Profiling

K. Gibert[(1)]

[(1)]*Department of Statistics and Operation Research*

Knowledge Engineering and Machine Learning group

Universitat Politècnica de Catalunya, Barcelona

karina.gibert@upc.edu

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

Profiling

Given a qualitative variable Y:

Identify relevant characteristics per group *(feature extraction)*

Use statistical tests and visualization over all X available variables

Tools depend on type of variables X

X numerical:

- Extension of t-test for mean comparisons
- Bar chart of global and local means

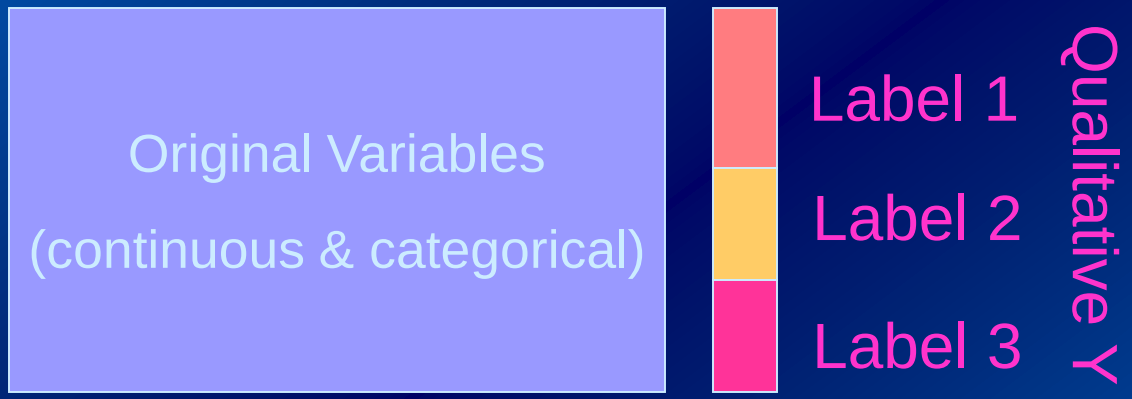
X qualitative:

- Extension of proportions comparison tests
- Line chart of global and local proportions

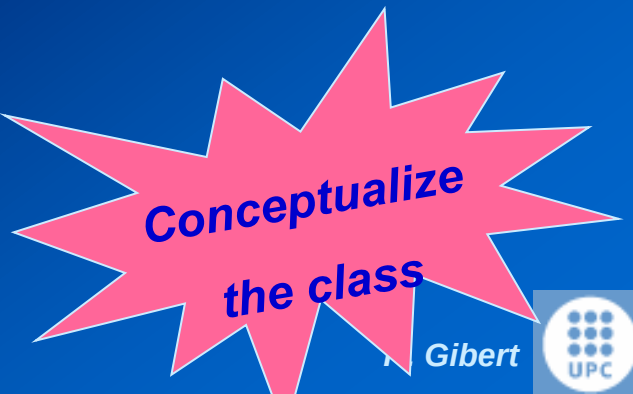
Repeat the test for every level of Y and rank by p-value

Profiling

- Diferential characterisation among different groups



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



Importance of a numerical variable in a class

Statistical assessment

Ludovic Lébart

French 1936-

Test-values

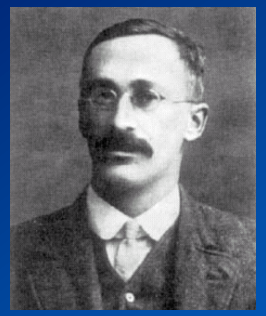
$$H_0 : \mu_k = \mu \quad k = 1, \ldots, q$$

