



# New Forms for Business Intelligence

A concrete approach to the use of data  
to improve and quicken business  
decisions

# Agenda

## Module 3: Univariate & Bivariate Analysis

- Introduction
- Univariate Analysis
- Bivariate Analysis

# Statistical analysis of the data

***Descriptive statistics:*** the methods regarding the representation and the synthesis of a group of data in order to point out their main characteristics

***Inferential Statistics:*** the methods which enable the estimate of the characteristic of a population based on sampling analysis

```
graph TD; A[All of the elements considered in the survey] --> B[Subset of the population Selected for the analysis]; B --> C[Summary statistics assessed on a sampling data needed to describe a unknown characteristic of the Population.];
```

All of the elements considered in the survey

Subset of the population  
Selected for the analysis

Summary statistics assessed on a sampling data needed to describe a unknown characteristic of the Population.

# Type of data

- **Qualitative:** data expressed verbally, classified in categories
- **Quantitative:** data expressed numerically and they are:
  - **Discrete:** data expressed by indivisible quantities i.e. family size
  - **Continuous:** data expressed by infinitely divisible unit of measurement i.e. (kilometers, meters, centimeters, millimeters,...)

# Type of data - *qualitative*

- **Nominal** it's used for qualitative data which are classified in defined categories with no a specific order

Where do you come from?	
a. North Italy	<input type="checkbox"/>
b. Center Italy	<input type="checkbox"/>
c. South Italy	<input type="checkbox"/>
d. Outside Italy	<input type="checkbox"/>

- **Ordinal** the categories have got a specific order; it enables to define a classification order among the categories but it does not enable to define any numeric assessment.

Education level (Currently Attending)	
a. High School	<input type="checkbox"/>
b. Undergraduate	<input type="checkbox"/>
c. Graduate	<input type="checkbox"/>

# Type of data - *quantitative*

- **Ratio scale** through this type of data it is possible to determine the different ratio between one category and an other; the value “0” of the scale is set.

i.e. The variables *average expense* and *amount of time needed* are measured in terms of ratio, that is they are within the framework of a comparative evaluation scale.

How long have you been a Facebook user for (Months):

\_\_\_\_\_

Approximately, how many friends do you have on Facebook:

\_\_\_\_\_

How many of these friends do you contact regularly:

\_\_\_\_\_

On average, how many times a week do you check Facebook:

\_\_\_\_\_

How much time do you spend on each visit (in minutes):

\_\_\_\_\_

## Type of data - *quantitative*

- **Interval scale** has the same characteristics as the previous scale, even though it has not got a fixed value “0”.  
i.e. In a survey done on the customers of a supermarket, their degree of satisfaction can be appropriately measured through an evaluation scale between 1 and 9. What I can say is the difference between values 2 and 3 is exactly the same as between values 8 and 9, but I cannot state that the value 8 is twice as much as value 4.

Where do you connect on Facebook more frequently?	Low			Medium			High		
	1	2	3	4	5	6	7	8	9
	a. Home,								
b. Work/ University									
c. Other places (internet point, friends' houses ..)									

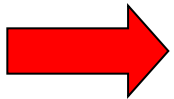
# Type of data

- **Qualitative**

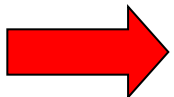
- **Nominal** it's used for qualitative data which are classified in defined categories with no a specific order.
- **Ordinal** the categories have got a specific order; it does not enable to define any numeric assessment.

- **Quantitative**

- **Ratio scale** through this type of data it is possible to determine the different ratio between one category and another; the value “0” of the scale is set.
- **Interval scale** has the same characteristics as the previous scale, even though it has not got a fixed value “0”.



Type of data guides the analyses



Most of the quantitative methods deal with quantitative data



# Keywords

- Descriptive Statistics
- Inferential Statistics
- Qualitative Data
  - Nominal
  - Ordinal
- Quantitative Data
  - Ratio Scale
  - Interval Scale
  - Discrete
  - Continuous

# Agenda

## Module 3: Univariate & Bivariate Analysis

➤ Introduction

➤ Univariate Analysis

➤ Bivariate Analysis

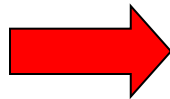
# Univariate descriptive statistics

In the univariate descriptive statistics we analyze one variable at a time.

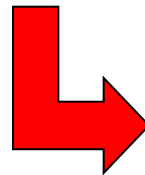
N_ID	D_8_2
H1	0.1
H2	0
H3	0
H4	0.2
H5	0.05
H6	0.2
H7	0.1
H8	0.1
H9	0.2
H10	0.05
H11	0
H12	0
H13	0
H14	0.15
H15	0
H16	0.1
H17	0
H18	0.2
H19	0
H20	0.05
H21	0.2
H22	0.2

...

H234	0.2
H235	0.1
H236	0.1



- Frequency distribution
- Synthesis measures
  - *Measures of location*
  - *Measures of spread*
  - *Measures of shape*



- Data Audit
  - Input errors
  - Missing values
- Basic insights

# Univariate descriptive statistics

In the univariate descriptive statistics we can use two major methods to analyze data:

- Frequency distributions
- Synthesis measures:
  - *Measures of location;*
  - *Measures of spread;*
  - *Measures of shape*

# Frequency distributions

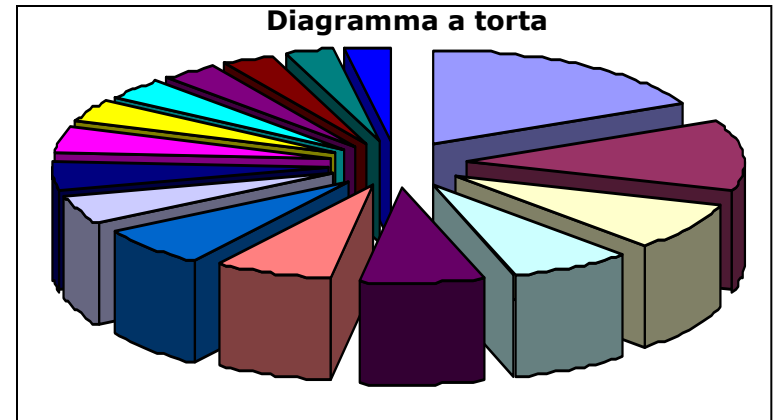
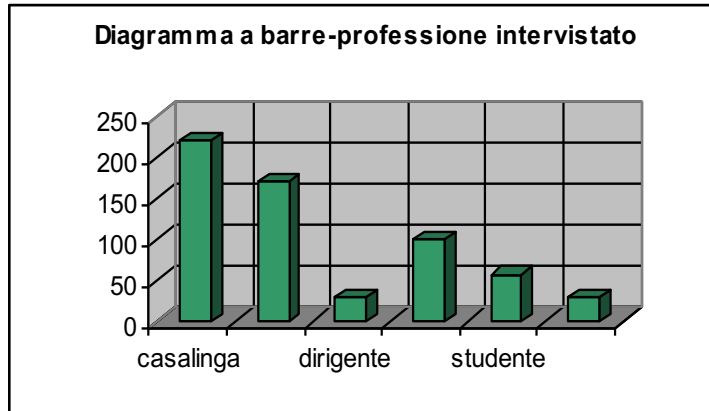
- *Frequency*: this is the first level of data synthesis and consists of counting the number of times each category is shown in the data
- *Frequency distribution*: the set of the categories and their frequencies
- *Relative frequency*: it is the ratio between the category frequency and total number of the units considered in the population.

$$p_i = n_i / N$$

Both types of frequencies are used with nominal, ordinal and quantitative data.

# Frequency distributions

*Graphic description of qualitative variables:*

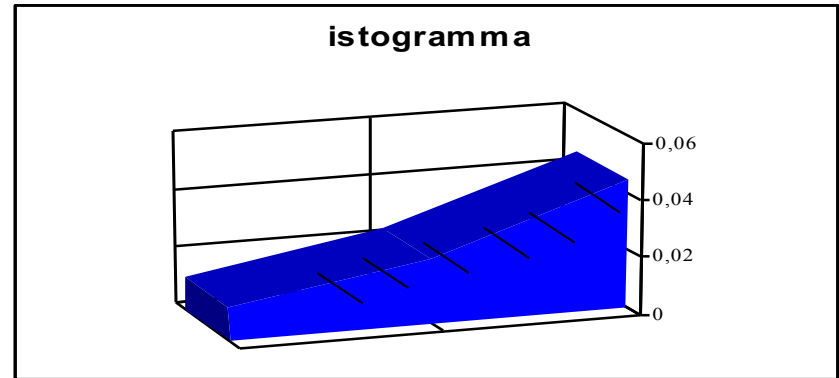
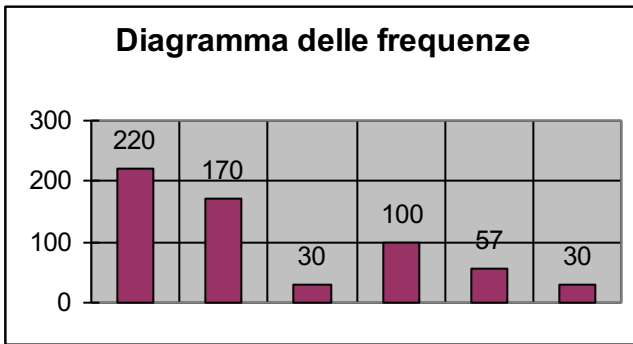


**Barred chart:** in the horizontal line we have got the categories at random, in the vertical their frequencies

**Pie chart:** the pie is proportionally divided according to the categories

# Frequency distributions

## *Graphic description of quantitative variables*

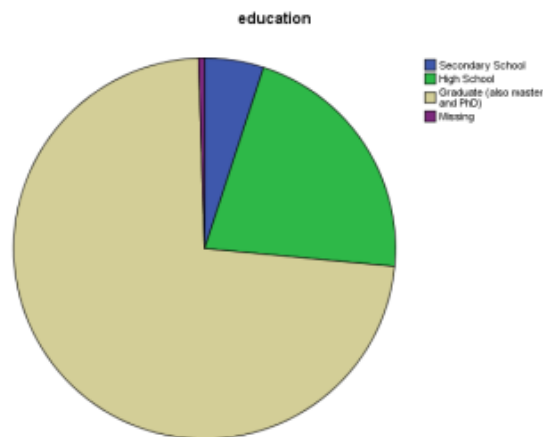


**Frequency Chart:** in the abscissa axis there the discrete variable values; the height of the bars is proportional to the frequencies (absolute or relative) of the value itself

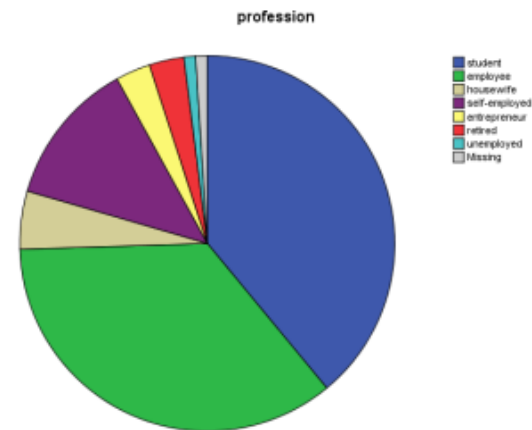
**Histogram:** in the horizontal axis the classes of the intervals taken into consideration, the vertical axis shows the density of the frequency, the rectangle area represents the frequency of the class itself.

# Frequency distributions

## Descriptive Statistics: Who is our sample?



It is a very educated sample, more than 73% is a graduate or post-graduate. This result is coherent with the image and offerings of the cinema.



Our respondents were mainly students and employees, and these two segments account for the approximately 74% of the sample.





