Descriptive and bivariate analysis

Julio Abad González

Department of
Economics & Statistics

Outline

1. Introduction
2. Types of variables and measurements scales
3. Univariate analysis
4. Analysis of two qualitative variables
5. Analysis of one qualitative variable and one quantitative variable
6. Analysis of two quantitative variables
7. References

1. Introduction

What is descriptive statistics?

Descriptive statistics are a set of techniques used to describe the basic features of the data in a study and summarise the variables measured on the sample

- **Univariate analysis**

  Statistical methods that aim to describe and summarise the distribution of a single variable

- **Bivariate analysis**

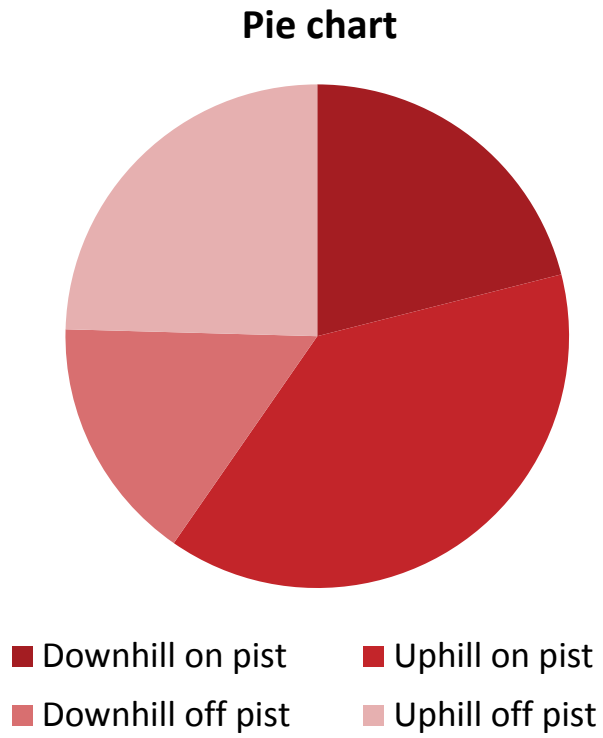
  Statistical methods that simultaneously analyse two variables measured on each individual under investigation for the purpose of determining the empirical relationship between these variables

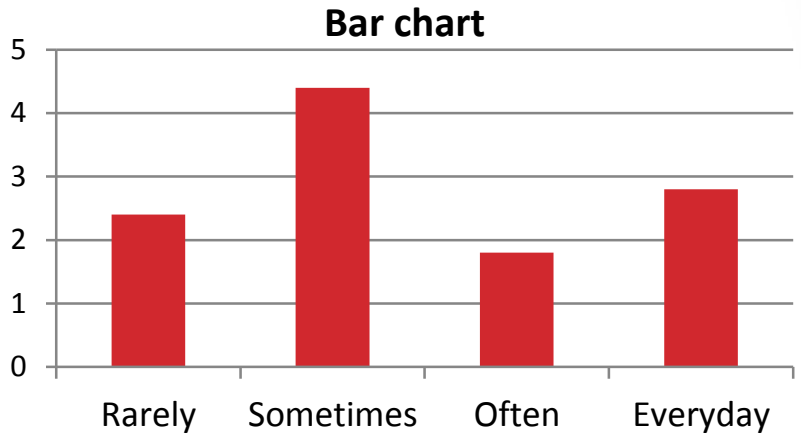
2. Types of variables and measurement scales

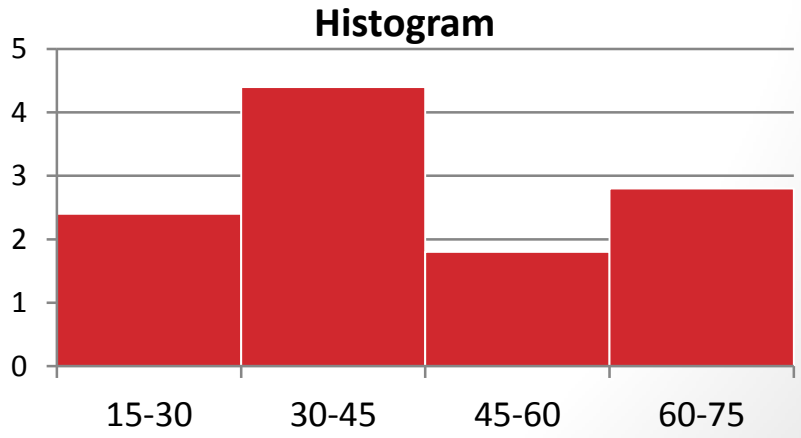
- **Qualitative, categorical or nonmetric variables:** Attributes, characteristics or properties used to identify or describe an individual whose values do not represent quantities or magnitudes
  - -- **Nominal scale** classifies data into distinct categories in which no ranking is implied
  - - **Ordinal scale** classifies data into values which could be ordered in a natural way
- **Quantitative, numerical or metric variables:** Characteristics whose values do represent quantities, amounts or magnitudes
  - + **Interval scale** is an ordered scale in which the difference between values is a meaningful quantity but whose zero point is arbitrary, not real (it does not indicate a zero amount or lack of variable)
  - ++ **Ratio scale** is an ordered scale in which the difference between values is a meaningful quantity and also has an absolute zero point that actually indicates the lack of variable

