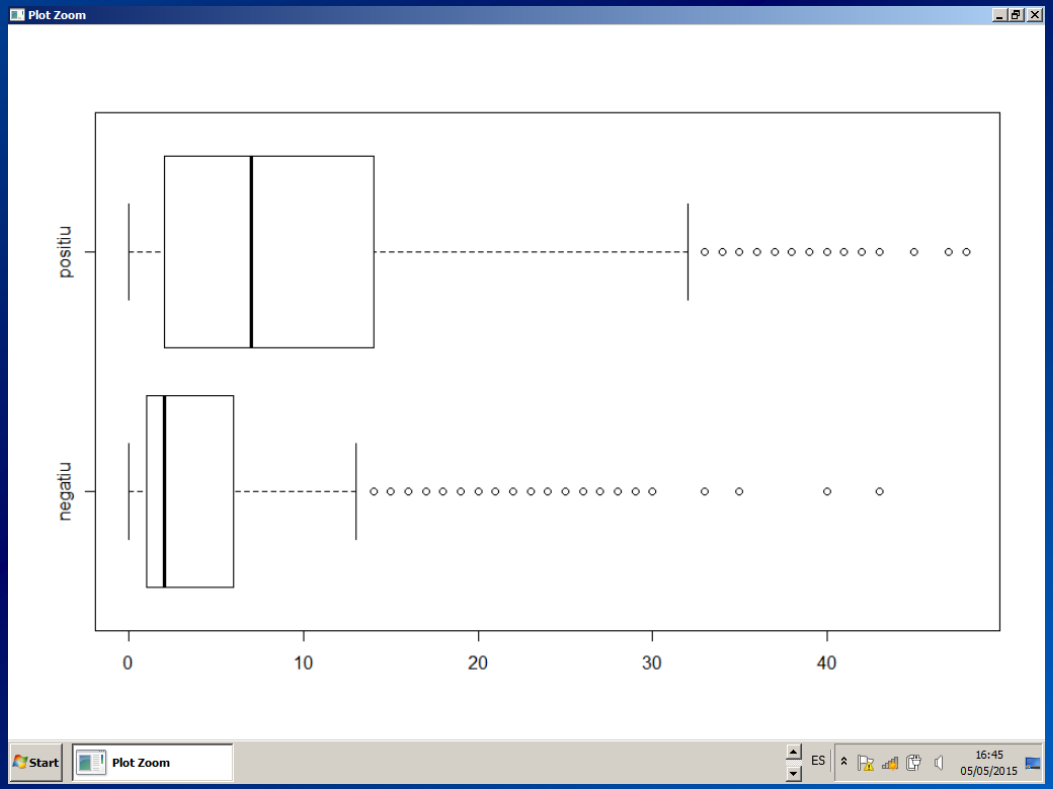
```
> catdes(X, num.var, proba = 0.05, row.w = NULL)
> condes(X, num.var, proba = 0.05, row.w = NULL)
```


UPC IDEAI

One categorical variable and one numerical

Multiple boxplot

Antiguitat a la feina



> boxplot(dades[,k]~P, horizontal=TRUE)

One categorical and one numerical

Descriptive by groups

aggregate(Antiguedad.Trabajo, by=list(Dictamen), FUN=mean)

Group.1 x

1 negatiu 4.586922

2 positiu 9.319062

aggregate(Antiguedad.Trabajo, by=list(Dictamen), FUN=sd)

Group.1 x

1 negatiu 6.118022

2 positiu 8.487919

aggregate(Antiguedad.Trabajo, by=list(Dictamen), FUN=max)

Group.1 x

1 negatiu 43

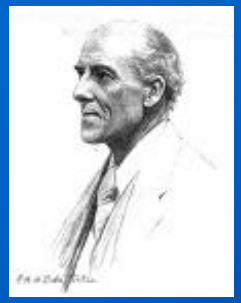
2 positiu 48

aggregate(Antiguedad.Trabajo, by=list(Dictamen), FUN=median)