# Univariate statistics

## *Measures of location:*

- Mean
- Median
- Mode
- Quantiles
- Percentiles

## *Measures of spread:*

- Range
- Interquantile Range
- Variance
- Standard Deviation
- Coefficient of Variation

## *Measures of shape:*

- Skewness
- Kurtosis

# Central Statistics Measures

## Statistics

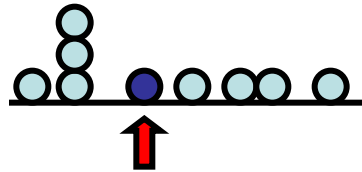
```
graph TD; Statistics[Statistics] --- Average[Average]; Statistics --- Median[Median]; Statistics --- Mode[Mode];
```

### Average

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

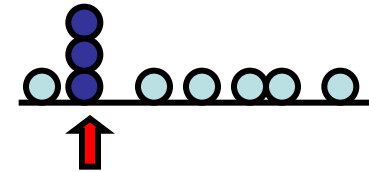
Arithmetic Mean

### Median



The central value of sorted observations

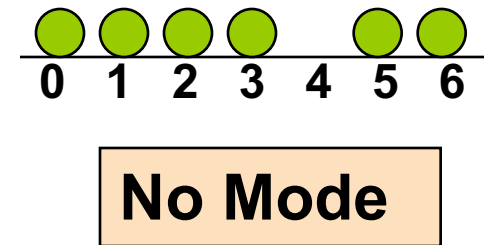
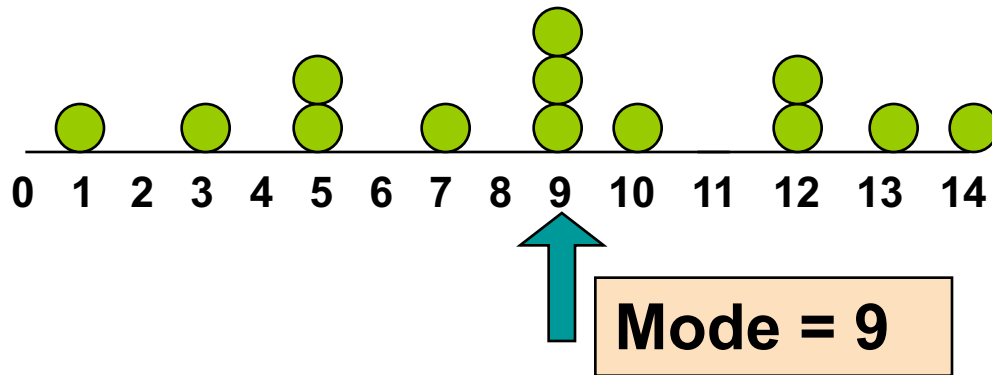
### Mode



The most frequent value

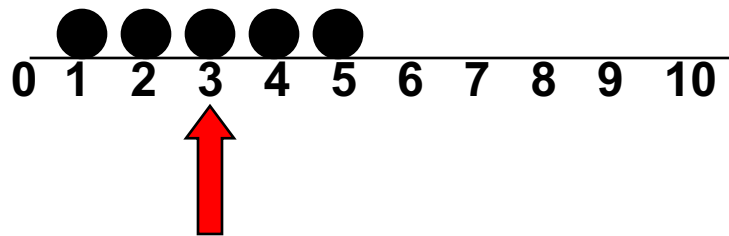
# Mode

- It is the value that occurs most frequently
- It is not influenced by any outliers
- It is used both for quantitative and qualitative data
- There may not be a mode
- There may be more than one mode



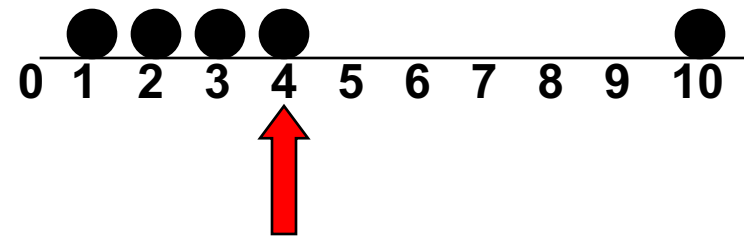
# Arithmetic Mean

- It is the most common central statistics of trend
- Average = sum of the values divided by the number of the observations
- Affected by outliers



**Mean = 3**

$$\frac{1 + 2 + 3 + 4 + 5}{5} = \frac{15}{5} = 3$$

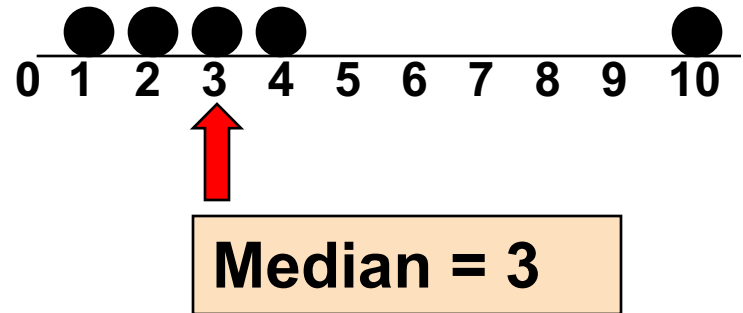
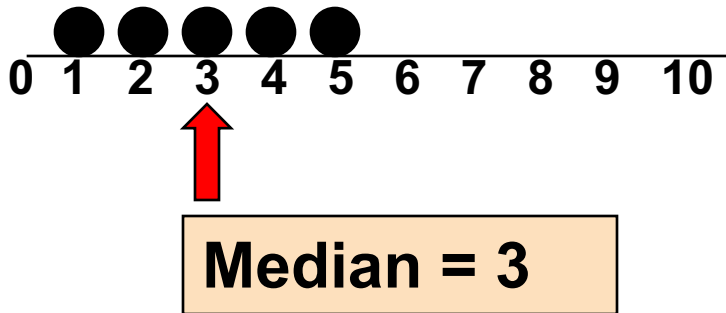


**Mean = 4**

$$\frac{1 + 2 + 3 + 4 + 10}{5} = \frac{20}{5} = 4$$

# Median

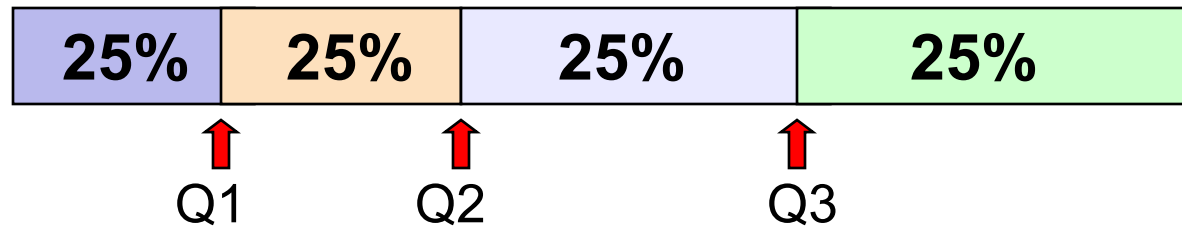
- In a sorted list, the median is the central value of the distribution (50% to its left, 50% to its right)



- It's not influenced by any outliers

# Non Central Statistics of Trend

- The quartiles split the sorted sequence of data into 4 segments with the same number of observations (units)



- The first quartile,  $Q_1$ , is the value which splits 25% of units to the left and 75% to the right
- $Q_2$  is the median value (50% to the left, 50% to the right)
- $Q_3$  splits only 25% of units to the right, consequently 75% is to the left

# Univariate statistics

## *Measures of location:*

- Mean
- Median
- Mode
- Quantiles
- Percentiles

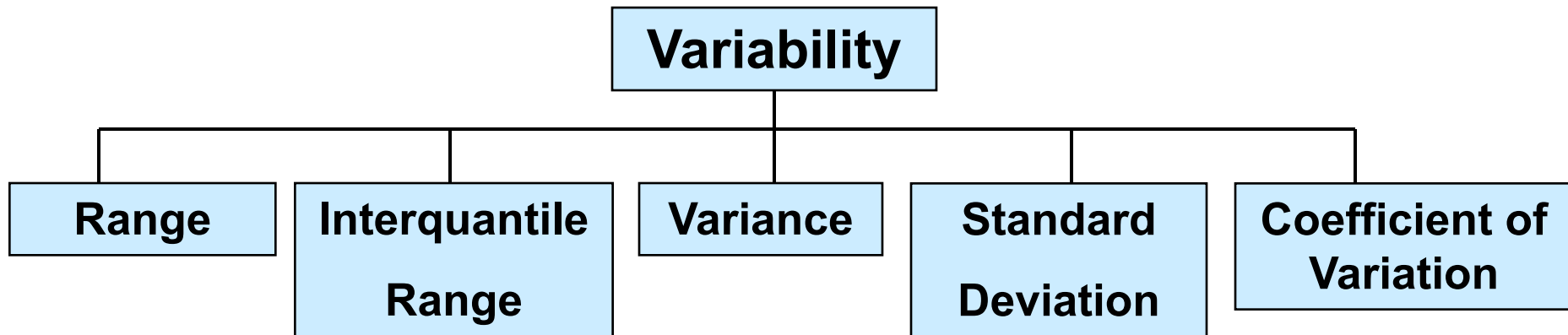
## *Measures of spread:*

- Range
- Interquantile Range
- Variance
- Standard Deviation
- Coefficient of Variation

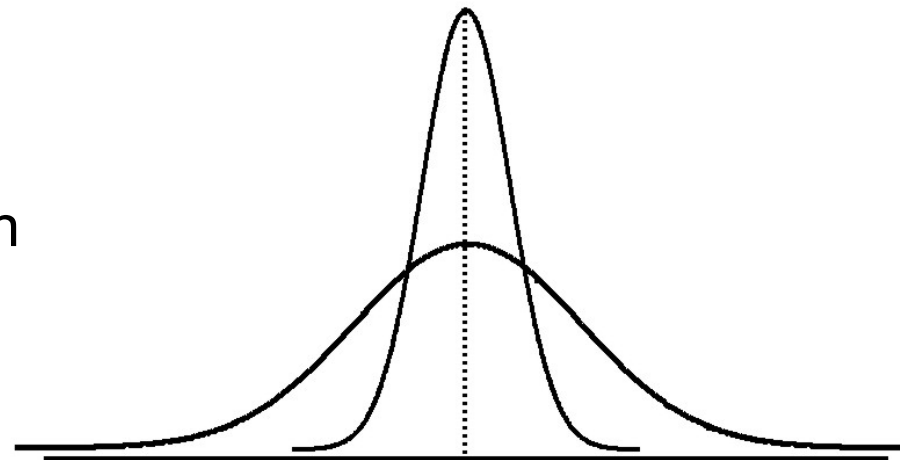
## *Measures of shape:*

- Skewness
- Kurtosis

# Measures of spread



- The measures of spread provide some information on the **variability** of the distribution values.