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

Student's *t*

William Gosset "Sudent",

English, 1876-1937

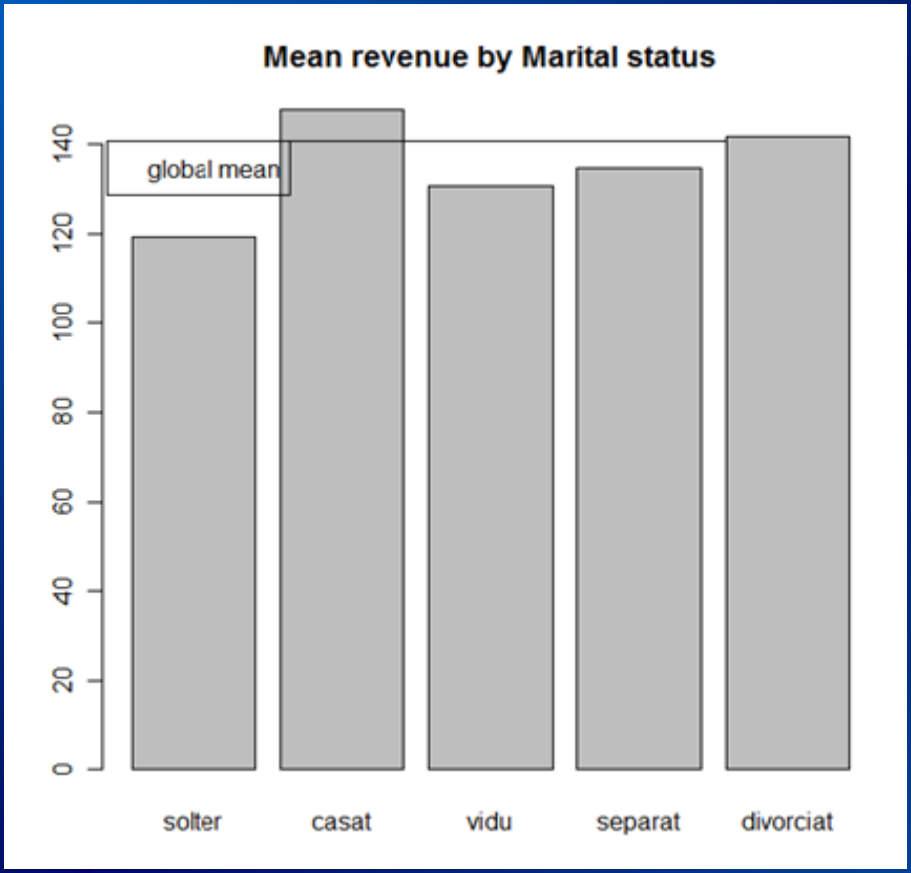


Normality

Rank the continuous variables by p.value (ascending)