Group.1 x

1 negatiu 2

2 positiu 7



Karl Pearson
English, 1857, 1936

Assessing association between categorical variables

The chi2 independence Test

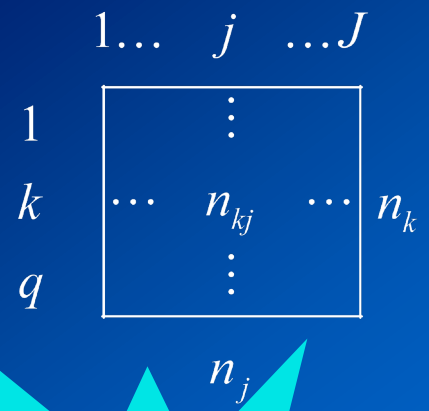
To be used for feature selection

Test:

H_0: X,Y are independent ($\pi_{kj}=\pi_k\pi_j \ \forall kj$)
H_1: X,Y are associated

Statistics:

$$X^2 = \sum_{k=1}^{p}\sum_{j=1}^{q} \frac{(n_{kj} - n_k n_j / n)^2}{n_k n_j / n} \sim \chi^2_{(p-1)(q-1)}$$



Missperformance if nkj<5

Ranking by ascending p.values

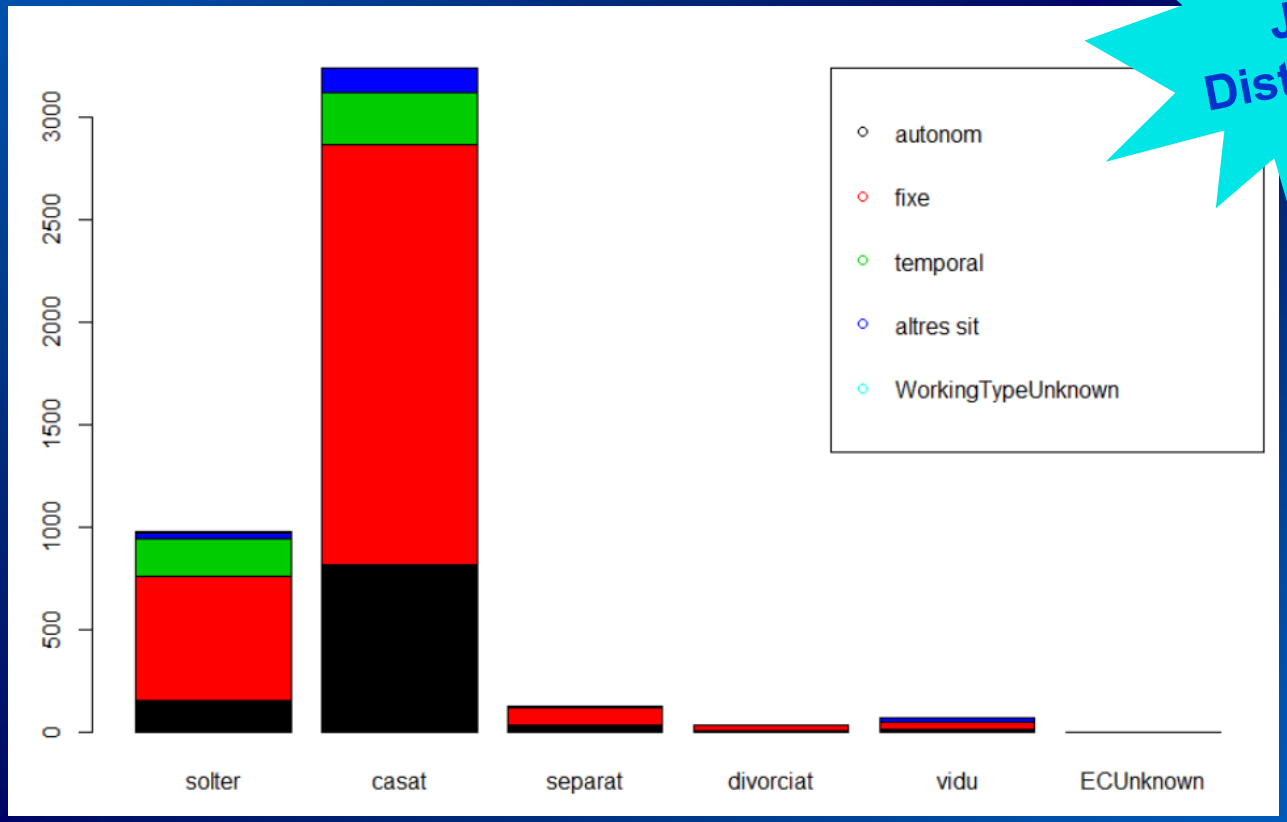
Care with Simpson's Paradox

```
> chisq.test
> catdes(X, num.var, proba = 0.05, row.w = NULL)
```