3. Univariate Analysis

- **Graphics:**







3. Univariate Analysis

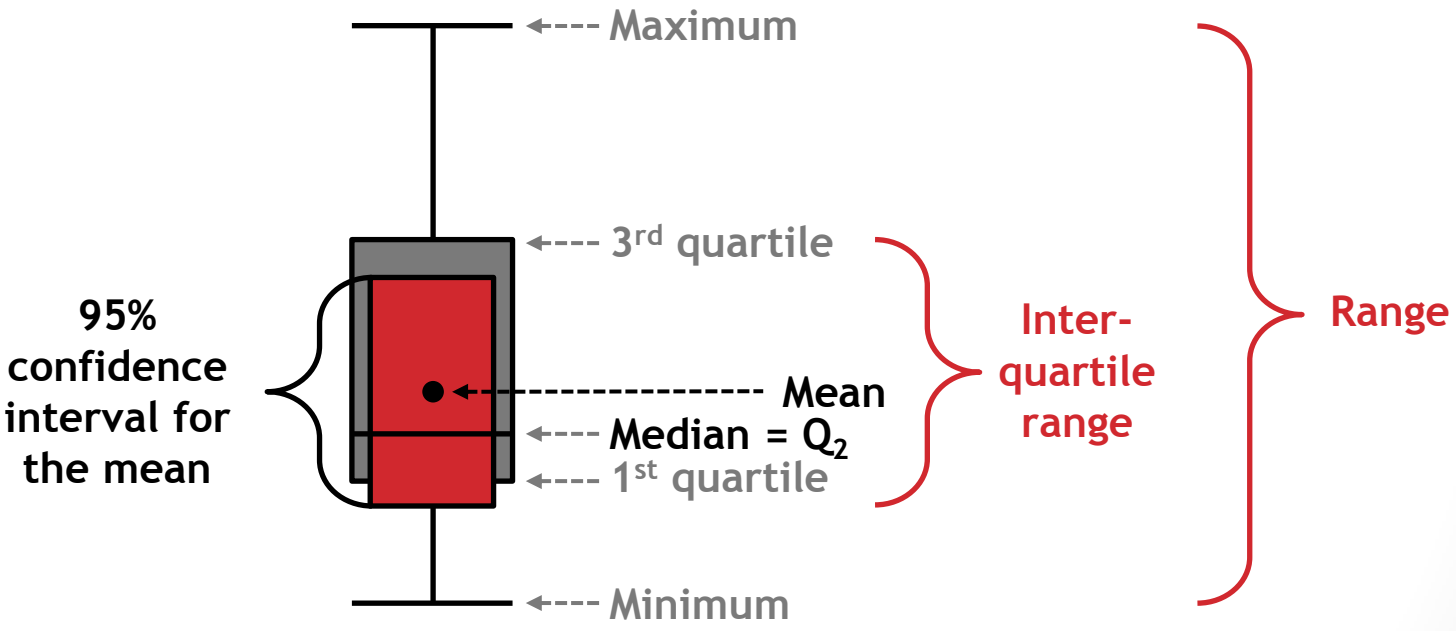
- **Graphics:** pie charts, bar charts and histograms
- **Summary statistics:**
  - **Central location:** a central or typical value for a data set
    - **Mode:** the most frequent value in the data set
    - **Median:** the middle value that divides the data set into to equal halves
    - **Mean:** the centre of gravity for all the data set
  - **Non-central location:** indicate the location of other key data values
    - **Minimum** and **Maximum**
    - **Quartiles:** three values (Q_1, Q_2 and Q_3) that divide the data set into four equal groups, each group comprising a quarter of the data
  - **Dispersion:** the spread or variability of the data
    - **Rank:** Maximum - Minimum
    - **Inter-quartile rank:** Q_3 - Q_1
    - **Standard deviation:** indicates how much variation or dispersion from the mean exists

3. Univariate Analysis

DESCRIPTIVE STATISTICS	Type of variable:	Qualitative		Quantitative	
	Measurement scale:	Nominal	Ordinal	Interval	Ratio
Graphs	Pie chart	x	x		
	Bar chart	x	x	x	x
	Histogram			x	x
	Box- plot			x	x
Measures of central location	Mode	x	x	x	x
	Median		x	x	x
	Mean			x	x
Measures of non-central location	Minimum / Maximum		x	x	x
	Quartiles		x	x	x
Measures of dispersion	Range			x	x
	Interquartile range			x	x
	Standard deviation			x	x

3. Univariate Analysis

- **Box-plot:** a graphic summary of the distribution



Central location measures
Non-central location measures
Dispersion measures

4. Analysis of two qualitative variables

a. Cross-tabulations and charts
b. Chi-square independence test
c. Simple Correspondence Analysis

4a. Cross-tabulations and charts

Contingency table (absolute frequencies):

- The X categories are located in rows
- The Y categories are located in columns
- Each cell contains the joint responses (n_{ij}) of the corresponding row (i) and column (j) categories

X \ Y	y_1	y_2	…	y_j	…	y_c	sum
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1c}	$\mathbf{n_{1.}}$
…	…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{ic}	$\mathbf{n_{i.}}$
…	…	…	…	…	…	…	…
x_r	n_{r1}	n_{r2}	…	n_{rj}	…	n_{rc}	$\mathbf{n_{r.}}$
sum	$\mathbf{n_{.1}}$	$\mathbf{n_{.2}}$	…	$\mathbf{n_{.j}}$	…	$\mathbf{n_{.c}}$	**n**

$$n_{i.} = \sum_{j=1}^{c} n_{ij} \qquad n_{.j} = \sum_{i=1}^{r} n_{ij} \qquad n = \sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{c} n_{.j} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}$$

4a. Cross-tabulations and charts

Contingency table (relative frequencies):

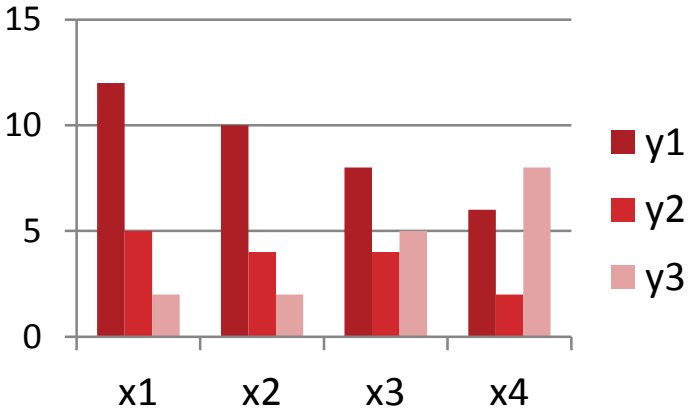
- Each cell contains the joint relative frequency or percentage of the overall total (f_{ij}) of the corresponding values/categories in row (i) and column (j):

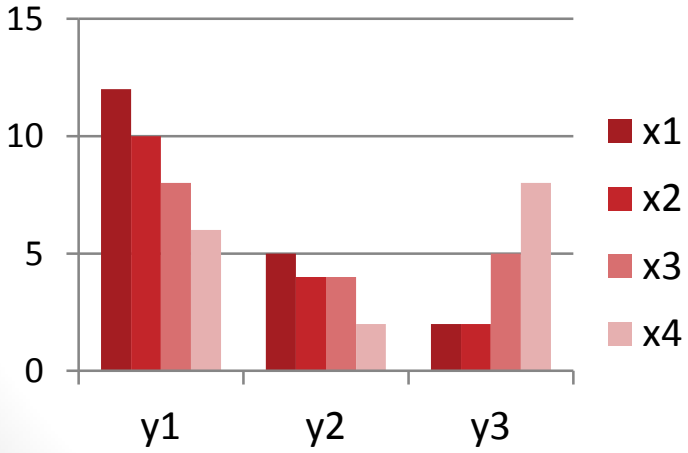
$$f_{ij} = n_{ij} / n$$

X \ Y	y_1	y_2	…	y_j	…	y_c	sum
x_1	f_{11}	f_{12}	…	f_{1j}	…	f_{1c}	$f_{1\cdot}$
…	…	…	…	…	…	…	…
x_i	f_{i1}	f_{i2}	…	f_{ij}	…	f_{ic}	$f_{i\cdot}$
…	…	…	…	…	…	…	…
x_r	f_{r1}	f_{r2}	…	f_{rj}	…	f_{rc}	$f_{r\cdot}$
sum	$f_{\cdot 1}$	$f_{\cdot 2}$	…	$f_{\cdot j}$	…	$f_{\cdot c}$	1

