Session 2:
Profiling: Interpreting the results

Anàlisi de Dades i Explotació de la Informació

Grau d'Enginyeria Informatica.

Information Tracking System

Prof. Mónica Bécue Bertaut & Lidia Montero

Monica.becue@upc.edu *lidia.montero@upc.edu*

FIB UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

1	Levels of Corporate Decision	17-feb	19-feb
2	Profiling	24-feb	26-feb
3 Oct	Multivariate analysis: PCA	03-mar	05-mar
4	Multivariate Analysis: Clustering	10-mar	12-mar
5	Multivariate Analysis: Interpretation rules	17-mar	19-mar
6	Multivariate Analysis: CA	24-mar	26-mar
7	Multivariate Analysis: MCA + clustering	07-abr	09-abr
			QUIZ - 1st Del
8	Statistical Modeling	14-abr	16-abr

Entrega deliverable I: Friday 17 abril

What is profiling?

We will see it by using a small data set: “chocolate data”

Description of the data set

Hall test for the evaluation of
10 types of chocolates

10 different chocolates
16 panelists
14 attributes
2 sessions

HACENDADO (55) HACENDADO (72)
HACENDADO (85)
LINDT-EF (70) LINDT-EF (85) LINDT-NS (70)
VALRHONA (64) VALRHONA (66) VALRHONA (70)
VALRHONA (85)

Average table

	o_cacao	o_leche	s_azucar	s_acido	s_amargo	a_cacao	a_leche	a_carame	a_vainilla	astringen	crujiente	fusion	pegajoso	granuloso	Marca	Gama	ConCacao
HACENDA	5,451613	3	7,5625	2,625	2,53125	4,84375	4,375	3,709677	3,1875	2,75	4,15625	5,59375	4,625	3,5	Hac	A	50to60
HACENDA	5,8125	2,625	5,0625	4,21875	5,0625	6,625	3,21875	3,53125	2,46875	4,59375	4,25	4,84375	4,0625	2,65625	Hac	A	61t070
HACENDA	6,4375	2,34375	2,84375	5,71875	7,59375	7,96875	1,71875	2,5625	2,125	6,5	4,625	4,3125	4,09375	2,96875	Hac	A	81to90
LINDT-EF (	6,8125	2,5	3,6875	5,90625	6,34375	7,03125	2,28125	2,71875	2,1875	5,78125	5,21875	4,90625	3,935484	2,15625	Lindt	B	61t070
LINDT-EF (	6,46875	2,21875	2,40625	5,71875	8,5625	8,125	1,59375	1,84375	1,90625	7,40625	5,34375	3,6875	3,59375	2,75	Lindt	B	81to90
LINDT-NS	5,9375	2,5	5,625	3,25	4,15625	5,59375	3,4375	3,5	2,96875	3,0625	3,875	4,875	3,677419	2	Lindt	B	61t070
VALRHON	5,5	2,625	5,28125	5,25	5,290323	6,125	3,1875	3,03125	2,5625	4,34375	4,25	5,65625	3,90625	2,15625	Val	B	61t070
VALRHON	5,5625	2,4375	5,5	4,125	5,483871	6,46875	3,1875	3,84375	2,6875	4,5	4,25	5,59375	4,0625	2,125	Val	B	61t070
VALRHON	5,375	2,5	4,875	5,3125	5,53125	5,71875	2,65625	2,9375	2,375	4,34375	4	5,15625	4,0625	2,1875	Val	B	61t070
VALRHON	6,6875	2,46875	2,53125	6,65625	8,25	7,625	1,6875	2,125	2	6,78125	4,125	4,78125	3,6875	2,40625	Val	B	81to90

Target variables

gama: categorical variable

olor a cacao: quantitative variable

1. Characterization of a categorical variable

1.1 Introduction: profiling groups of individuals

1.2 by the other categorical variables

1.2 by the categories of the other categorical variables

1.3 by the quantitative variables

1.1 Profiling one partition of the individuals and/or one group (=one category) of individuals

The groups of individuals are defined from a categorical variable , whose different values are called “levels”