Identical center,  
Different variability

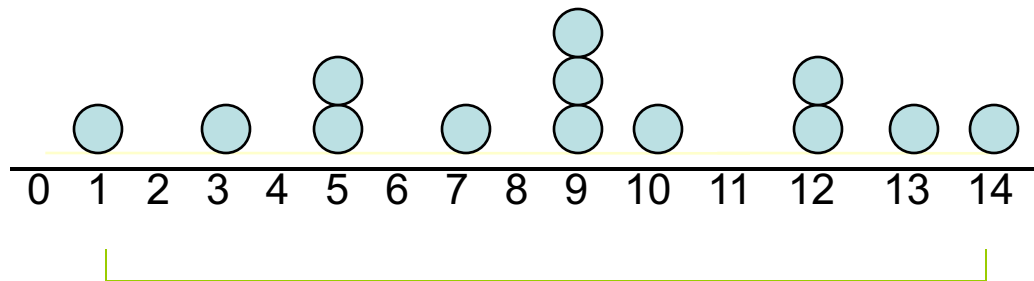


# Range

- The simplest measure of variability
- The difference between the maximum value and minimum value observed:

$$\text{Range} = X_{\max} - X_{\min}$$

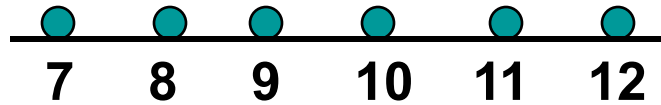
Example:



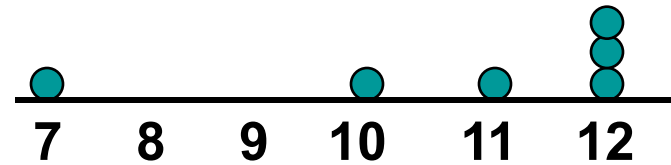
$$\text{Range} = 14 - 1 = 13$$

# Range

- Ignores the distribution of data



$$\text{Range} = 12 - 7 = 5$$



$$\text{Range} = 12 - 7 = 5$$

- Vastly affected by outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5

$$\text{Range} = 5 - 1 = 4$$

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

$$\text{Range} = 120 - 1 = 119$$

# Interquartile Range

- We can solve the problem concerning the outliers by using the interquartile range
- It does not take the highest and lowest values into consideration and computes the range of the middle 50% of the data distribution
- Interquartile Range = 3<sup>o</sup> quartile – 1<sup>o</sup> quartile

$$\text{IQR} = Q_3 - Q_1$$

# Variance

- Mean of squares related to differences between each observation and the mean

– Variance of Population:

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

whereas  $\mu$  = mean of population

$N$  = size of population

$x_i$  =  $i^{\text{th}}$  value of the variable  $X$

# Standard Deviation

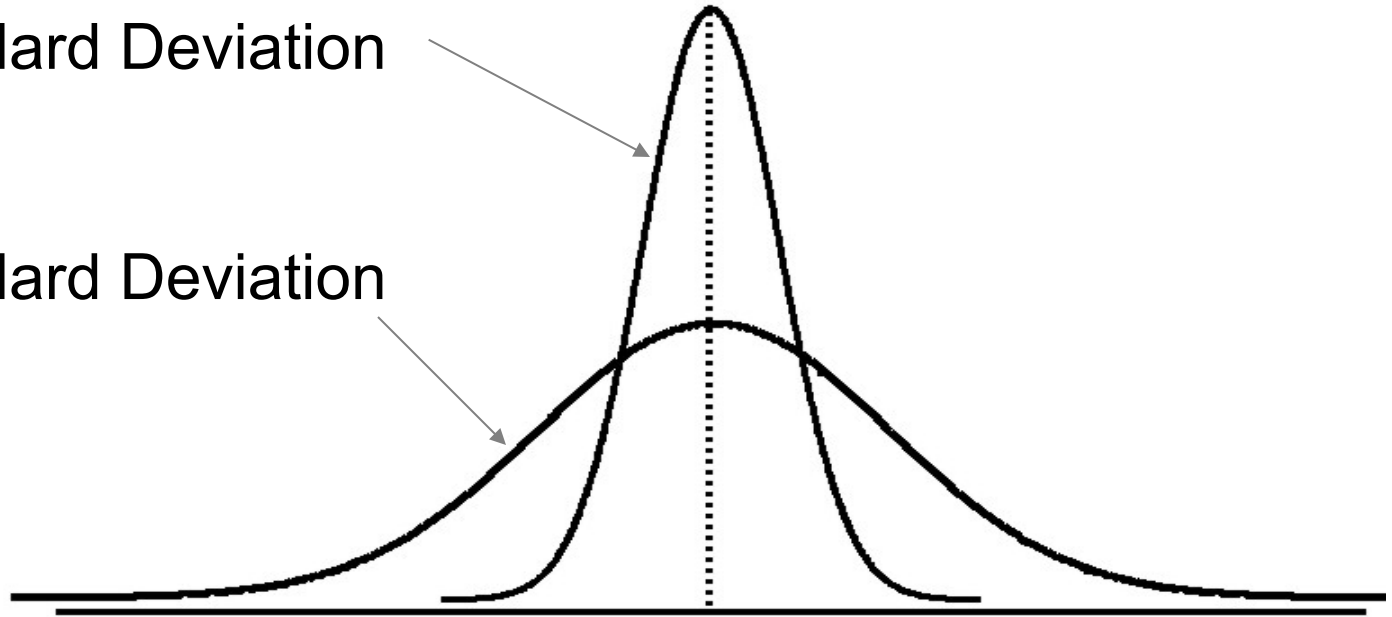
- Variability indicator commonly used
  - It shows the variability relative to the mean
  - It has the same unit of measurement as the original data
- Standard Deviation of population:

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$$

# Standard Deviation

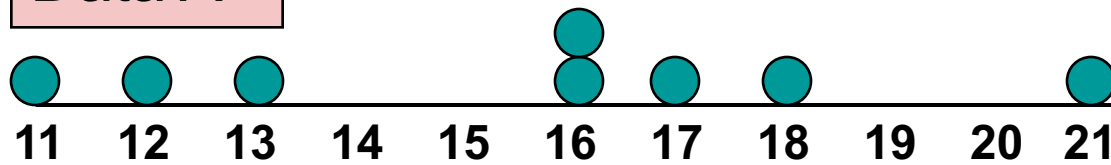
Minor Standard Deviation

Major Standard Deviation



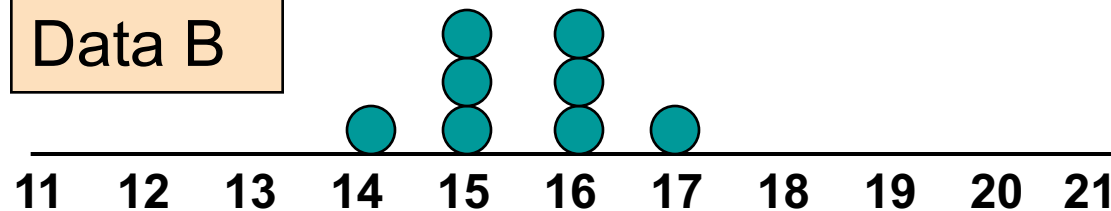
# Standard Deviation

Data A



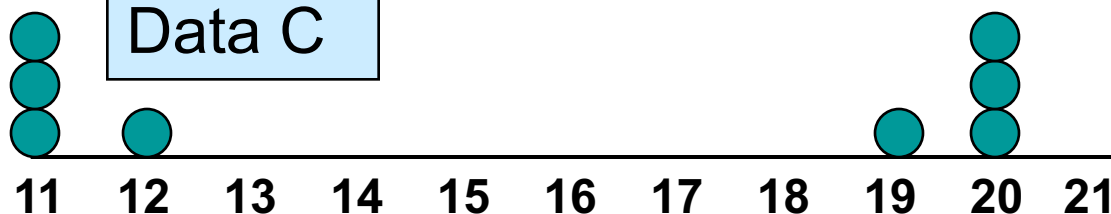
mean = 15.5  
 $s = 3.338$

Data B



mean = 15.5  
 $s = 0.926$

Data C



mean = 15.5  
 $s = 4.570$

# Standard Deviation

- It is computed by using the whole data set
- The more distant the values are from the mean, the more influence they have  
(this is because the square of the deviations from the mean is used)
- Identical remarks can be made with regards to the variance computation



# Coefficient of Variation

- It measures the relative variability
- It assumes values  $\geq 0$
- It is always measured in percentage figures
- It shows the relative variability around the mean
- It can be used to compare two or more data sets gauged with different units of measurement

$$CV = \left( \frac{s}{\bar{x}} \right) \cdot 100\%$$

# Coefficient of Variation

- Share A:
  - Last year average price = \$50
  - Standard Deviation = \$5

$$CV_A = \left( \frac{s}{\bar{x}} \right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Share B:
  - Last year average price = \$100
  - Standard Deviation = \$5

$$CV_B = \left( \frac{s}{\bar{x}} \right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

Both the shares have the same standard deviation, but share B shows less variability relative to its price

# Univariate statistics

## *Measures of location:*

- Mean
- Median
- Mode
- Quantiles
- Percentiles

## *Measures of spread:*

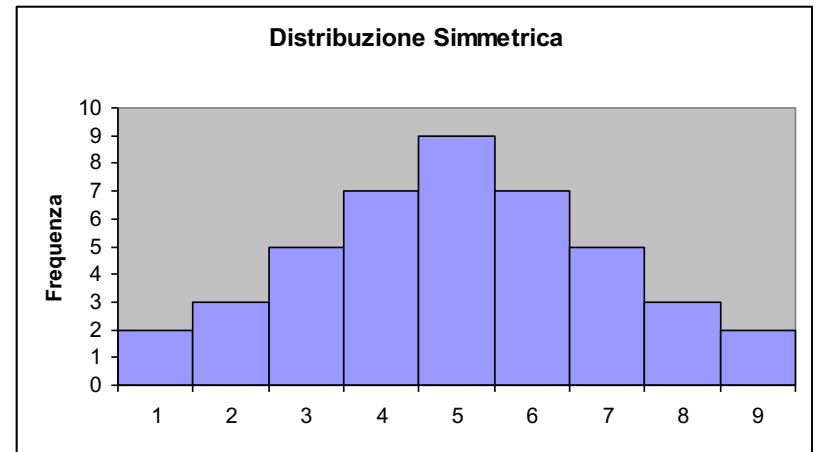
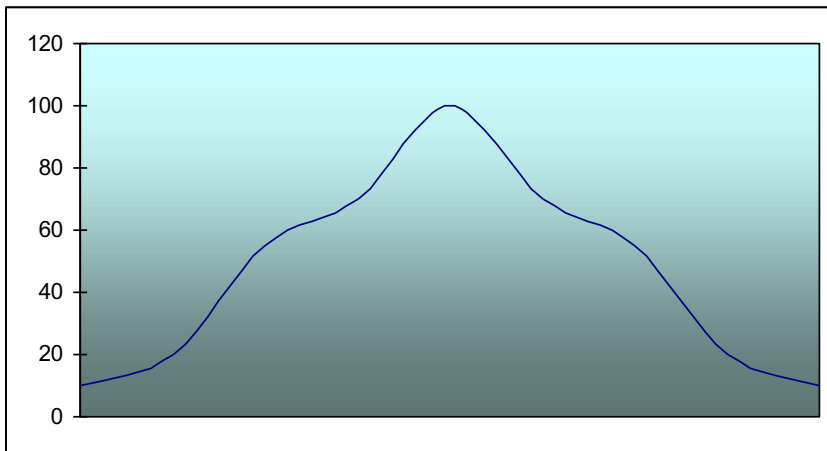
- Range
- Interquantile Range
- Variance
- Standard Deviation
- Coefficient of Variation

## *Measures of shape:*

- Skewness
- Kurtosis

# Form of Distribution

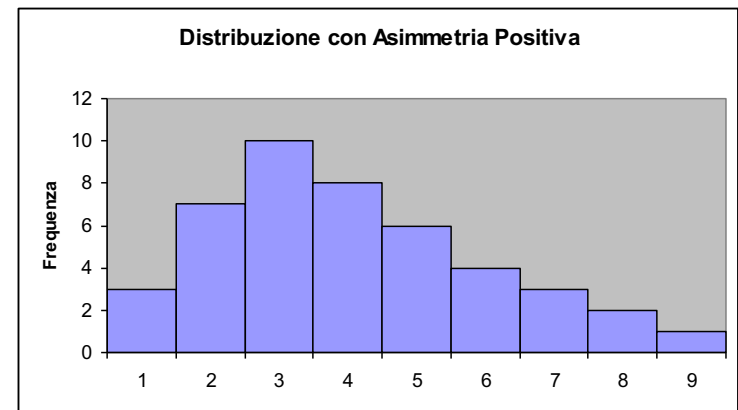
- The form of distribution is said to be symmetrical if the observations are all equally spread around the center.