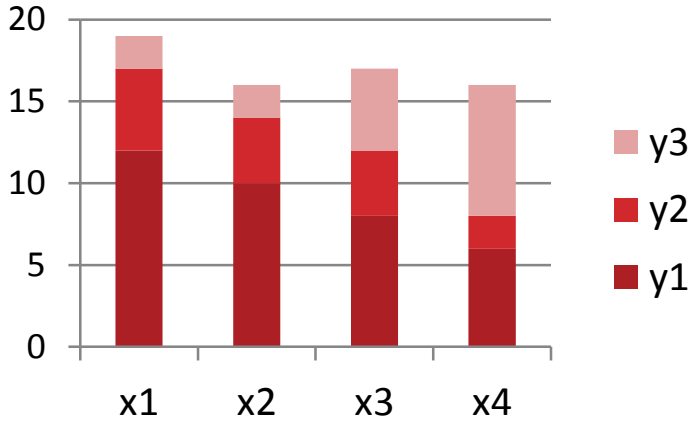
4a. Cross-tabulations and charts

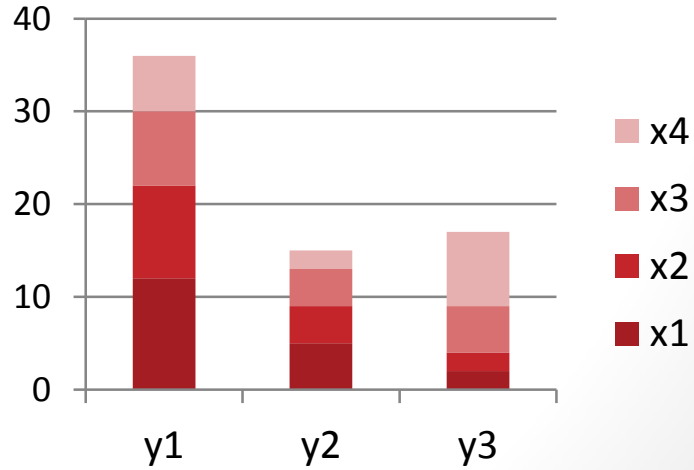
Adjacent bar charts





Stacked bar charts

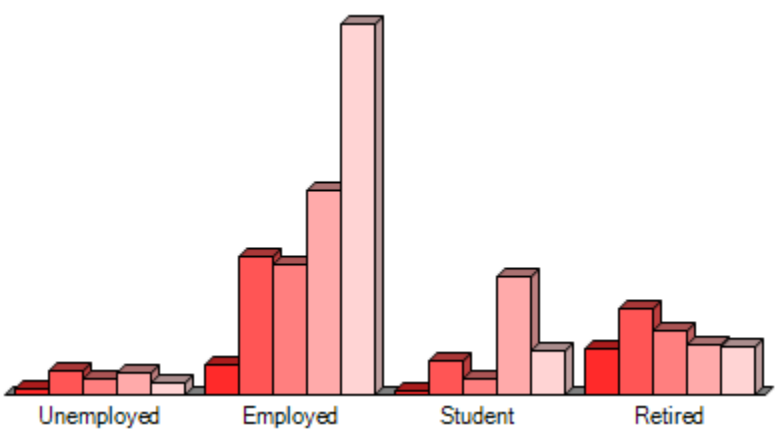


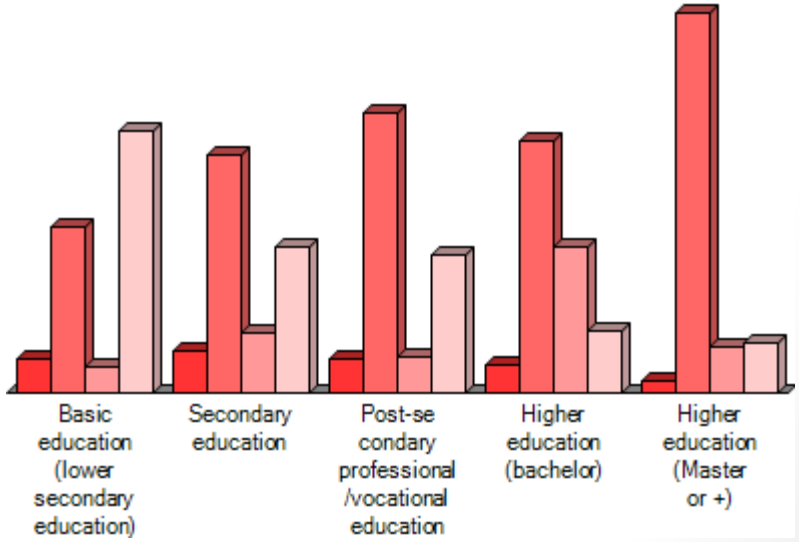


4a. Cross-tabulations and charts

Your education level (including in progress)
Your occupational status

	Basic education (lower secondary education)		Secondary education		Post-secondary professional /vocational education		Higher education (bachelor)		Higher education (Master or +)	
	N	% cit.	N	% cit.	N	% cit.	N	% cit.	N	% cit.
Unemployed	18	0.4%	72	2%	46	1%	65	2%	37	0.9%
Employed	86	2%	395	9%	370	9%	580	14%	1046	25%
Student	14	0.3%	100	2%	49	1%	338	8%	130	3%
Retired	135	3%	244	6%	183	4%	146	3%	139	3%





4a. Cross-tabulations and charts

Your education level (including in progress)
Your occupational status

	Basic education (lower secondary education)		Secondary education		Post-secondary professional /vocational education		Higher education (bachelor)		Higher education (Master or +)	
	N	% cit.	N	% cit.	N	% cit.	N	% cit.	N	% cit.
Unemployed	**18**	**0.4%**	**72**	**2%**	**46**	**1%**	**65**	**2%**	**37**	**0.9%**
Employed	**86**	**2%**	**395**	**9%**	**370**	**9%**	**580**	**14%**	**1046**	**25%**
Student	**14**	**0.3%**	**100**	**2%**	**49**	**1%**	**338**	**8%**	**130**	**3%**
Retired	**135**	**3%**	**244**	**6%**	**183**	**4%**	**146**	**3%**	**139**	**3%**