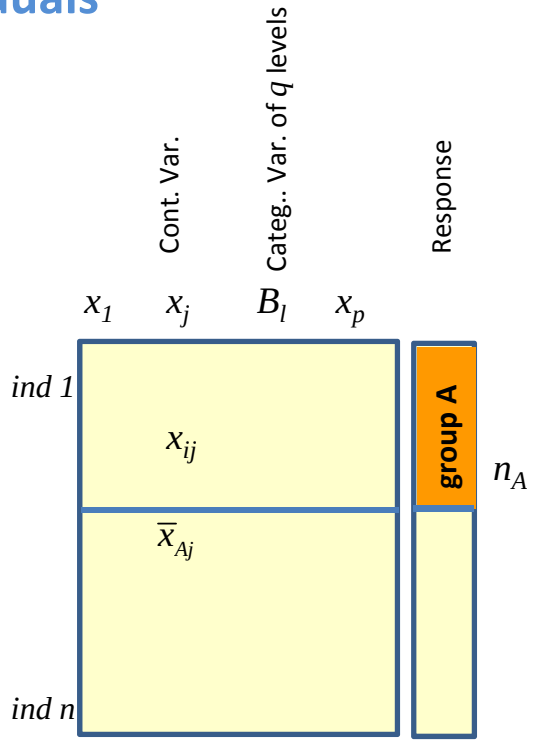
Goal:

- We consider as a response variable the variable whose categories define the groups that we want to “profile”.
- The *explanatory variables* are either categorical or quantitative.

Tool: To identify the variables whose behavior differs for the individuals belonging to a group as compared to the whole sample.

- Categorical : frequencies comparison
- quantitative variable : mean comparison

But using a **probabilistic test** to assess the significance of the result



1.2 Characterization of a categorical variable by the other categorical variables

Example: describing the variable *Marca*

```
> catdes(base,num.var=16)

$test.chi2
       p.value df
Marca 0.006737947  2
```

```
Marca    Gama

Marca     Gama
Hac  :3   A:3
Lindt:3   B:7
Val  :4
```

Chi.square test
performed through building all the cross-tables
also called contingency tables

between *Gama* and all the other categorical variables

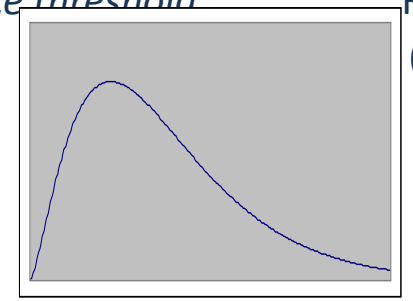
$$\chi^2_{obs} = \sum_i \sum_j \frac{\left(n_{ij} - n_{i.}n_{.j}/n\right)^2}{n_{i.}n_{.j}/n} = \sum_i \sum_j \frac{\left(n_{ij} - np_{i.}p_{.j}\right)^2}{np_{i.}p_{.j}}$$

Independence chi-2 test

Test

We want to see if variable *j* is related to another categorical variables *I*

Null hypothesis	H_0: conservative hypothesis. Both variables are independent
Alternative hypothesis	H_1: Both variables are not independent
Test statistic:	a statistic is a function of the sample (=observed data) and thus a variate or random variable in this case, the statistics chi-square is computed from comparing the differences between the observed counts and the expected counts under H_0
Reference distribution:	Distribution of the *test statistic* under H_0 (that is, if H_0 is true). Chi-2 distribution, with the convenient degrees of freedom
Significance threshold:	Risk of rejecting H_0 although H_0 being true (significance depends on the number of individuals) P-value



Chi2-law

1.3 Characterization of a categorical variable by the categories of the other categorical variables

$category

–**Relationship between each category** of the variable target and another category of another categorical variable: comparison of two proportions, taking into account an hypergeometric model

```
$category
$category$A
          Cla/Mod Mod/Cla Global     p.value   v.test
Marca=Hac     100     100     30 0.008333333 2.638257

$category$B
          Cla/Mod Mod/Cla Global     p.value    v.test
Marca=Hac       0       0     30 0.008333333 -2.638257
```

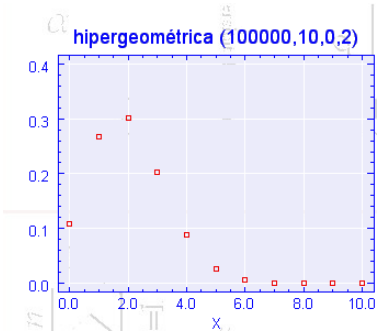
Hypothesis test

Null hypothesis	H_0: the proportion of "Hacendado" in "Gama A" group does not differ from this in the whole sample
Alternative hypothesis	H_1: the proportion of ""Hacendado" in "Gama A" group differs from this in the whole sample
Test statistic:	linked these proportions/ counts
Reference distribution:	Distribution of the *test statistic* under H_0 (if the H_0 is true).
Significance threshold:	as in the other cases