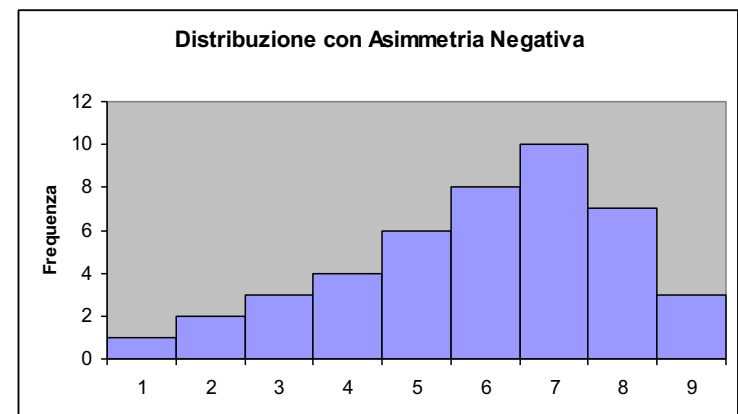
# Form of Distribution

- The form of distribution is said to be asymmetrical if the observations are not equally spread around the center

A distribution with **positive skewness** (**skewed to the right**) has a tail stretching to the right, towards positive values.



A distribution with **negative skewness** (**skewed to the left**) has a tail stretching to the left, towards negative values.



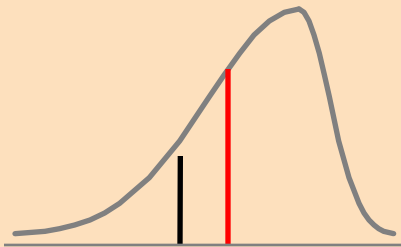
# Measures of the distribution shape

**Skewness:** it is an index which measures the degree of symmetry or skewness of a distribution:

- $\gamma=0$  symmetrical distribution: median=mean;
- $\gamma<0$  negative skewness: skewed to the left (mean<median);
- $\gamma>0$  positive skewness: skewed to the right (mean>median);

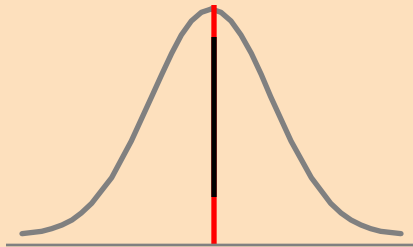
## Skewed to the left

Mean < Median



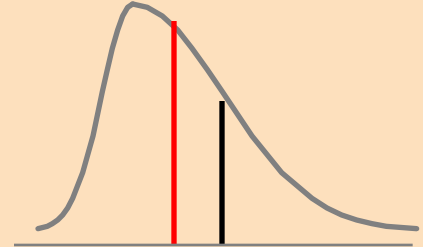
## Symmetrical

Mean = Median



## Skewed to the right

Median < Mean



# Measures of the distribution shape

*Kurtosis*: it is an index which enables us to check whether the data have a Normal distribution (symmetrical):

- $\beta=3$  if the distribution is “Normal”;
- $\beta<3$  if the distribution is ipo-normal (compared to the Normal distribution it has a lower density of frequency with regard to values very distant from the mean value);
- $\beta>3$  if the distribution is iper-normal (compared to the Normal distribution it has a higher density of frequency with regard to values very distant from the mean value).

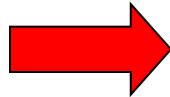
# Univariate descriptive statistics

In the univariate descriptive statistics we analyze one variable at a time.

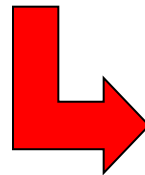
N_ID	D_8_2
H1	0.1
H2	0
H3	0
H4	0.2
H5	0.05
H6	0.2
H7	0.1
H8	0.1
H9	0.2
H10	0.05
H11	0
H12	0
H13	0
H14	0.15
H15	0
H16	0.1
H17	0
H18	0.2
H19	0
H20	0.05
H21	0.2
H22	0.2

...

H234	0.2
H235	0.1
H236	0.1



- Frequency distribution
- Synthesis measures
  - *Measures of location*
  - *Measures of spread*
  - *Measures of shape*



- Data Audit
  - Input errors
  - Missing values
- Basic insights



# Example: Nike vs Adidas

Data Matrix:

[...\Example\Final\\_Data\\_Nike\\_Adidas\\_V1.sav](#)

- Consider variables:
  - Q25: Education
  - Q6\_1-Q6\_7: Importance of characteristics
  - Q20: Money for sports items

# Univariate statistics



## Market Research “SHAMPOO”



# Univariate statistics

## Univariate Descriptive Statistics: Place of Purchase

80% of the interviewees purchase shampoo in supermarkets. This is mainly due to the nature of the product: shampoo is seen as a mass market product, and only customers that are interested in a more specialized product are willing to invest time in different places of purchase (such as pharmacy, hair dresser and perfumery).

Those results are probably affected by the fact that the research sample is composed by people under 35 years old: if the research would have taken into consideration an older sample probably the findings would be different.



Statistics

		D7A_ supermarket	D7B_ Pharmacy	D7C_ Herbalist	D7D_ Catalogue	D7E_ Hairdress	D7F_ Specialized	DTG_ Parfumery	D7H_Web	D7I_D2D	D7L_Market
N	Valid	172	172	172	172	172	172	172	172	172	172
	Missing	0	0	0	0	0	0	0	0	0	0
Mean		74,94	4,30	1,83	,17	5,53	2,88	3,85	,03	,32	,32
Median		100,00	,00	,00	,00	,00	,00	,00	,00	,00	,00
Mode		100	0	0	0	0	0	0	0	0	0
Variance		1208,769	255,943	56,129	2,893	344,648	117,254	181,423	,145	14,663	14,663
Minimum		0	0	0	0	0	0	0	0	0	0
Maximum		100	100	50	20	100	70	100	5	50	50

# Univariate statistics

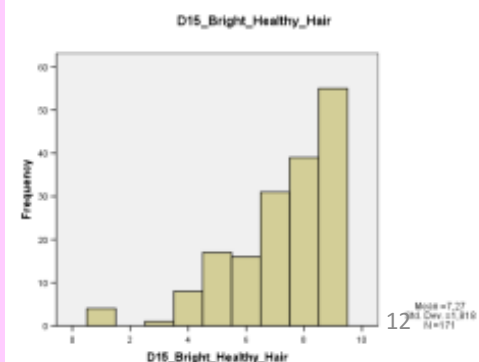
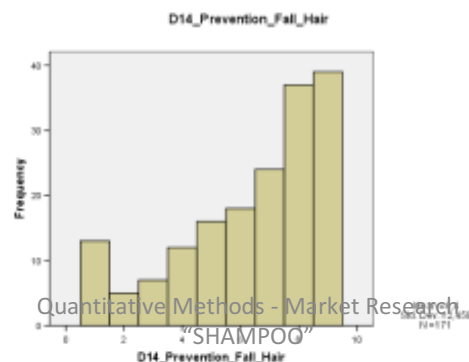
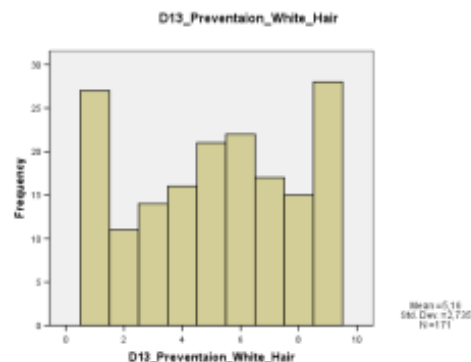
## Descriptive Statistics :

### Differences in the Relevance of the Main Characteristics of the New Product

The histograms show the frequency of the relevance given by the interviewees to the three main innovative features of the product we are planning to launch. The prevention of the fall of hair together with maintenance of an healthy and bright hair received the highest preference while the problem of white hair is distributed with an higher variety.

Statistics

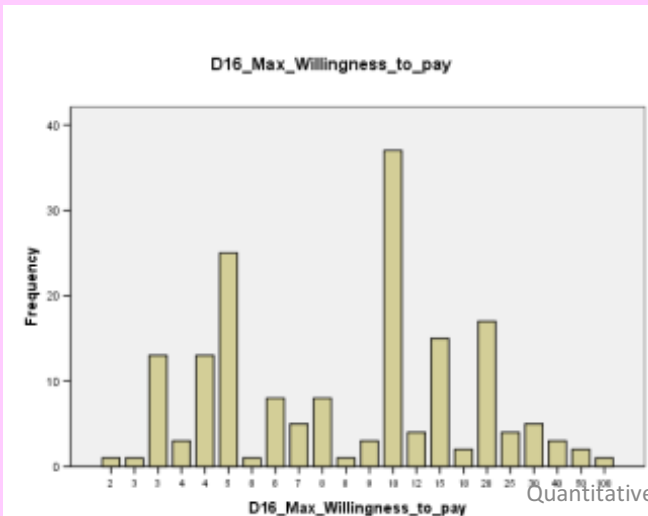
		D13_ Preventaion_ White_Hair	D14_ Prevention_ Fall_Hair	D15_Bright_ Healthy_Hair
N	Valid	171	171	171
	Missing	1	1	1
Mean		5,16	6,40	7,27
Median		5,00	7,00	8,00
Mode		9	9	9
Std. Deviation		2,735	2,458	1,818
Variance		7,479	6,042	3,306
Minimum		1	1	1
Maximum		9	9	9



# Univariate statistics

## Descriptive Statistics: Maximum Willingness to Pay

The price our sample is willing to pay for such an innovative product is on average 11,68€, and this value is not that biased because both the median and mode are close to this price (10€).



Statistics		
D16 Max Willingness to pay		
N	Valid	172
	Missing	0
Mean		11,68
Median		10,00
Mode		10
Std. Deviation		11,039
Variance		121,869
Minimum		2
Maximum		100

# Univariate statistics

## Univariate Descriptive Statistics: Level of Involvement

To see how high is the level of involvement in the product, we calculate the mean and the standard deviation of the level of agreement interviewees gave to these 5 statements:

	N	Minimum	Maximum	Mean	Std. Deviation
D19a_Attention_in_choose	172	1	9	6,83	1,761
D19b_All_brands_are_similar	172	1	9	4,82	2,140
D19c_pick_the_first	172	1	9	2,62	1,942
D19d_look_for_my_favorite	172	1	9	6,69	2,065
D19e_try_different_shampoos	172	1	9	5,02	2,263
Valid N (listwise)	172				

As it emerges from the chart the interviewees give great attention to the choice of the shampoo, tend to prefer their favorite one but they often try new brands.

# Univariate statistics

## Univariate Descriptive Statistics: Subset Remembered

The majority of the interviewees showed particular interest to mass market brands (such as Pantene, Garnier, L'Oreal and Sunsilk). This is probably mainly due to the fact that these companies highly invest in advertising through mass market channels.

Sunsilk position is also really interesting: this Spanish company is present in Italy since a few years but became well known thanks to an aggressive advertising campaign.