4a. Cross-tabulations and charts

Row-profile table

- Contains the conditional frequencies or percentages of each combination row-column (i, j) with respect to its row total (i):

$$f_{j/x_i} = n_{ij} / n_{i\cdot}$$

X \ Y	y_1	y_2	…	y_j	…	y_c	sum
x_1	f_{1/x_1}	f_{2/x_1}	…	f_{j/x_1}	…	f_{c/x_1}	1
x_2	f_{1/x_2}	f_{2/x_2}	…	f_{j/x_2}	…	f_{c/x_2}	1
…	…		…	…	…	…	1
x_i	f_{1/x_i}	f_{2/x_i}	…	f_{j/x_i}	…	f_{c/x_i}	1
…	…		…	…	…	…	1
x_r	f_{1/x_r}	f_{2/x_r}	…	f_{j/x_r}	…	f_{c/x_r}	1
Average profile	$f_{\cdot 1}$	$f_{\cdot 2}$	…	$f_{\cdot j}$	…	$f_{\cdot c}$	1

4a. Cross-tabulations and charts

Column-profile table

- Contains the conditional frequencies or percentages of each combination row-column (i, j) with respect to its column total (j):

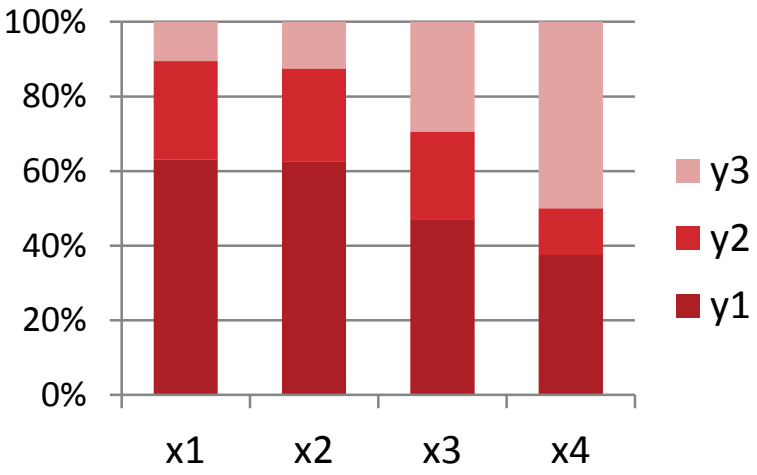
$$f_{i/y_j} = n_{ij} / n_{\cdot j}$$

X \ Y	y_1	y_2	…	y_j	…	y_c	Average profile
x_1	f_{1/y_1}	f_{1/y_2}	…	f_{1/y_j}	…	f_{1/y_c}	$f_{1\cdot}$
x_2	f_{2/y_1}	f_{2/y_2}	…	f_{2/y_j}	…	f_{2/y_c}	$f_{2\cdot}$
…	…	…	…	…	…	…	…
x_i	f_{i/y_1}	f_{i/y_2}	…	f_{i/y_j}	…	f_{i/y_c}	$f_{i\cdot}$
…	…	…	…	…	…	…	…
x_r	f_{r/y_1}	f_{r/y_2}	…	f_{r/y_j}	…	f_{r/y_c}	$f_{k\cdot}$
sum	1	1	1	1	1	1	1

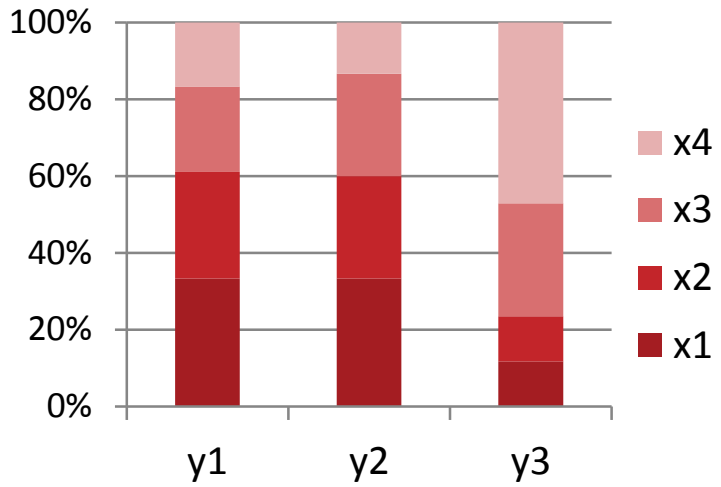
4a. Cross-tabulations and charts

100% Stacked bar charts

Row-profiles



Column-profiles



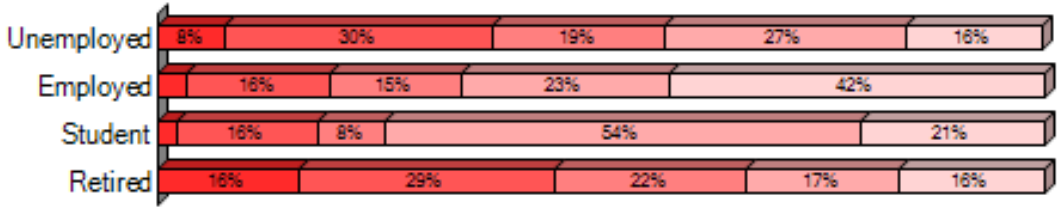
4a. Cross-tabulations and charts

Row-profiles

Your education level (including in progress)
Your occupational status

	Basic education (lower secondary education)		Secondary education		Post-secondary professional /vocational education		Higher education (bachelor)		Higher education (Master or +)		**Total**	
	N	% cit.	N	% cit.	N	% cit.	N	% cit.	N	% cit.	N	% cit.
Unemployed	**18**	**8%**	**72**	**30%**	**46**	**19%**	**65**	**27%**	**37**	**16%**	**238**	**100%**
Employed	**86**	**3%**	**395**	**16%**	**370**	**15%**	**580**	**23%**	**1046**	**42%**	**2477**	**100%**
Student	**14**	**2%**	**100**	**16%**	**49**	**8%**	**338**	**54%**	**130**	**21%**	**631**	**100%**
Retired	**135**	**16%**	**244**	**29%**	**183**	**22%**	**146**	**17%**	**139**	**16%**	**847**	**100%**

The relation is very significant



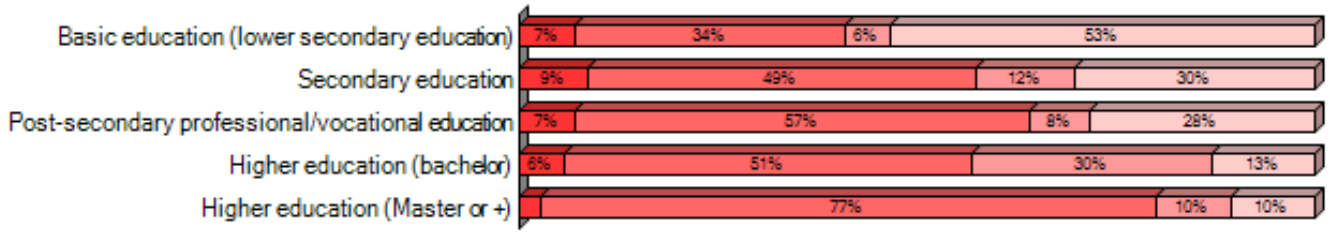
4a. Cross-tabulations and charts

Column-profiles

Your occupational status
Your education level (including in progress)