To test if the frequency of "Hacendado" **is due or not to chance (=H0)**, that is the labels "Hacendado" are put at random, or individuals of group *k* are taken at random

we have to compute the probability, under H0, to obtain equal or more respondents with the label "Hacendado" than the observed number of respondents with the label "Hacendado"

$$\sum_{x=10}^{x=79} \text{Prob}(x, n_{i.}, n_{.j}, n) = \sum_{x=10}^{x=79} \frac{\binom{n_{i.}}{x}\binom{n-n_{i.}}{n_{.j}-x}}{\binom{n}{k_{.j}}}$$



1.4 Characterization of a categorical variable by a quantitative variable

Characterization of the categorical variable "*gama*" by the quantitative variables

Model : $Y_{ki} = \mu + \alpha_k + \varepsilon_{ki}$

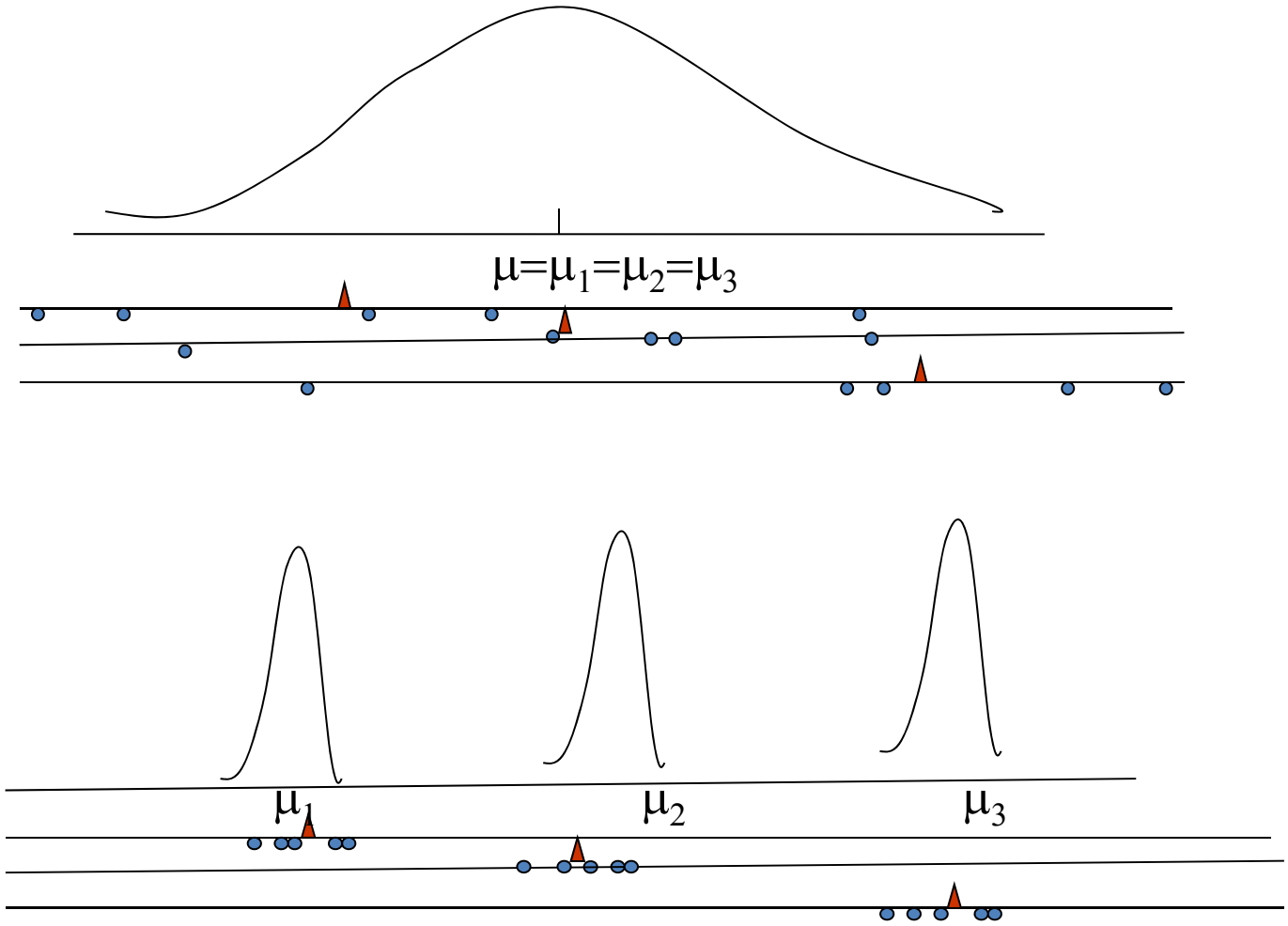
H_0 (no class effect): $\alpha_1 = \ldots = \alpha_k = \ldots = \alpha_K = 0$