	<i>First Brand Remembered</i>	<i>Second Brand Remembered</i>	<i>Third Brand Remembered</i>
<i>Pantene</i>	<b>26,20%</b>	17,40%	9,90%
<i>Garnier</i>	16,90%	<b>23,30%</b>	<b>15,10%</b>
<i>L'Oreal</i>	14%	14%	5,80%
<i>Sunsilk</i>	7%	3,50%	11,60%

## TOTAL AMOUNT OF MONEY ON AN INDIVIDUAL BASIS

### Basic Statistical Measures

#### Location

**Mean**

106.1410

**Median**

103.2900

**Mode**

0.0000

#### Variability

**Std Deviation**

81.01306

**Variance**

6563

**Range**

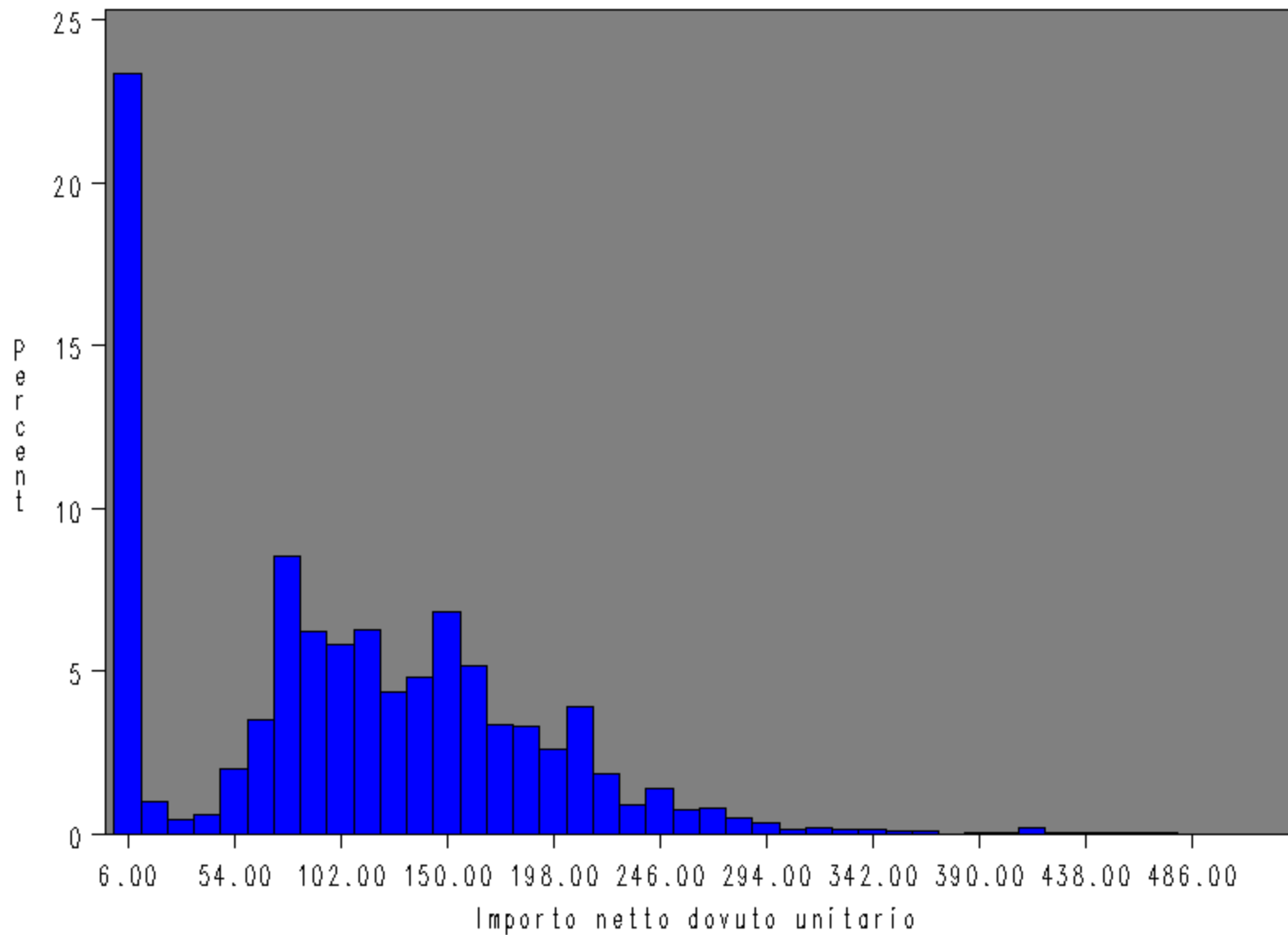
523.69000

**Interquartile Range**

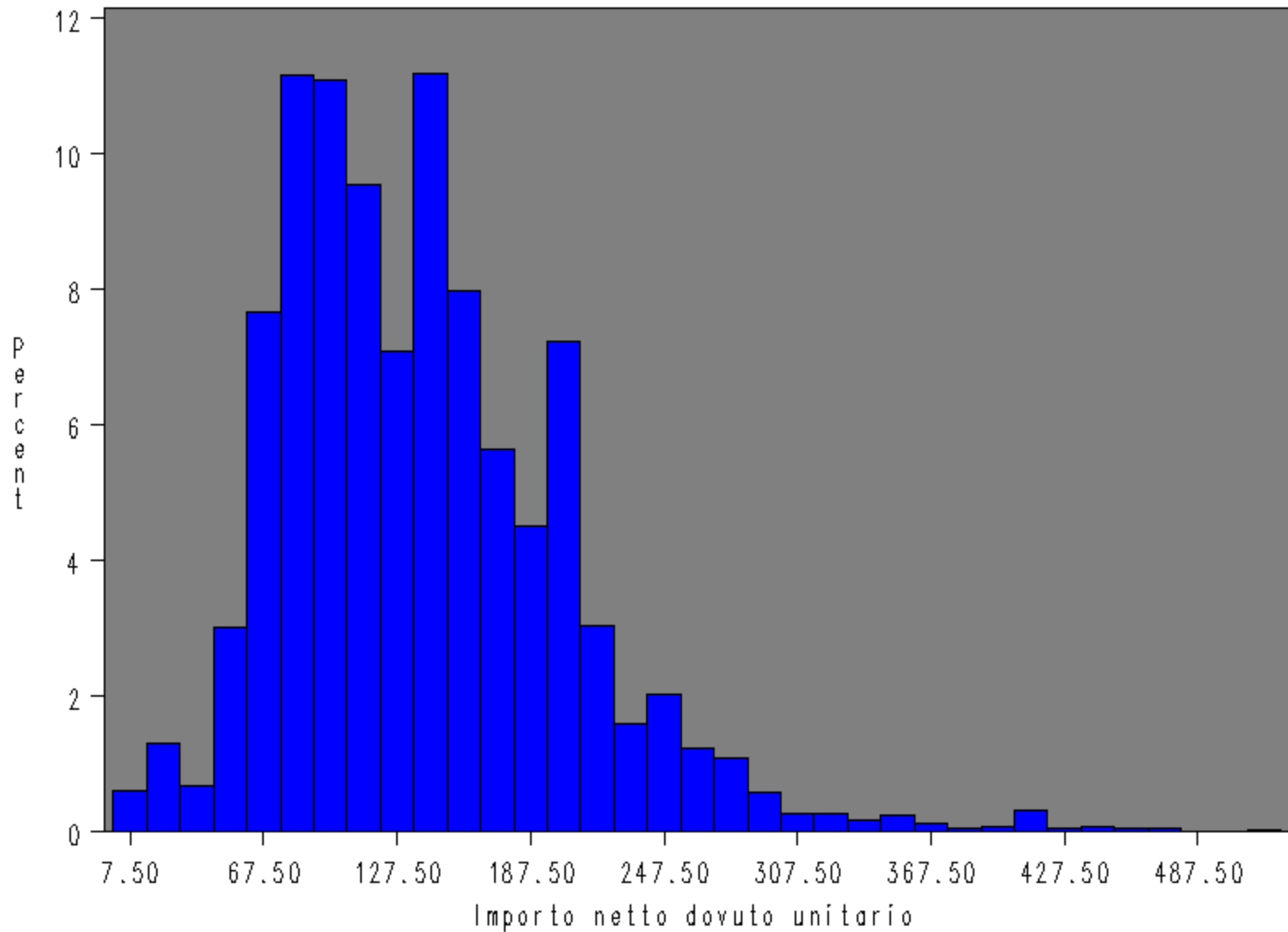
118.62500



## TOTAL AMOUNT OF MONEY ON AN INDIVIDUAL BASIS



## TOTAL AMOUNT OF MONEY ON AN INDIVIDUAL BASIS



## TOTAL AMOUNT OF MONEY ON AN INDIVIDUAL BASIS

### Basic Statistical Measures

#### Location

**Mean**      138.0247

**Median**      129.1100

**Mode**      149.0000

#### Variability

**Std Deviation**      64.29397

**Variance**      4134

**Range**      521.77000

**Interquartile Range**      82.62000

# Univariate Analysis

	Distribution	Mode	Percentiles	Moments	Shape
Nominal	X	X			
Ordinal	X	X	X		
Quantitative	X	X	X	X	X

# Keywords

- Frequency distribution
  - Absolute distribution
  - Relative distribution
  - Cumulative distribution
- Pie/Bar Chart
- Histogram
- Measures of location
  - Mean
  - Median
  - Mode
  - Quantiles
  - Percentiles
- Box Plot
- Missing values
- Outliers
- *Connect results/outputs to business case*
- *Connect results/outputs to sample*
- *Interpret & comment results/outputs*

# Text Book

Naresh K. Malhotra, *“Marketing Research – An Applied Orientation”*,  
Pearson – Prentice Hall, Ed. 2010

- Chapter 15 – pag 480 - 488

## Related Bibliography

- *E. Jane Miller*, *The Chicago Guide to Writing about Numbers*, *The University of Chicago press*, 2004

# Agenda

## Module 3: Univariate & Bivariate Analysis

- Introduction
- Univariate Analysis
- Bivariate Analysis

# Descriptive bivariate statistics

It is used to describe the relationship between two variables jointly. It differs from the typology of the variables analysed.

- qualitative/quantitative discrete variables: contingency tables (two-way table)
- quantitative variables: linear correlation analysis
- one qualitative variable and one quantitative variable: comparison between the means



# Contingency tables

The Contingency Table is a two way (or cross) table and the values indicated inside the chart represent the absolute joint frequencies, their total sum is equal to the overall number of the units observed.

If considering a contingency table, like the one below, it's possible to achieve a great amount of information: **Marginal distributions**, by means of adding joint frequencies by row and by column; the joint relative frequencies, by carrying out the ratio between the absolute joint frequencies and the overall number of the units observed.

Gender \* Go to French Bar Crosstabulation

			Go to French Bar		Total
			No	Yes	
Gender	Male	Count	7	32	39
		% within Gender	17.9%	82.1%	100.0%
		% within Go to French Bar	46.7%	42.1%	42.9%
		% of Total	7.7%	35.2%	42.9%
	Female	Count	8	44	52
		% within Gender	15.4%	84.6%	100.0%
		% within Go to French Bar	53.3%	57.9%	57.1%
		% of Total	8.8%	48.4%	57.1%
Total	Count	15	76	91	
	% within Gender	16.5%	83.5%	100.0%	
	% within Go to French Bar	100.0%	100.0%	100.0%	
	% of Total	16.5%	83.5%	100.0%	

# Contingency tables

It is possible to achieve further unidimensional distributions:

- *Conditioned frequencies*: are the frequencies of the character X conditioned to the character Y and vice-versa

$$P_{y|x}(x_i, y_j) = P(x_i, y_j) / P_x(x_i)$$

$$P_{x|y}(x_i, y_j) = P(x_i, y_j) / P_y(y_j)$$

*Statistical independence*: if X changes and the conditioned distributions  $(Y|X)=x_i$  are all equal we can come to the conclusion that the distribution of character Y is not affected by X. In case of statistical independence the relative joint frequency is equal to the multiplication (cross product) of the marginal frequencies

$$P(x_i, y_j) = P_x(x_i)P_y(y_j)$$

The statistical independence is obviously a symmetrical concept. If it is true for X it is also true for Y. In case of statistical independence the bivariate analysis of X (Y) does not provide additional information regarding the univariate analysis.

# Contingency tables

- *Perfect unilateral dependence* each value of X corresponds with only one value of Y, although the reverse situation may not happen. Generally speaking, when the number of columns (Y) is lower than the number of rows (X) the variable X can never be entirely affected by variable Y.
- *Perfect bilateral dependence* each X value corresponds with only one Y value and viceversa; the perfect bilateral dependence might only be obtained in case of square matrixes.

# Indexes of connection

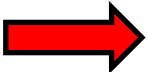
In the real analyses statistical independence is rather unlikely to occur. Therefore it is required to have indexes to measure the degree of connection between the variables.