	Unemployed		Employed		Student		Retired		**Total**	
	N	% cit.	N	% cit.	N	% cit.	N	% cit.	N	% cit.
Basic education (lower secondary education)	**18**	**7%**	**86**	**34%**	**14**	**6%**	**135**	**53%**	**253**	**100%**
Secondary education	**72**	**9%**	**395**	**49%**	**100**	**12%**	**244**	**30%**	**811**	**100%**
Post-secondary professional/vocational education	**46**	**7%**	**370**	**57%**	**49**	**8%**	**183**	**28%**	**648**	**100%**
Higher education (bachelor)	**65**	**6%**	**580**	**51%**	**338**	**30%**	**146**	**13%**	**1129**	**100%**
Higher education (Master or +)	**37**	**3%**	**1046**	**77%**	**130**	**10%**	**139**	**10%**	**1352**	**100%**

The relation is very significant



4b. Chi-square independence test

- Allows to test if two categorical variables (X and Y) are independent (there is no relationship between them) or not
- It is based on the comparison between the observed frequencies (n_{ij}) and the expected frequencies if X and Y were independent (e_{ij})

$$e_{ij} = \frac{n_{i.} \times n_{.j}}{n}$$

- When the expected frequency of a cell is much higher than the corresponding observed one ($n_{ij} >> e_{ij}$) there is an **attraction** between the categories x_i and y_j (that combination row-column is **over-represented**)
- When an expected frequency of a cell is much lower than the corresponding observed one ($n_{ij} << e_{ij}$) there is a **repulsion** between the categories x_i and y_j (that combination row-column is **under-represented**)

4b. Chi-square independence test

- The chi-square test statistic is computed by adding all these differences after having them squared and standardised:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

- This test statistic follows a Chi-Square Distribution: $\chi^2 \underset{e_{ij} \geq 5}{\rightarrow} \chi^2_{(r-1)\times(c-1)}$
- Since this test statistic measures the squared standardised differences between the real frequencies and the theoretical ones (under the independence hypothesis), **the bigger the computed test statistic is, the more significant the relationship between X and Y is**
- It is also possible to determine which cells are the most significantly overrepresented and underrepresented, i.e., have the highest contributions to the chi-square test statistic

4b. Chi-square independence test

Your occupational s... → / Your education leve... ↓	Unemployed		Employed		Student		Retired		Total
	Freq.	Deviation	Freq.	Deviation	Freq.	Deviation	Freq.	Deviation	Freq.
Basic education (lower secondary education)	18		86	- VS	14	- VS	135	+ VS	253
Secondary education	72	+ VS	395	- VS	100	- S	244	+ VS	811
Post-secondary professional/vocational education	46	+ LS	370		49	- VS	183	+ VS	648
Higher education (bachelor)	65		580	- VS	338	+ VS	146	- VS	1129
Higher education (Master or +)	37	- VS	1046	+ VS	130	- VS	139	- VS	1352
Total	238		2477		631		847		4193

Values in blue/pink are significantly over-represented/under-represented (with a level of risk of 5%)

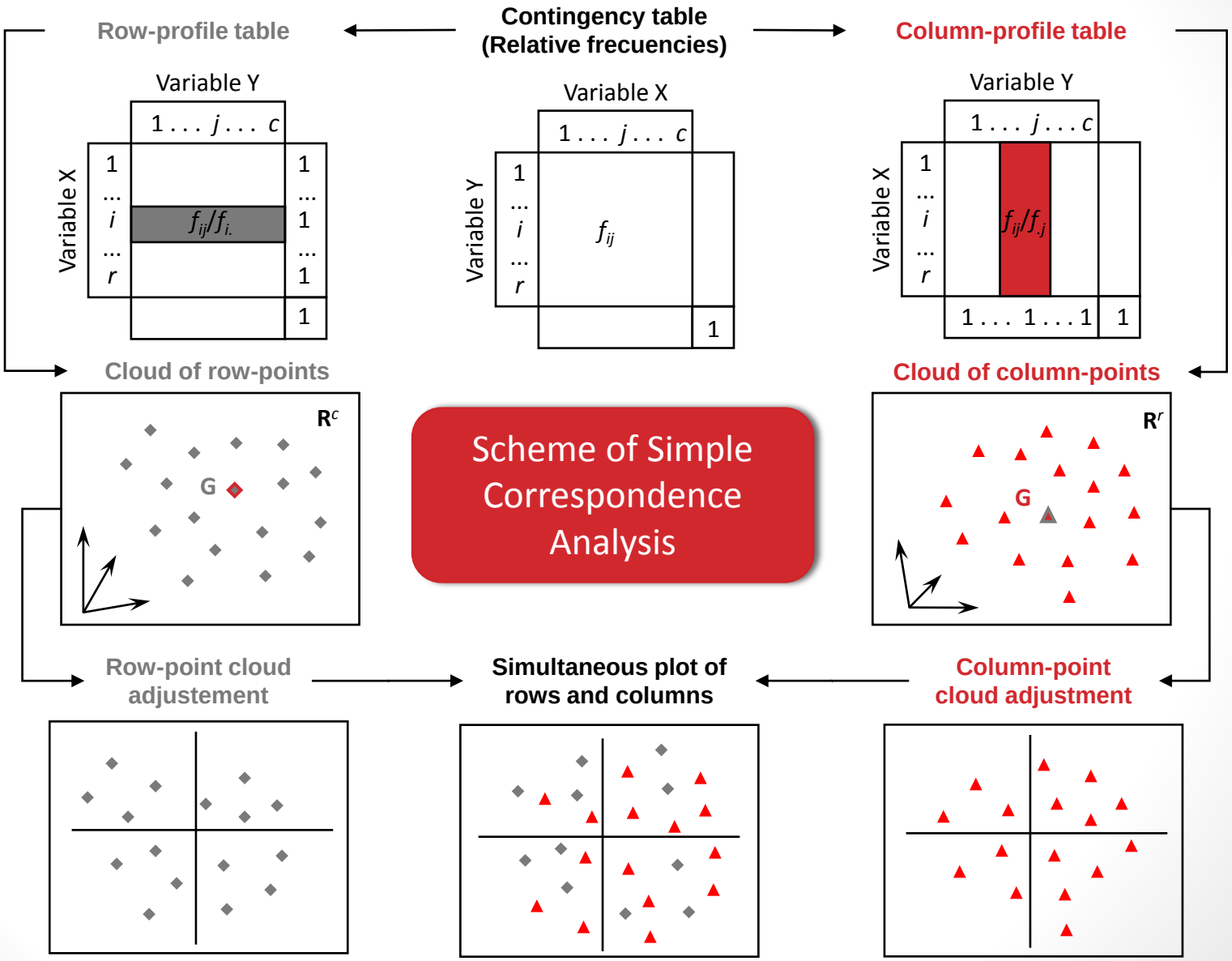
Actual responses : 4193 Non-response(s) : 72 Response rate : 98,31%
P-Value: = < 0.01% ; Khi2 = 697,77 ; dof = 12 (The relationship is very significant)