Importance of a numerical variable in a class

Visual assessment




```
barplot(
   tapply(Revenue, Dictamen, 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

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

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

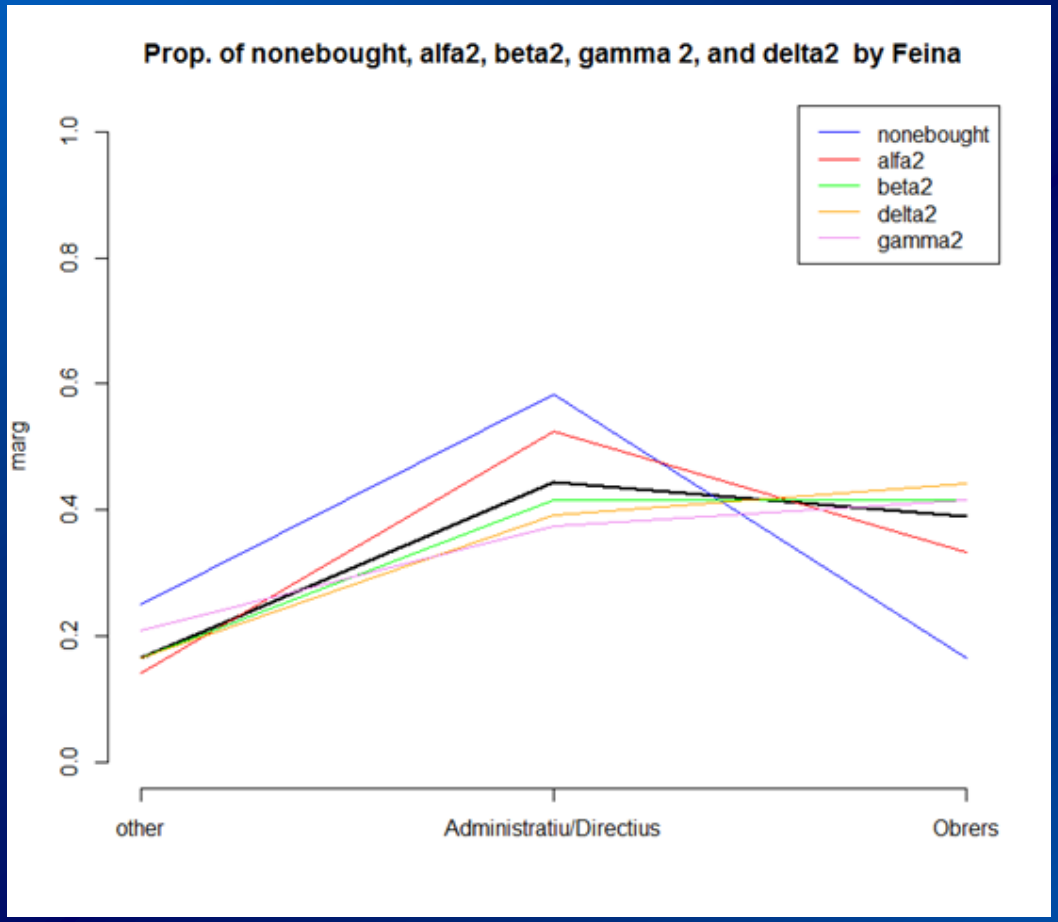
Non-rare phenomenon

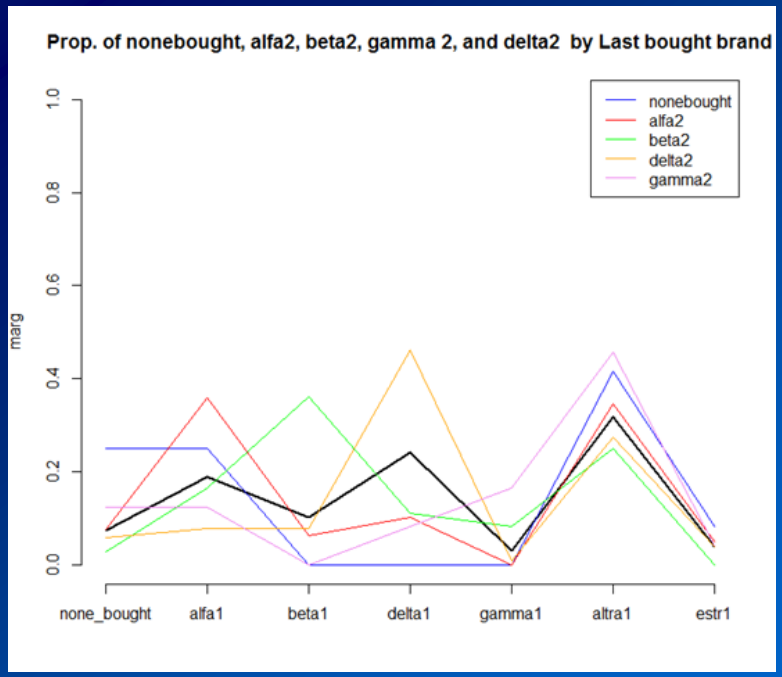
$$z = \frac{\frac{n_{kj}}{n_k} - \frac{n_j}{n}}{\sqrt{\left(1 - \frac{n_k}{n}\right)\left(\frac{p_j \left(1 - p_j\right)}{n_k}\right)}} \sim N(0,1)$$

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

Importance of a numerical variable in a class

Visual assessment





Importance of a numerical variable in a class

Visual assessment