–  $\chi^2$  (chi-square) has value “0” when X and Y are independent. It is affected by:

- number of observations available: the higher the number of observations (N), the higher the index
- the dimensions of the data matrix: the higher the number of rows/ columns , the higher the index

$$\chi^2 = N \sum \sum [P(x_i, y_j) - P_x(x_i) P_y(y_j)]^2 / P_x(x_i) P_y(y_j)$$

Chi-Square Tests



	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.106 <sup>a</sup>	1	.744	.781	.480
Continuity Correction <sup>b</sup>	.002	1	.967		
Likelihood Ratio	.106	1	.745		
Fisher's Exact Test					
Linear-by-Linear Association	.105	1	.746		
N of Valid Cases	91				



a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 6.43.

b. Computed only for a 2x2 table

# Indexes of connection

- Cramer's V Index, based on Chi-Square statistic, is a more effective index in that it is relative and it's not so affected by the number of observations or the dimension of the contingency table. It's range is between 0 and 1:
  - It assumes 0 in case of statistical independence,
  - It assumes 1 in case of perfect dependence, at least on a unilateral. It grows according to the degree of connection between the two variables.

**Symmetric Measures**

		Value	Approx. Sig.
Nominal by Nominal	Phi	.034	.744
	 Cramer's V	.034	.744
N of Valid Cases		91	

# Contingency tables

facebook

In order to analyze the future potentials of monetization for the social network we wanted to analyze the current reactions of actual users towards two key aspects in this sense such as advertising and privacy concerns.

There seems to be differences towards this two sensible issues both in terms of gender as well in terms of geographical origin

# Contingency tables

facebook

## Privacy Read vs Gender

Crosstab

			Privacy_read		Total
			No	Yes	
Gender	Female	Count	82	26	108
		% within Gender	75,9%	24,1%	100,0%
		% of Total	42,5%	13,5%	56,0%
	Male	Count	53	32	85
		% within Gender	62,4%	37,6%	100,0%
		% of Total	27,5%	16,6%	44,0%
Total	Count		135	58	193
	% within Gender		69,9%	30,1%	100,0%
	% of Total		69,9%	30,1%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4,169 <sup>b</sup>	1	,041		
Continuity Correction <sup>a</sup>	3,548	1	,060		
Likelihood Ratio	4,153	1	,042		
Fisher's Exact Test				,057	,030
Linear-by-Linear Association	4,147	1	,042		
N of Valid Cases	193				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 25,54.

The V Cramer coefficient is Significant ; its value (0,147) signals some statistical dependence between the considered variables



We can then assume that there is a difference in terms of behavior between genders.

**Males are more privacy concerned**

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,147	,041
	Cramer's V	,147	,041
N of Valid Cases		193	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

# Contingency tables

facebook

## Privacy Share vs Region

With regards to privacy, both the Phi squared and the Cramer's index are acceptable and the test remains significant:

- We can therefore assume that there are different degree of concerns with privacy both within Italy as outside
- In particular **users from the north of Italy seems to be the most privacy conscious ones.** Users from the south and foreigners seems to be far less privacy conscious.

Crosstab

			Privacy Share		Total
			No	Yes	
Region	North Italy	Count	34	57	91
		% within Region	37,4%	62,6%	100,0%
		% of Total	17,6%	29,5%	47,2%
	Center Italy	Count	10	11	21
		% within Region	47,6%	52,4%	100,0%
		% of Total	5,2%	5,7%	10,9%
	South Italy	Count	21	9	30
		% within Region	70,0%	30,0%	100,0%
		% of Total	10,9%	4,7%	15,5%
	Outside Italy	Count	33	18	51
		% within Region	64,7%	35,3%	100,0%
		% of Total	17,1%	9,3%	26,4%
Total	Count	98	95	193	
	% within Region	50,8%	49,2%	100,0%	
	% of Total	50,8%	49,2%	100,0%	

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	15,030 <sup>a</sup>	3	,002
Likelihood Ratio	15,292	3	,002
Linear-by-Linear Association	13,033	1	,000
N of Valid Cases	193		

a. 0 cells (,0%) have expected count less than 5. The

Symmetric Measures

	Value	Approx. Sig.
Nominal by Nominal	Phi	,279
	Cramer's V	,279
N of Valid Cases	193	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.



# Example: Nike vs Adidas

Data Matrix:

[...\Example\Final\\_Data\\_Nike\\_Adidas\\_V1.sav](#)

- Consider variables:
  - Q2 x (Q22, Q24)

# Descriptive bivariate statistics

It is used to describe the relationship between two variables jointly. It differs from the typology of the variables analysed.

- qualitative/quantitative discrete variables: contingency tables (two-way table)
- **quantitative variables**: linear correlation analysis
- one qualitative variable and one quantitative variable: comparison between the means

# Linear Correlation

The connection measurements can be applied to qualitative variables. If you wish to measure the degree of concordance between two quantitative variables, you need to use other indexes:

- *Covariance*  **$\text{Cov}(X,Y)$**  is an index which may have positive values if there is a concordance between X and Y (with high values of the former we have high values of the latter); it may have negative values when a discordance occurs (with high values of the former we don't have high values of the latter).

In case of statistical independence it assumes “0” value. It is an absolute index, which means it shows the presence and the direction of a link between two variables but nothing can be stated on the nature of their link.

$$\text{Cov}(X,Y) = \sum \sum (x_i - \mu_x) (y_j - \mu_y) p(x_i, y_j)$$

# Linear Correlation

- Covariance between two variables:

$\text{Cov}(x,y) > 0 \rightarrow$  x and y tend to go in the same direction

$\text{Cov}(x,y) < 0 \rightarrow$  x and y tend to go in opposite directions

$\text{Cov}(x,y) = 0 \rightarrow$  x and y have no linear connection

- The covariance is only related to the strength of the connection but does not involve a causal effect.

# Linear Correlation

- The *Coefficient of linear correlation*  $\rho(X,Y)$  is a relative index which overcomes the problem of the previous index.

$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

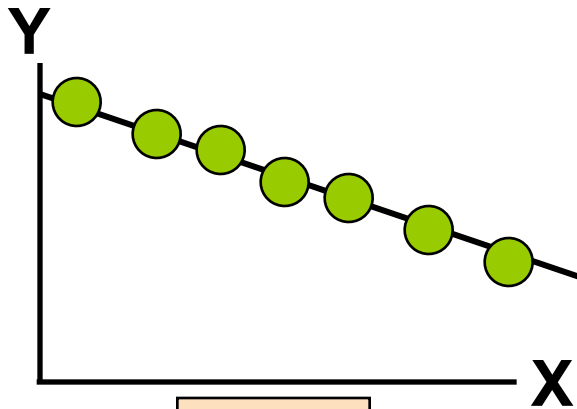
$\sigma_{X/Y}$  = Standard Deviation of X/Y

- It may have values between:  $[-1, 1]$ .
- It is equal to -1 or to 1 if and only if there a perfect linear relation between X and Y and vice versa.
- It is equal to 0 when there is no type of linear relation between X and Y.

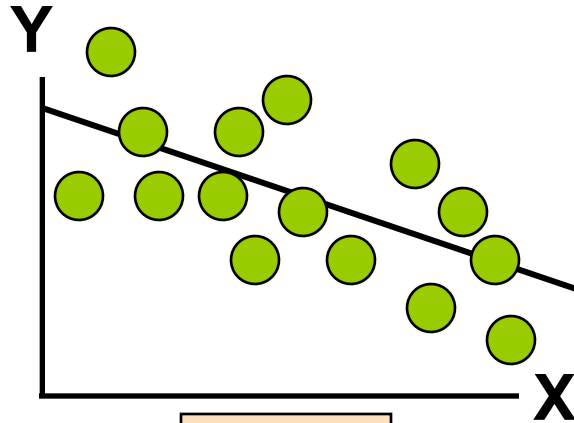
# Linear Correlation

- *Coefficient of linear correlation  $\rho(X, Y)$  :*
  - $\rho = 0 \Rightarrow$  there is no linear relation between  $X$  and  $Y$
  - $\rho > 0 \Rightarrow$  there is a positive linear relation between  $X$  and  $Y$ 
    - » When  $X$  has high values it means  $Y$  may also have high values
    - »  $\rho = +1 \Rightarrow$  there is a perfectly positive linear dependence
  - $\rho < 0 \Rightarrow$  there is a negative linear connection between  $X$  and  $Y$ 
    - » When  $X$  has high values it means  $Y$  may also have low values
    - »  $\rho = -1 \Rightarrow$  there is a perfectly negative linear dependence

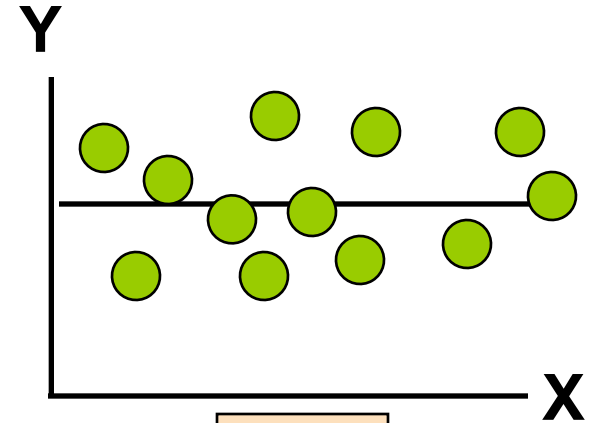
# Linear Correlation



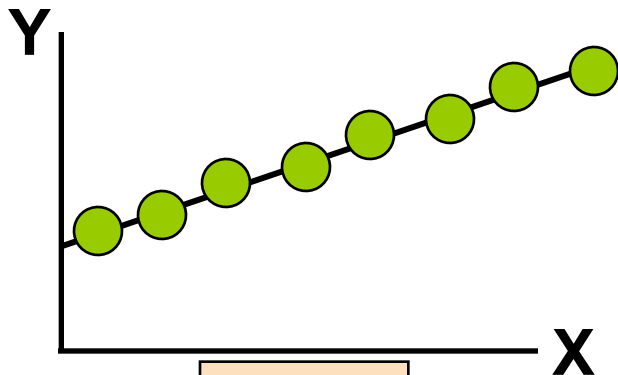
$$r = -1$$



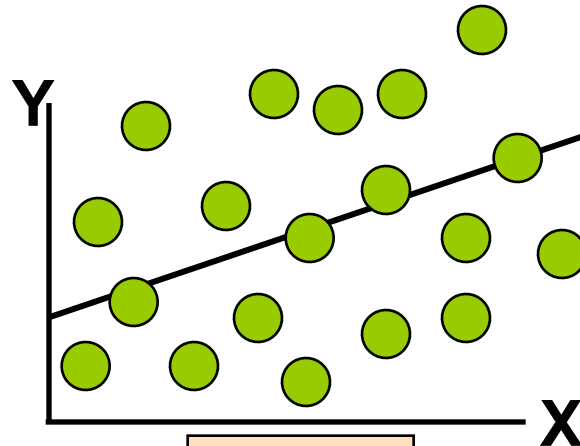
$$r = -.6$$



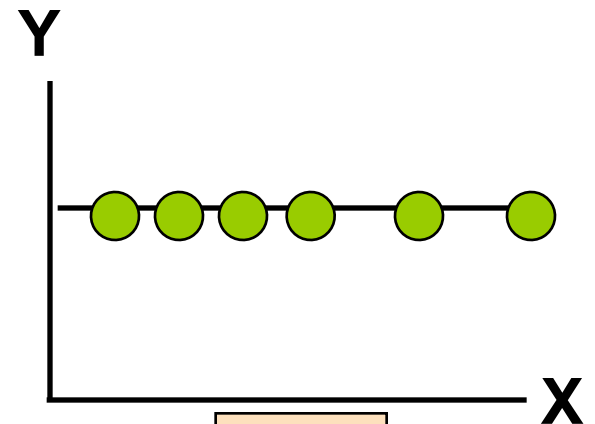
$$r = 0$$



$$r = +1$$



$$r = +.3$$



$$r = 0$$

# Linear Correlation

## Correlations

		Quality of the ingredients	Genuineness	Lightness	Taste
Quality of the ingredients	Pearson Correlation	1	.629**	.299**	.232**
	Sig. (2-tailed)		.000	.000	.001
	N	220	220	218	220
Genuineness	Pearson Correlation	.629**	1	.468**	.090
	Sig. (2-tailed)	.000		.000	.181
	N	220	220	218	220
Lightness	Pearson Correlation	.299**	.468**	1	.030
	Sig. (2-tailed)	.000	.000		.657
	N	218	218	219	219
Taste	Pearson Correlation	.232**	.090	.030	1
	Sig. (2-tailed)	.001	.181	.657	
	N	220	220	219	221

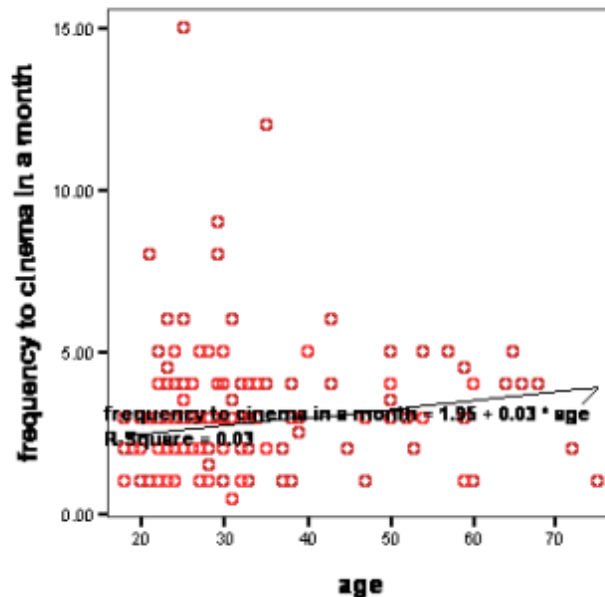
\*\*. Correlation is significant at the 0.01 level (2-tailed).