H_1: There are at least two "gama" levels k and k' such as: $\alpha_k \neq \alpha_{k'}$

```
$quanti.var
                Eta2     P-value
granuloso 0.6381764 0.005573292
pegajoso  0.4599793 0.031110432
```

Example:

Assuming 3 levels



1.5 Profiling categories from quantitative variables

Characterization of the categories of "*gama*" by the quantitative variables

H_0 The coefficient of class *k* is null (mean in the category=global mean)

H_1: The coefficient of class *k* is not null (mean in the category≠global mean)

```
$quanti
$quanti$A
            v.test Mean in category Overall mean sd in category Overall sd
granuloso 2.396578         3.041667     2.490625      0.3482969  0.4515706
pegajoso  2.034653         4.260417     3.970665      0.2581148  0.2796842
              p.value
granuloso 0.01654895
pegajoso  0.04188579

$quanti$B
              v.test Mean in category Overall mean sd in category Overall sd
pegajoso  -2.034653         3.846486     3.970665      0.1783063  0.2796842
granuloso -2.396578         2.254464     2.490625      0.2311103  0.4515706
              p.value
pegajoso  0.04188579
granuloso 0.01654895
```

$$H_o \quad \mu_k = \mu$$

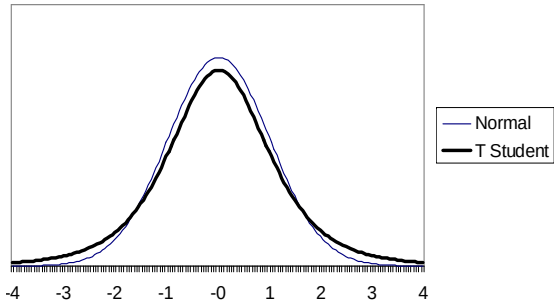
groups	*means*	*counts*
1	$\overline{x}_1$	n_1
⋮	⋮	⋮
p	$\overline{x}_p$	n_p
Global	$\overline{x}$	n

Test statistic: Diference between the mean in group *k* and the global mean but relative to the global variance and to the effectives

Given that $$s^2_{\overline{x}_k} = \frac{s^2}{n_k} \cdot \frac{n - n_k}{n-1}$$

William Gosset “Sudent”, English, 1876, 1937

Statistic $$\frac{\overline{x} - \overline{x}_k}{s_{\overline{x}_k}} = \frac{\left(\overline{x} - \overline{x}_k\right)\sqrt{n_k}}{s} \sqrt{\frac{n - n_k}{n-1}}$$ Student’s *t*



2. Characterization of a quantitative variable

2.1 by the other quantitative variables

2.2 by the categorical variables

2.3 by the categories of the other categorical variables

2.1 Relationship between the quantitative variable "aroma a cacao" and the other quantitative variables

```
> condes(base,num.var=6)
$quanti
             correlation      p.value
astringencia   0.9687514 4.017203e-06
s amargo       0.9498014 2.614201e-05
o_cacao        0.7912125 6.406573e-03
s_acido        0.7885560 6.715689e-03
crujiente      0.6596458 3.796569e-02
fusion        -0.7741130 8.584167e-03
a_caramelo    -0.7861651 7.002816e-03
o_leche       -0.7958340 5.893113e-03
a_vainilla    -0.9077084 2.835949e-04
a_leche       -0.9213895 1.518206e-04
s_azucar      -0.9457315 3.553049e-05
```

H_0 $\rho=0$

H_1: $\rho\neq 0$

Linear correlation

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

2.2 Relationship between the quantitative variable "aroma a cacao" and the categorical variables

$$\text{Model} \quad : \quad Y_{ki} = \mu + \alpha_k + \varepsilon_{ki}$$

H_0 (no class effect): $\alpha_1 = \ldots = \alpha_k = \ldots = \alpha_K = 0$

H_1: There are at least two "gama" levels k and k' such as: $\alpha_k \neq \alpha_{k'}$

```
$quali
                  R2     p.value
ConCacao 0.8429035 0.001536677
```

2.3 Relationship between the quantitative variable “aroma a cacao” and the categories

H_0 The coefficient of class *k* is null

H_1: The coefficient of class *k* is not null

```
$category
       Estimate     p.value
81to90 1.569444 0.003335757
```