4c. Simple correspondence analysis

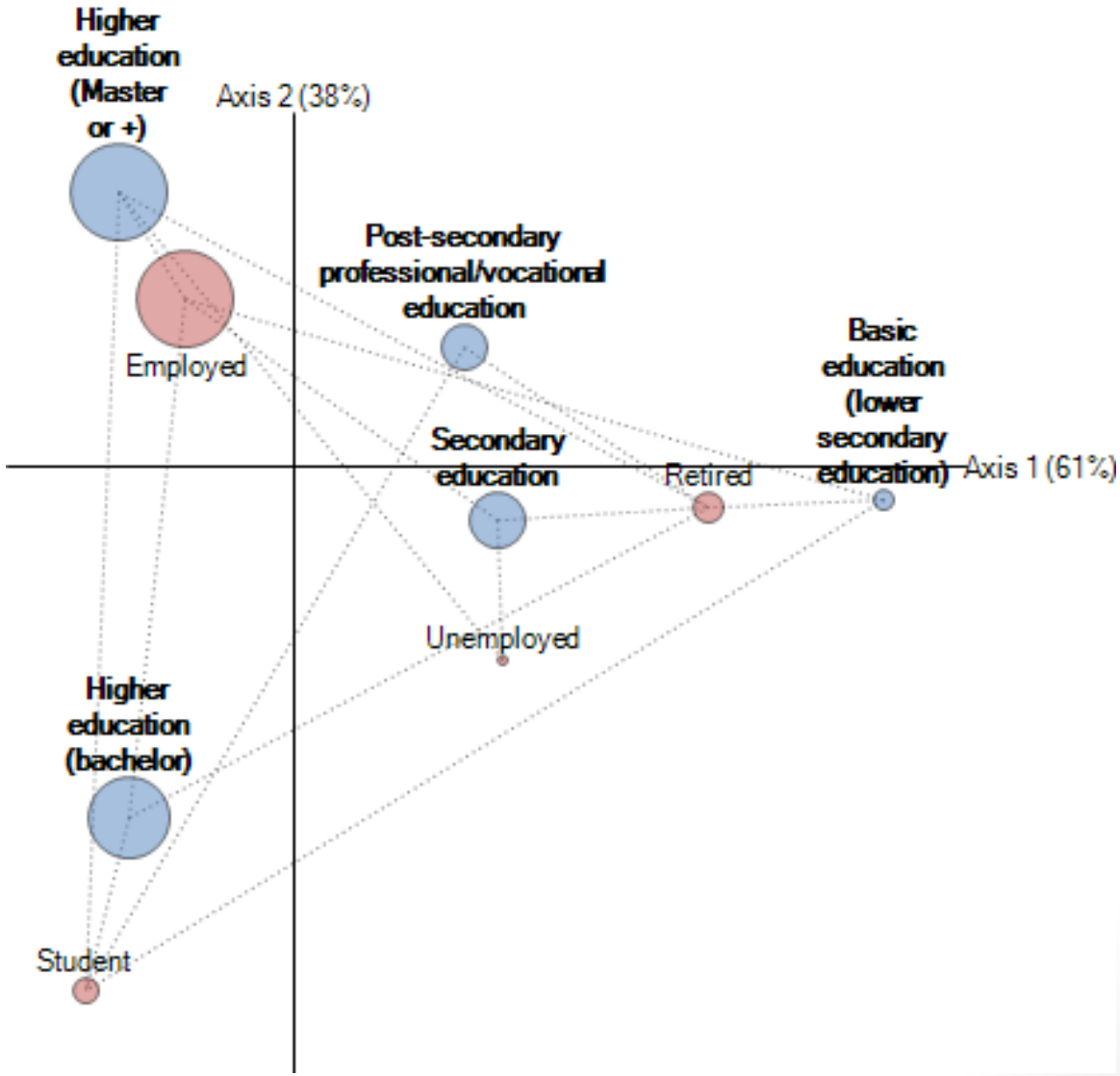
SCA (a.k.a. Correspondence Factor Analysis) is a data analysis technique that:

- Performs a **dimensional reduction** on a contingency table obtaining new factors or dimensions with a minimum loss of information
- Allows to **simultaneously graph on a scatterplot the categories of the two variables according to the attraction/repulsion among them** (in the same vein as in the chi-square independence test):
  - The more attraction exists between a row and a column, the closer they will be on the plot
  - The more repulsion exists between a row and a column, the further they will be on the plot
  - Rows and columns on central positions correspond to average profiles (close to independence)
  - Rows and columns on peripheral positions correspond to atypical profiles (far from independence)

4c. Simple correspondence analysis



4c. Simple correspondence analysis



5. Analysis of one qualitative variable and one quantitative variable

a. Table of means and box-plot
b. Analysis Of Variance (one-way ANOVA)