IDEAI

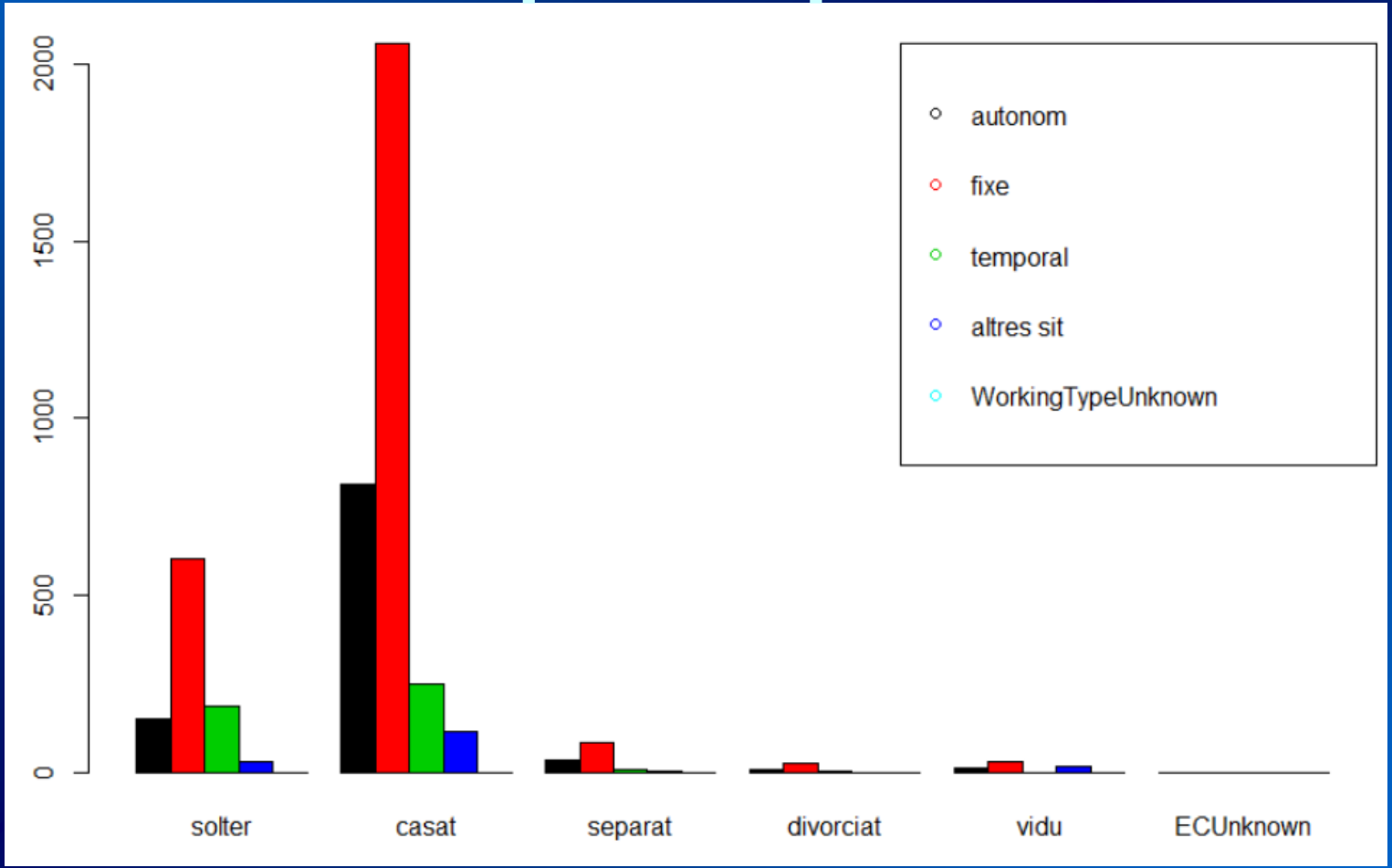
Two categorical variables

Multiple barplot



Two categorical variables

Multiple barplot

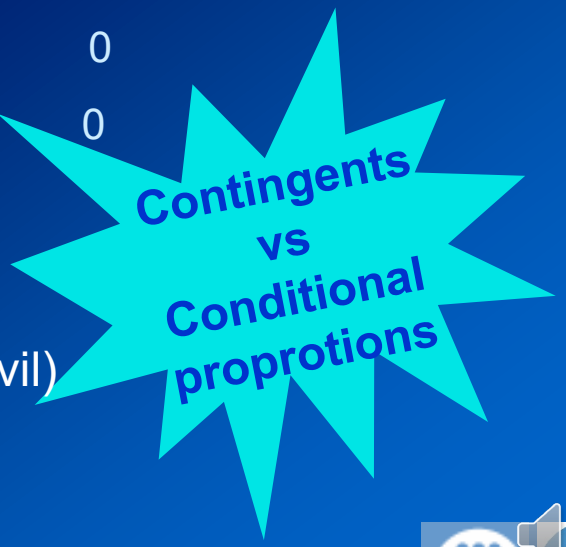


Contingency tables

(Cross Tables)

	Estado.civil					
Tipo.trabajo	solter	casat	separat	divorciat	vidu	ECUnknown
autonom	154	815	34	7	13	1
fixe	605	2056	84	28	33	0
temporal	188	252	8	3	1	0
altres sit	29	118	4	0	20	0
WTUnknown	2	0	0	0	0	0

> table<-table(Tipo.trabajo,Estado.civil)



Contingency tables

(Margins)

	Estat_civil						
Vivenda	solter	casat	vidu	separat	divorciat	Total	*Row %*
lloguer	174	723	11	50	15	973	*21.9%*
escriptura	167	1839	50	38	12	2106	*47.4%*
contr_privat	26	212	3	4	1	246	*5.5%*
ignora_cont	1	18	0	0	1	20	*0.4%*
pares	507	238	0	30	7	782	*17.6%*
altres viv	98	208	3	8	2	319	*7.2%*
Total	973	3238	67	130	38	4446	
Columns %	*21.9%*	*72.8%*	*1.5%*	*2.9%*	*0.9%*		

```
> rowperc<-prop.table(table,1)
> colperc<-prop.table(table,2)
```

Characterizing a Qualitative Variable

2. For significant variables: find sense of differences

   (characterize significant differences) Test-values (Lebart)

   1. X num: Extension of t-test for the means comparison

      means profiling graph (barchart of local and global means)

1 test per variable and class

   2. X quali: Extension of proportions comparison

      snake graphs for local and global proportions

1 test per modality and class

Importance of a numerical variable in a class

Statistical assessment

Ludovic Lébart

French 1936-

Test-values

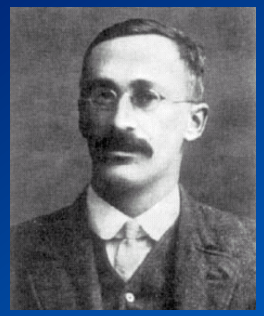
$$H_0 : \mu_k = \mu \quad k = 1, \ldots, q$$

$$t = \frac{\bar{x}_k - \bar{x}}{\sqrt{(1 - \frac{n_k}{n}) \frac{s^2}{n_k}}} \sim t_{n-1}$$