# Linear Correlation

## 2.2 Data elaboration through SPSS Univariate analysis

Correlation between AGE and FREQUENCY TO CINEMA IN A MONTH



The Pearson Correlation is positive and the P-Value is lower than 0.05: the test is significant and there is a positive correlation

The sample presents a positive linear correlation between the age and the frequency they go to cinema in a month (the Pearson correlation is positive). This is a quite weak correlation and there are some outliers (as it can be seen from the scatterplot).

Anyway this result is significant for our objectives: the older the customers are, the more often they go to cinema, while the youngs tend to be rare users.

Correlations

		frequency to cinema in a month	age
frequency to cinema in a month	Pearson Correlation	1	.175*
	Sig. (2-tailed)		.015
	N	194	193
age	Pearson Correlation	.175*	1
	Sig. (2-tailed)	.015	
	N	193	199

\*. Correlation is significant at the 0.05 level (2-tailed).



# Linear Correlation

## Correlation between AGE and how many movies are seen in a certain kind of cinema

Correlations

		age	A multisala cinema
age	Pearson Correlation	1	-.408**
	Sig. (2-tailed)		.000
	N	199	192
A multisala cinema	Pearson Correlation	-.408**	1
	Sig. (2-tailed)	.000	
	N	192	192

\*\* . Correlation is significant at the 0.01 level (2-tailed).

The age is **indirectly correlated** to the percentage of movies seen in a multiplex cinema: the younger tend to prefer this kind of cinema that is quite different from Apollo in terms of offering and business model. This is the basis to understand the needs of the customers in this specific demographic segment

Correlations

		age	Another essay cinema
age	Pearson Correlation	1	.347**
	Sig. (2-tailed)		.000
	N	199	189
Another essay cinema	Pearson Correlation	.347**	1
	Sig. (2-tailed)	.000	
	N	189	189

\*\* . Correlation is significant at the 0.01 level (2-tailed).

On the other side, the percentage of movie seen in an essay cinema is **directly correlated** to the age. The eldest respondents are the ones who frequent more essay cinemas. Even in this case the result is relevant for specific managerial implications.



# Example: Nike vs Adidas

Data Matrix:

[...\Example\Final\\_Data\\_Nike\\_Adidas\\_V1.sav](#)

- Consider variables:
  - Q5\_1, Q5\_2, Q20

# Descriptive bivariate statistics

It is used to describe the relationship between two variables jointly. It differs from the typology of the variables analysed.

- qualitative/quantitative discrete variables: contingency tables (two-way table)
- quantitative variables: linear correlation analysis
- one qualitative variable and one quantitative variable: comparison between the means

# Comparison between the means

If you want to cross over a quantitative variable with a qualitative variable the relation can be described by comparing the means of the numerical variable within the categories defined by the variable measured on a nominal/ordinal basis.

<i>Swiftiness Typology of Customers</i>	<i>Mean</i>	<i>N</i>
<i>Individual persons</i>	7.8403	357
<i>Companies/Busi nesses</i>	8.5132	76
<i>Total Amount</i>	7.9584	433

# Comparison between the means

## Report

Produzione artigianale

Età	Mean	N	Std. Deviation
18-25	5.01	78	2.224
26-35	5.53	55	2.609
36-50	6.00	41	2.098
Over 50	6.09	47	2.320
Total	5.55	221	2.352

## Report

Attenzione a bisogni specifici

Età	Mean	N	Std. Deviation
18-25	4.05	78	2.772
26-35	4.53	53	2.791
36-50	5.00	41	2.837
Over 50	5.83	47	8.168
Total	4.73	219	4.536

# Comparison between the means

A synthetic index of the intensity of the relation is based on the **decomposition of the variance** of the quantitative variable Y whose dependence on the categorical variable X is examined. The total variability of Y is:

$$SS_T = SS_{\text{Between}} + SS_{\text{In}}$$

$SS_T$ : Total Sum of Squares, describes the overall variability of Y,

$SS_{\text{Between}}$ : Sum of the Squares between the groups, describes how much variability of Y may be linked to the degree of variation of the X categories,

$SS_{\text{In}}$ : Sum of the Squares in the groups, describes the variability of Y which is not influenced by X.

# Comparison between the means

## Report

### Handicraft

Age	Mean	N	Std. Deviation
18-25	5.01	78	2.224
26-35	5.53	55	2.609
36-50	6.00	41	2.098
Over 50	6.09	47	2.320
Total	5.55	221	2.352

### Measures of Association

	Eta	Eta Squared
Handicraft * Age	.191	.036



# Comparison between the means

## *Compared means*

*CD one-song price \* download*

There is significance. People that don't download are willing to pay more (mean 6 euro) than people who usually download. Our explanation is that the latter ones are used to download music free with peer-to-peer programs: they are willing to pay less in order to buy the physical CD, then. However, the importance of CDs as collectible object explain why even this category of person has a desired price who is not so far from the other group. The same behavior was observed also within the means between download and price of an entire album.



# Comparison between the means

prezzo\_CD\_single

download	Mean	N	Std. Deviation
no	6,5769	26	4,91669
yes	4,0557	174	2,72502
Total	4,3835	200	3,19608

There is significance, even if Eta Squared is low.

ANOVA Table

		Sum of Squares	df	Mean Square	F	Sig.
prezzo_CD_single	Between Groups (Combined)	143,780	1	143,780	15,071	,000
* download	Within Groups	1888,995	198	9,540		
	Total	2032,776	199			

Measures of Association

	Eta	Eta Squared
prezzo_CD_single * download	,266	,071

We examined also other variables, but we didn't find any interesting value in significance and eta squared.



# Example: Nike vs Adidas

Data Matrix:

[...\Example\Final\\_Data\\_Nike\\_Adidas\\_V1.sav](#)

- Consider variables:
  - Q2 x Q20

# Tests for the analysis of variables relation

- We can have two typologies of Errors:

Possible outcome		
	State of the matter	
Decision	$H_0$ True	$H_0$ False
Non Rejection $H_0$	No error	Second Type Error
Rejection $H_0$	First Type Error	No Error

# Tests for the analysis of variables relation

- **First Type Error**
  - To reject the True Null Hypothesis
  - This type of error is regarded as very serious error

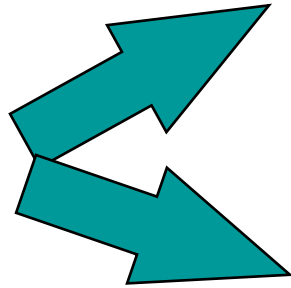
The probability of the first type error is  $\alpha$

- $\alpha$  is called the level of test significance
- It is set by the analyst in advance

# How to read a statistical test (1)

Example:

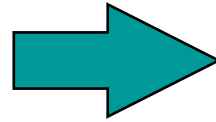
1) Hypothesis



$H_0$ : X & Y are independent

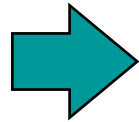
$H_1$ : X & Y are not independent

2) Test Statistics



Chi-Square Statistics

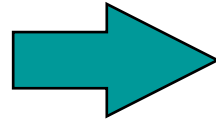
3) p-value



It represents the probability to make the first type error according to the value observed in the test statistics

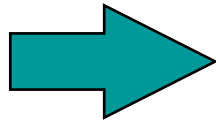
# How to read a statistical test (2)

If p-value is small



REJECTION  $H_0$

Otherwise



ACCEPTANCE  $H_0$

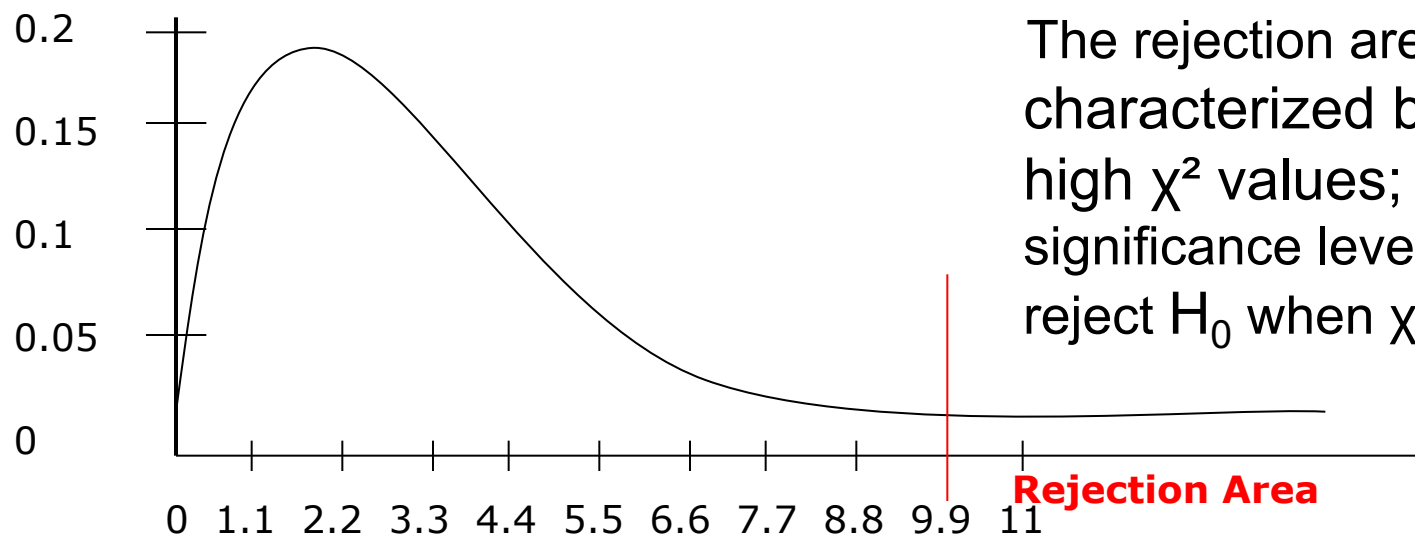
# $\chi^2$ Test for statistical independence

We regard the distribution of  $\chi^2$  (Chi-Square) with a number of degrees of freedom equal to  $(k-1)(h-1)$ , whereas  $k$  is the number of rows and  $h$  is the number of columns in the contingency table.

The following:

- $H_0$  : statistical independence between  $X$  and  $Y$
- $H_1$  : statistical dependence between  $X$  and  $Y$

The rejection area is situated in the right tail of the distribution.