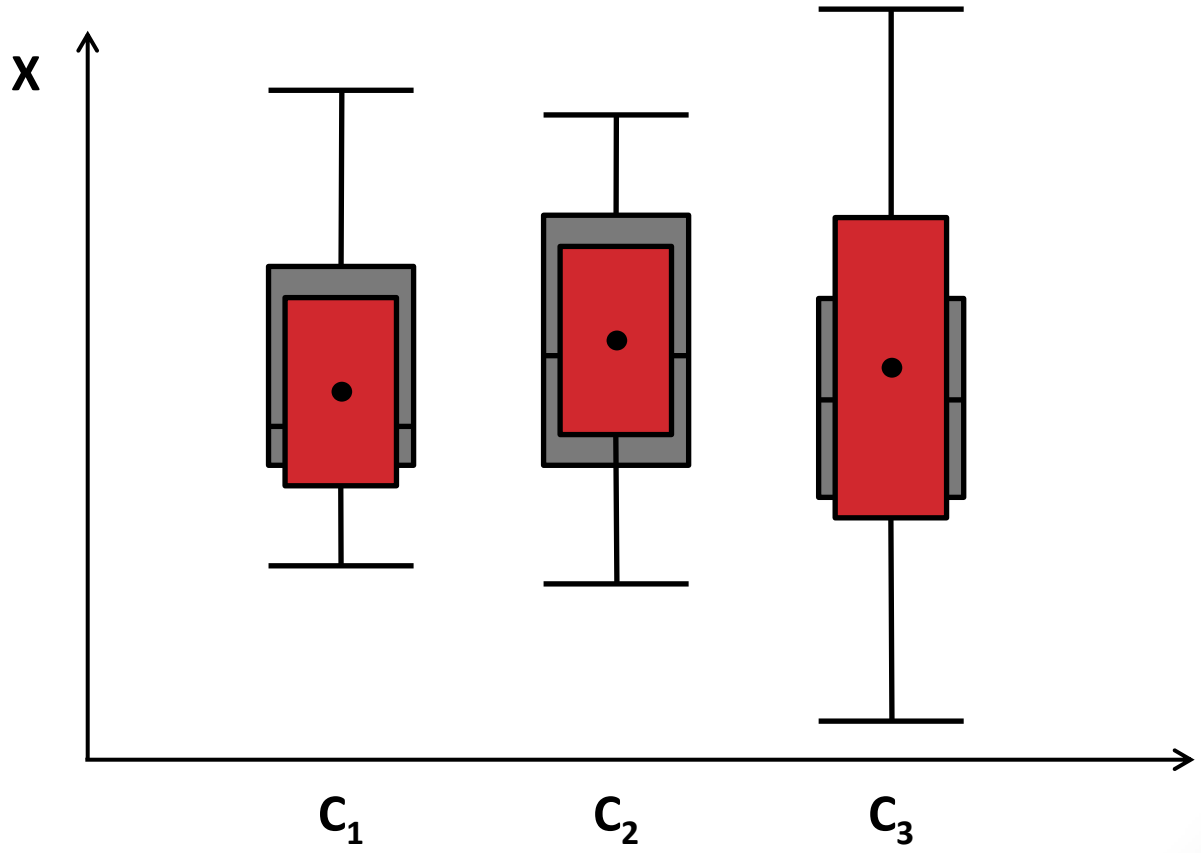
5a. Table of means and box-plots

- The aim is to compare summary statistics of the numerical variable (X) across the k different groups of individuals that define the categorical variable (C)

Summary statistics of X	Categories					Overall
	C_1	…	C_j	…	C_k	
Mean	$\bar{x}_1$	…	$\bar{x}_j$	…	$\bar{x}_k$	$\bar{x}$
Standard deviation	S_1	…	S_j	…	S_k	S
Median	Me_1	…	Me_j	…	Me_k	Me
Minimum	Min_1	…	Min_j	…	Min_k	Min
Maximum	Max_1	…	Max_j	…	Max_k	Max

5a. Table of means and box-plots

- A multiple box-plot allows to make this comparison graphically



5b. Analysis of variance (ANOVA)

- The aim of one-way ANOVA is to test if there are significant differences among the means of the numerical variable (called dependent or response variable) across the k different groups of individuals that define the categorical variable (usually called factor or treatment)
- Two independent estimates of the variance for the dependent variable are compared:
  - **Mean square between groups(MS_B):** reflects the variability among the means of the different groups/categories
  - **Mean square within groups (MS_W):** reflects the variability among the individuals belonging to the same group/category
- The test statistic: **$F = MS_B/MS_W$** follows a **$F_{k-1, n-k}$**
- The bigger the computed test statistic is, the more significant the differences among the group means are

5b. Analysis of variance (ANOVA)

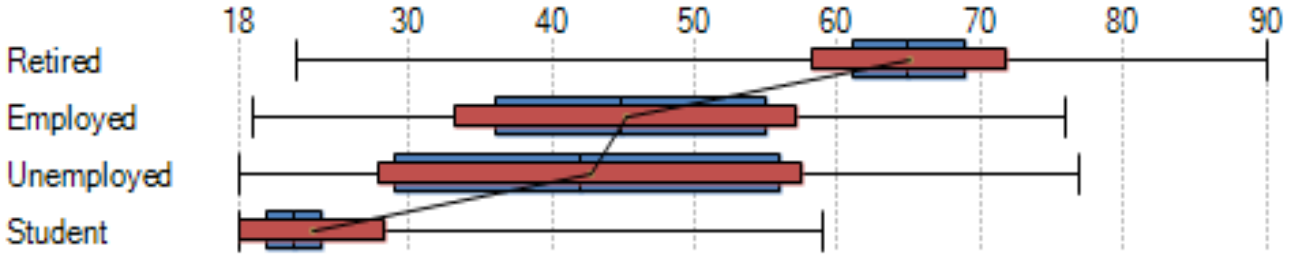
Age

Your occupational status

	Age				
	Mean	Std deviation	Min	Max	Median
Retired	64.9	6.9	22	90	65.0
Employed	45.2	12.1	19	76	45.0
Unemployed	42.7	14.9	18	77	42.0
Student	23.1	5.2	18	59	22.0

p = **<1%** ; F = **1906.0** (**VS**)

The relation is very significant
elements over (under) represented are coloured.

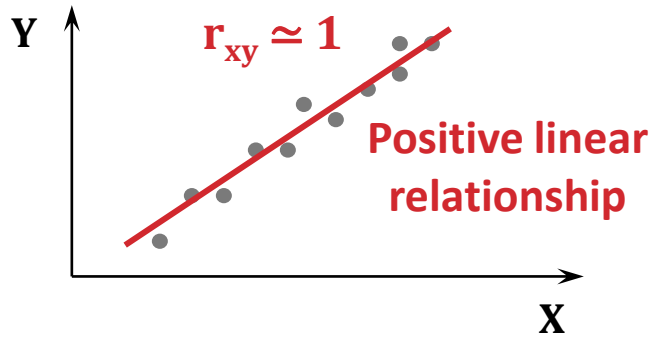


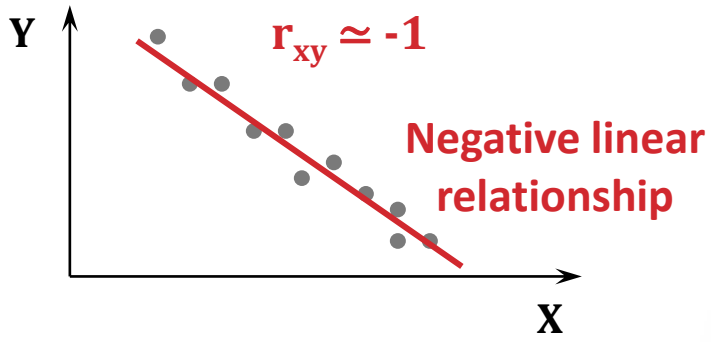
6. Analysis of two quantitative variables