Student's *t*

William Gosset "Sudent",

English, 1876-1937

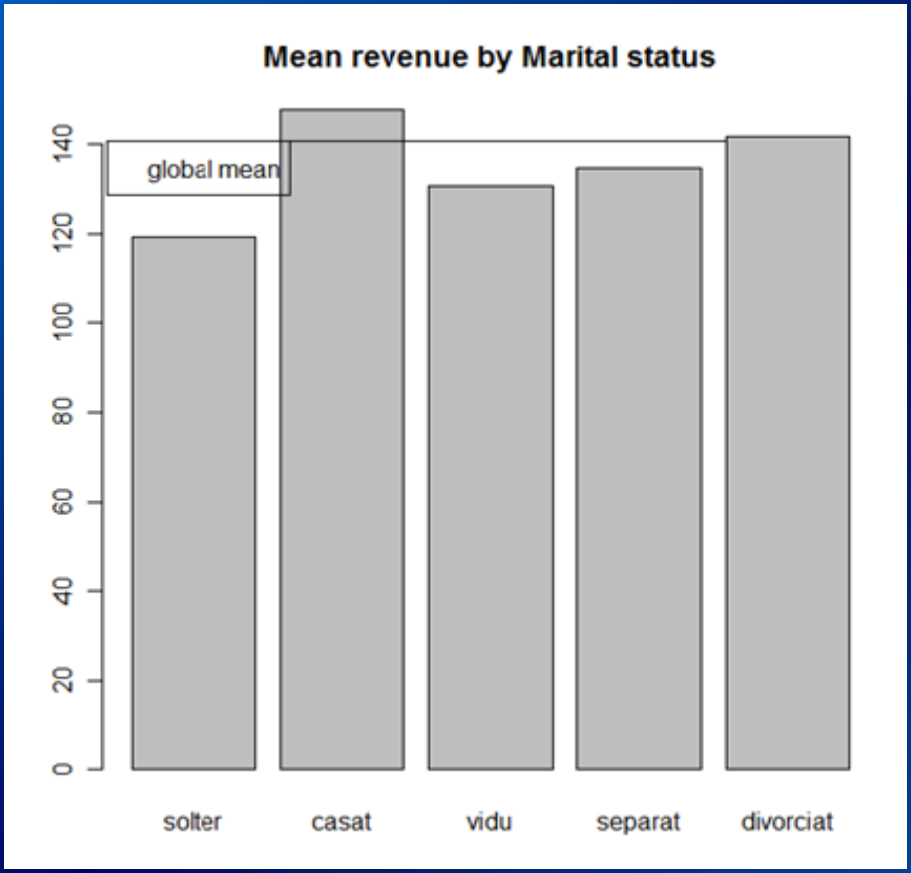


Normality

Rank the continuous variables by p.value (ascending)

Importance of a numerical variable in a class

Visual assessment




```
barplot(
  tapply(Revenue, Marital Status, mean),
  main=paste( "Means of", "Revenue",
                "by","Marital.Status"))
abline(h=mean(Revenue))
legend(0,mean(Revenue),
          "global mean",bty="n")
```

Importance of a modality in a class

Statistical assessment

Ludovic Lébart

French 1936-

Test-values

$$H_0 : p_{j \cdot k} = p_j \quad k = 1, \ldots, p; j = 1, \ldots, q$$

$$\frac{n_{kj}}{n_k} \sim N\left(p_j = \frac{n_j}{n}, \left(1 - \frac{n_k}{n}\right) \frac{p_j (1 - p_j)}{n_k} \right)$$