The rejection area is characterized by relatively high  $\chi^2$  values; if the significance level is at 5%, we reject  $H_0$  when  $\chi^2 > \chi^2_{0.95}$



# $\chi^2$ Test for statistical independence

## Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	5.471 <sup>a</sup>	3	.140
Likelihood Ratio	5.402	3	.145
N of Valid Cases	221		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 15.95.

## Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	26.304 <sup>a</sup>	8	.001
Likelihood Ratio	28.928	8	.000
N of Valid Cases	221		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 5.47.

# t test for linear independence

The t test is designed to assess the hypothesis of linear independence between two variables, by using the linear correlation index  $\rho$ .

The following:

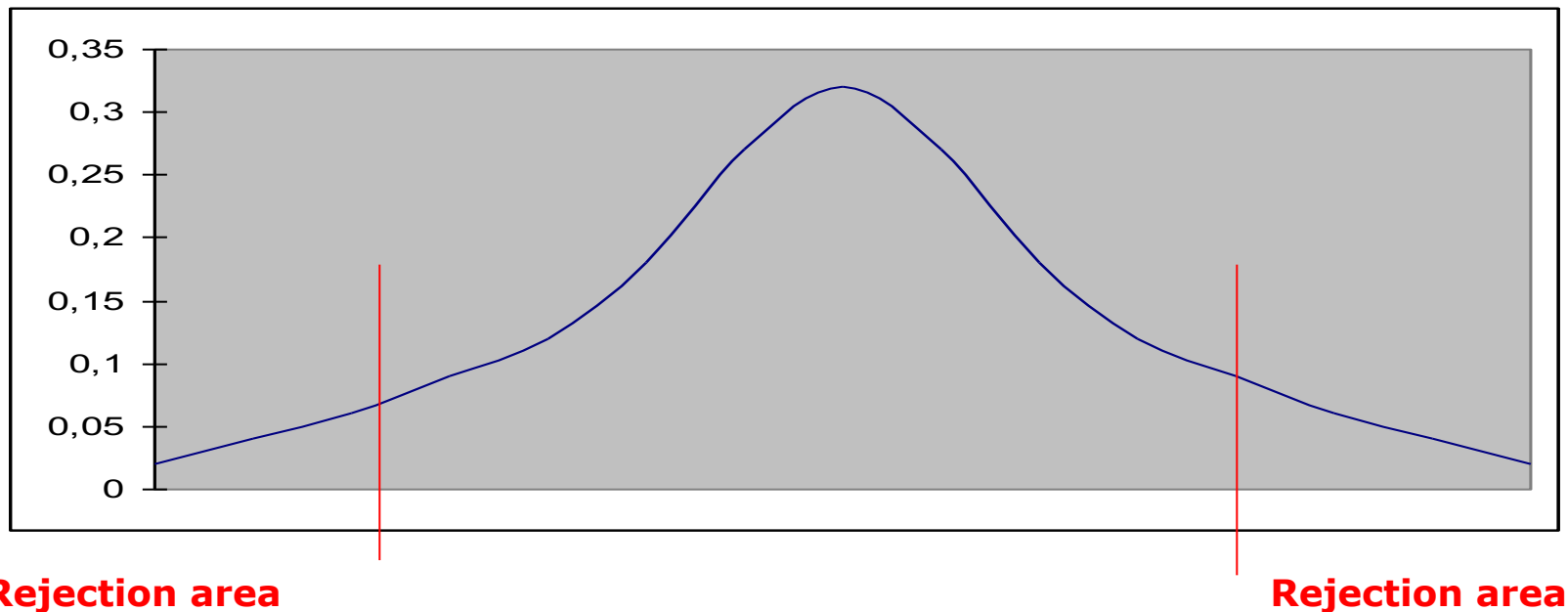
- $H_0$ : linear independence between X and Y ( $\rho_{\text{popolaz}} = 0$ )
- $H_1$ : linear dependence between X and Y ( $\rho_{\text{popolaz}} \neq 0$ )

The test statistic has the distribution of the t-Student with a number of degrees of freedom equal to  $n-2$ . The more the sample grows the higher the test statistic is.

$$t = \rho \sqrt{(n-2) / (1 - \rho^2)}$$

# t test for linear independence

The rejection area is characterized by relatively high absolute t values; if the significance level is at 5%, we reject  $H_0$  when  $|t| > t_{0,975}$



# t test for linear independence

**Correlations**

		Quality of ingredients	Genuineness	Lightness	Taste
Quality of ingredients	Pearson Correlation	1	.629**	.299**	.232**
	Sig. (2-tailed)		.000	.000	.001
	N	220	220	218	220
Genuineness	Pearson Correlation	.629**	1	.468**	.090
	Sig. (2-tailed)	.000		.000	.181
	N	220	220	218	220
Lightness	Pearson Correlation	.299**	.468**	1	.030
	Sig. (2-tailed)	.000	.000		.657
	N	218	218	219	219
Taste	Pearson Correlation	.232**	.090	.030	1
	Sig. (2-tailed)	.001	.181	.657	
	N	220	220	219	221

\*\* . Correlation is significant at the 0.01 level (2-tailed).

# F test to assess the hypotheses on the difference between means

We take the decomposition of variance into account.

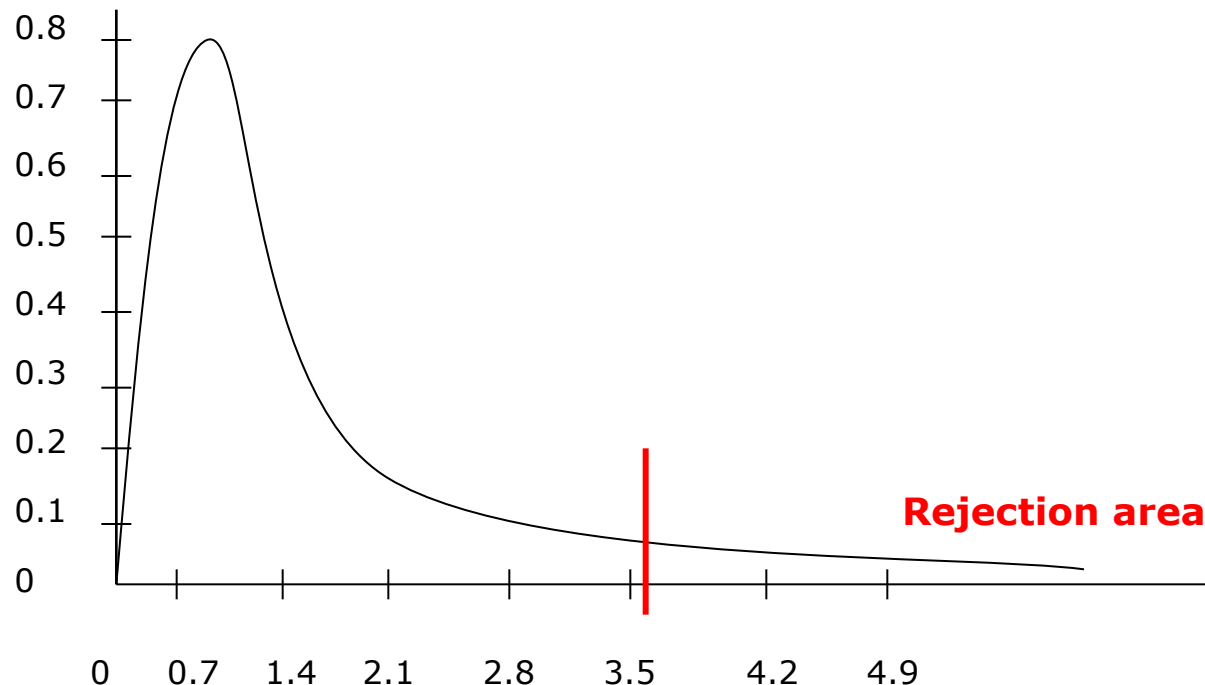
The following:

- $H_0$ : the means are all identical
- $H_1$ : there are at least two means that differ from one another

The test statistic to be used, under  $H_0$ , has the distribution of F-Fisher with a number of degrees of freedom equal to  $(c-1, n-1)$ . The more the variance between means grows and the less the variability within the categories drops, the higher the test statistic is. In addition, the more the sample grows the higher the test statistic is.

# F test to assess the hypotheses on the difference between means

The rejection area is situated in the right tail of the distribution and it is characterized by relatively high F values; if the significance level is at 5%, we reject  $H_0$  when  $F > F_{0,95}$ .



# F test to assess the hypotheses on the difference between means

**Report**

Handicraft

Age	Mean	N	Std. Deviation
18-25	5.01	78	2.224
26-35	5.53	55	2.609
36-50	6.00	41	2.098
Over 50	6.09	47	2.320
Total	5.55	221	2.352

**Measures of Association**

	Eta	Eta Squared
Handicraft * Age	.191	.036

**ANOVA Table**

		Sum of Squares	df	Mean Square	F	Sig.
Handicraft * Age	Between Groups (Combined)	44.296	3	14.765	2.733	.045
	Within Groups	1172.356	217	5.403		
	Total	1216.652	220			

**Report**

Produzione artigianale

Età	Mean	N	Std. Deviation
18-25	5.01	78	2.224
26-35	5.53	55	2.609
36-50	6.00	41	2.098
Over 50	6.09	47	2.320
Total	5.55	221	2.352

**ANOVA Table**

		Sum of Squares	df	Mean Square	F	Sig.
Between Groups (Combined)		44.296	3	14.765	2.733	.045
Handicraft * Age	Within Groups	1172.356	217	5.403		
Total		1216.652	220			

**Report**

Care for specific needs

Età	Mean	N	Std. Deviation
18-25	4.05	78	2.772
26-35	4.53	53	2.791
36-50	5.00	41	2.837
Over 50	5.83	47	8.168
Total	4.73	219	4.536

**ANOVA Table**

		Sum of Squares	df	Mean Square	F	Sig.
Between Groups (Combined)		97.921	3	32.640	1.599	.191
Care for specific needs * Age	Within Groups	4387.641	215	20.408		
Total		4485.562	218			



# Bivariate Analysis

## Objective

To describe the relationship between two variables jointly.

- **qualitative variables**: Analysis of Connection
- **quantitative variables**: Analysis of Correlation
- **mixed variables**: Analysis of Variance

# Bivariate Analysis

	Descriptive Tools	Descriptive Indexes	Statistical Test	Null Hypothesis
Connection	Contingency Table	Chi-Square Kramer's V	Chi-Square test	Statistical Indipend.
Correlation	Scatter Plot	Linear Correlation Coeffcient	t-Test	No linear relation
ANOVA	Means by Classes	Spearman Coefficient	F-Test	Indipend. by mean

# Keywords

- Connection
  - Contingency Table
  - Joint, Marginal, Conditioned Distributions
  - Chi-Square Index
  - Cramer's V Index
- Correlation
  - Covariance
  - Linear correlation index
  - Correlation matrix
- Dependence in mean
  - Comparison of means
  - Analysis of Variance (ANOVA)
  - Total Sum of Squares
  - Sum of Squares Between Groups
  - Sum of Squares Within Groups
  - Spearman Index

# Cod. 20173 - Keywords

- Statistical Test
  - Null Hypothesis vs Alternative Hypothesis
  - First Type Error
  - Second Type Error
  - p-value
  - Chi-Square Test
  - t Test
  - F Test

# Text Book

Naresh K. Malhotra, “*Marketing Research – An Applied Orientation*”,  
Pearson – Prentice Hall, Ed. 2010

- Chapter 15 – pag 493 – 502
- Chapter 16 – pag 528 – 545
- Chapter 17 – pag 561 – 566

Naresh K. Malhotra, “*Marketing Research – An Applied Orientation*”,  
Pearson – Prentice Hall, Ed. 2007

- Chapter 15 – pag 468 – 477
- Chapter 16 – pag 502 – 514
- Chapter 17 – pag 532 – 540

## Related Bibliography

• *E. Jane Miller*, The Chicago Guide to Writing about Numbers, *The University of Chicago press*, 2004