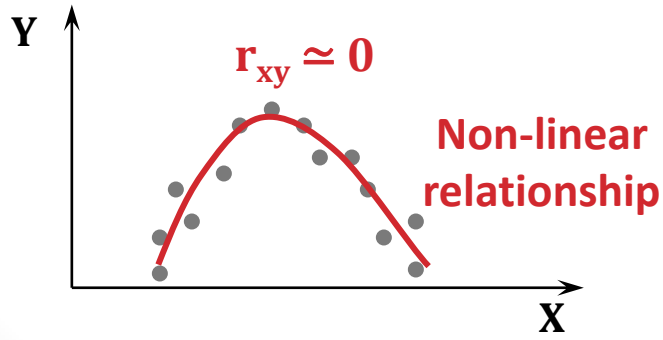
a. Scatterplot and coefficient of correlation
b. Simple regression analysis

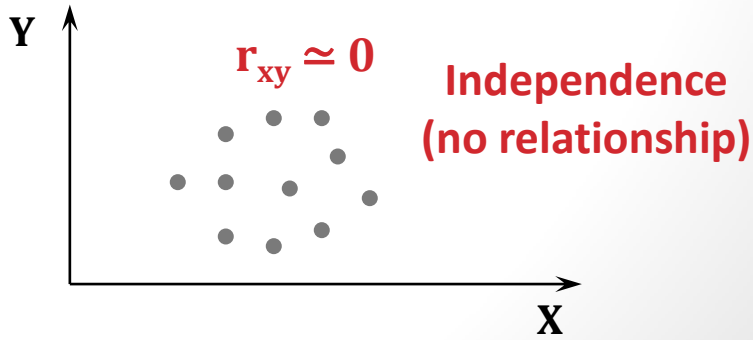
6a. Scatterplot and coefficient of correlation

- Scatterplot shows graphically the nature of the relationship between two quantitative variables (if there is any)
- Coefficient of correlation measures the strength and direction of a linear relationship between two quantitative variables



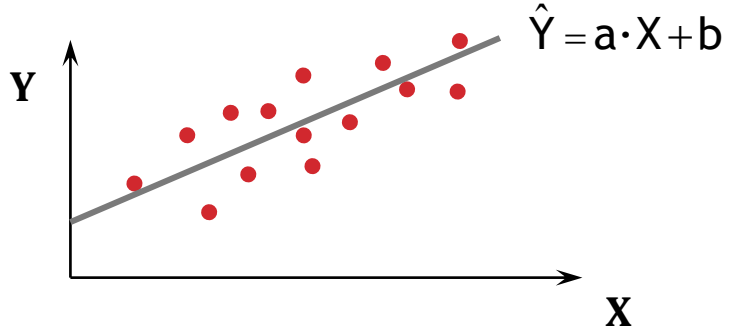






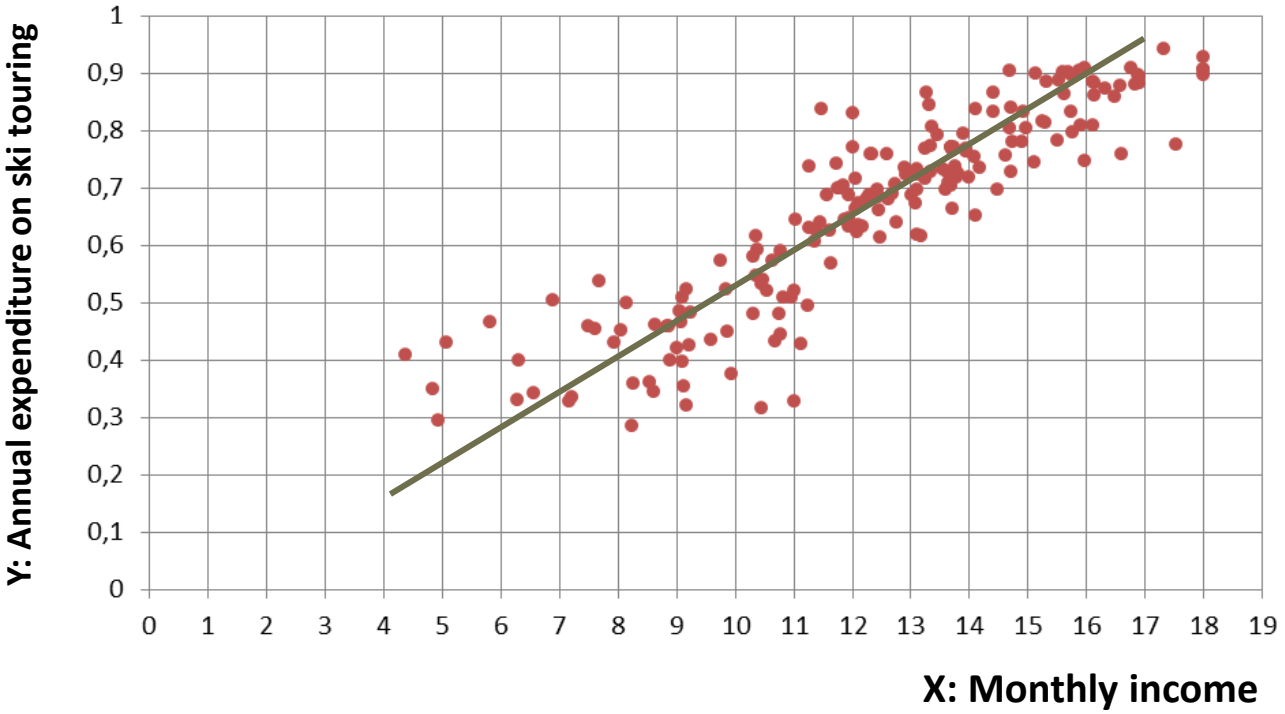
6b. Simple linear regression

- Aims to develop a linear model to predict the values of a numerical variable, called dependent or response (Y), based on the value of another numerical variable, called independent or explanatory (X)
- Simple Linear Regression Model: **$Y=a\cdot X+b+\varepsilon$**



- **Intercept (b):** Mean value of Y when X equals 0
- **Slope (a):** Expected change in Y per unit change in X
  - **a>0** → Positive linear relationship (Y increases as X increases)
  - **a<0** → Negative linear relationship (Y increases as X decreases)
- **Coefficient of determination R^2:** goodness of fit measure that indicates the percentage of the variance of Y that is explained by the model

6b. Simple linear regression



7. References

- Greenacre, M.J. (1993): *Correspondence Analysis in Practice*. London: Academic Press Ltd.
- Le Sphinx Développement (2006): *Sphinx Handbook.* http://www.lesphinx-developpement.fr
- Lebart L.; Morineau A.; Warwick K. (1984): *Multivariate Descriptive Statistical Analysis*. New York: Wiley.
- Newbold, P.; Carlson, W.L.; Thorne, B.M. (2013): *Statistics for Business and Economics* (8th ed.) Essex: Pearson Education Ltd.

Thanks for your attention

Julio Abad González

Department of
Economics & Statistics