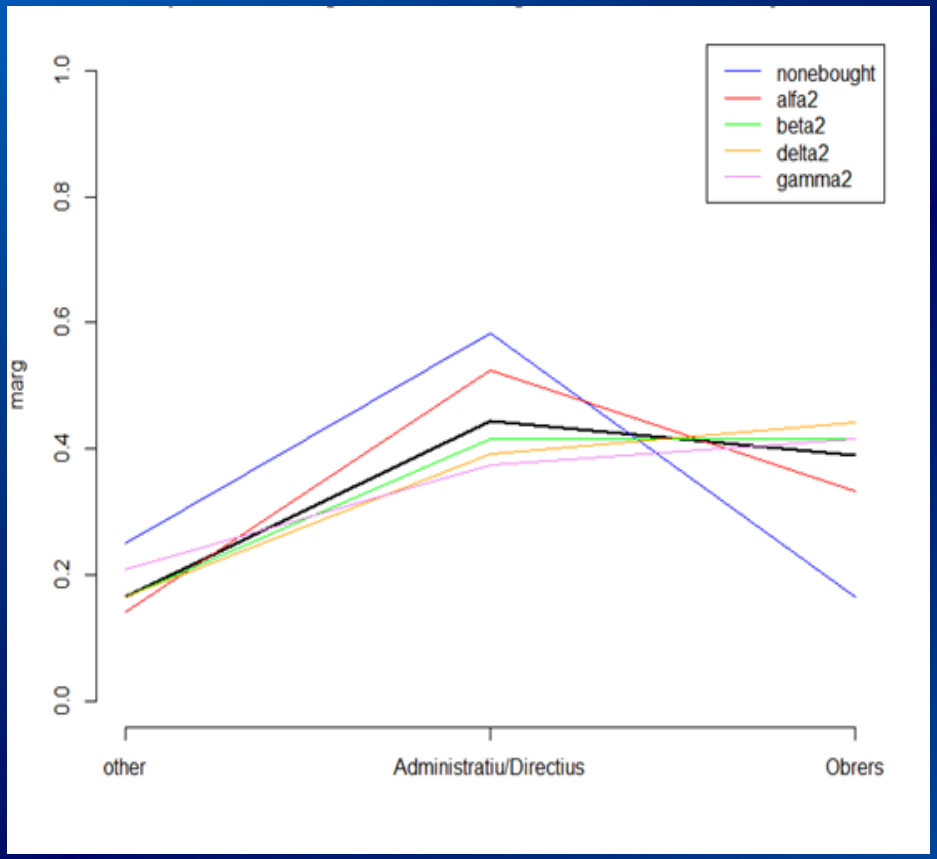
Non-rare phenomenon

$$z = \frac{\frac{n_{kj}}{n_k} - \frac{n_j}{n}}{\sqrt{\left(1 - \frac{n_k}{n}\right)\left(\frac{p_j \left(1 - p_j\right)}{n_k}\right)}} \sim N(0,1)$$

Rank the levels of the categorical variables by p.value (ascending)

Importance of a numerical variable in a class

Visual assessment




```
> plot(marg,type="n",ylim=c(0,1),
   main=paste("Prop. of brand by",
   names(dades)[k]))
> paleta<-
rainbow(length(levels(dades[,k])))
> for(c in
1:length(levels(dades[,k]))){

lines(rowperc[,c],col=paleta[c])
    }
> legend("topright", levels(dades[,k]),
   col=paleta, lty=2, cex=0.6)
```

Characterizing a Class Variable

1. Find significant variables wrt the class variable Y

   (ANOVA; K-W; Chi2…..
2. For significant variables: find sense of differences

   (characterize significant differences) Test-values (Lebart)
3. For each class:

   Collect the numeric variables with signifficant test-value

   Collect the modalities of qualitative variables with signifficant test-value

   Use the corresponding profiling graphs to identify the sense of the signifficance

   Build a class concept with sentences like

   Class C has Num var X higher (or lower)

   Class C has more (or less) presence of Modality

   If posible, assign a "label" to class that represents the global concept

Class Conceptualization

UPC IDEAI