Statistics for business and decision making

Prof. Ilaria Benedetti

02. Data presentation

Frequency distribution

Suppose you have collected exam scores and country of residence for 30 students.

Individual	Country of residence	Score (0-100)	Individual	Country of residence	Score (0-100)
1	Italy	90	**16**	Tunisia	80
2	Spain	80	**17**	Spain	100
3	France	80	**18**	Spain	90
4	Lebanon	75	**19**	Spain	90
5	Lebanon	80	**20**	France	100
6	Italy	80	**21**	Ethiopia	100
7	Spain	80	**22**	Ethiopia	80
8	Portugal	90	**23**	Ethiopia	90
9	Pakistan	100	**24**	Italy	80
10	Pakistan	100	**25**	Italy	100
11	Pakistan	90	**26**	Germany	80
12	Portugal	90	**27**	Germany	90
13	Syria	78	**28**	Germany	90
14	Syria	100	**29**	Germany	90
15	France	90	**30**	Tunisia	100

You wish to examine the distribution of individuals by country of residence (i.e. how many individuals come from a certain country) and by exam score.

The information can be presented in tabular form through the construction of a frequency distribution

Frequency distribution

Country of residence	n_i Freq.	p_i Percent
Ethiopia	3	10.00
France	3	10.00
Germany	4	13.33
Italy	4	13.33
Lebanon	2	6.67
Pakistan	3	10.00
Portugal	2	6.67
Spain	5	16.67
Syria	2	6.67
Tunisia	2	6.67
Total	30	100.00

Frequency distribution

A first step toward understanding a set of data on a given variable is to explore the data and describe the data in summary form.

A **frequency distribution** is a display of the **frequency**, or number of occurrences, of each value in the data set.

The information can be presented in *tabular form*.

In a frequency distribution, a **frequency** describes the number of time or how often a category, score or range of score occurs in a certain distribution.

The **absolute frequency n_i** describes the number of times a particular value for a variable (data item) has been observed to occur.

A **relative frequency f_i** describes the number of times a particular value for a variable has been observed to occur in relation to the total number of values for that variable:

$$f_i = \frac{n_i}{n}$$

The relative frequency is calculated by dividing the absolute frequency by the total number of values for the variable.

By multiplying relative frequency per 100 we obtained **percent frequency**:

$$p_i = f_i \cdot 100$$

X	Freq.
x_1	n_1
x_2	n_2
...	...
x_j	n_j
...	...
x_k	n_k
Totale	n

x_1; x_2; … ; x_j; …; x_n they are the distinct modalities that the character X assumes in the collective of N units examined.

n_1; n_2; … ; n_j; …; n_k are the absolute frequencies associated with each modality.
n_1 indicates how many units have the x_1 modality of the character X.

$$\sum_{j=1}^{K} n_j = n$$

The sum of the absolute frequencies is equal to the total number of units of the collective

Frequency table

- **Absolute frequency "n_i"**
- **Relative frequency "f_i"**

Cumulative frequency distribution shows the total number of occurrences that lie above or below certain key values.

We construct a cumulative frequency distribution by adding the frequencies of all frequency distribution classes up to and including the present class.

- **cumulative absolute frequency "N_i"**
- **cumulative relative frequency "F_i"**
- **cumulative percentage frequency "P_i"**

Score (0-100)	n_i Freq.	p_i Percent	P_i Cum.
75	1	3.33	3.33
78	1	3.33	6.67
80	9	30.00	36.67
90	11	36.67	73.33
100	8	26.67	100.00
Total	30	100.00	

EXERCISE **Frequency distribution**

The daily household expenses for food products (expressed in euro) of a sample 20 Italian households was measured.
The data collected are:

23	20	15	18	21
15	14	18	20	20
20	18	20	24	23
20	15	21	20	20

- Declare the type of data
- Construct a frequency distribution and display it as a table (absolute and relative, and cumulative)

With R:

```
#create a vector
data<-c(23, 20, 15,18, 21, 15, 14, 18, 20, 20, 18, 20, 20, 20, 18, 20, 24, 23, 20, 15, 21, 20, 20)
# absolute freq:
nj<-table(data)
# relative freq
fj<-prop.table(nj)
# cum. Rel- freq
Fj<-cumsum(fj)
```

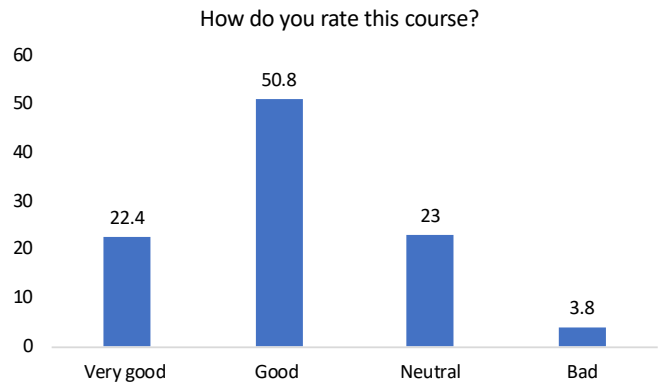
Data presentation

- **FREQUENCY DISTRIBUTIONS**

How do you rate this course?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Very Good	41	22.4	22.4	22.4
	Good	93	50.8	50.8	73.2
	Neutral	42	23.0	23.0	96.2
	Bad	7	3.8	3.8	100.0
	Total	183	100.0	100.0	

- **GRAPHICAL DISTRIBUTIONS**



From tabular to graphical representation

- Frequency distributions are good ways to present the essential aspects of data collections in concise and understable terms
- Pictures are always more effective in displaying large data collections

Histogram

A **histogram** is a graph that consists of vertical bars constructed on a horizontal line that is marked off with intervals for the variable being displayed.

The **intervals** correspond to the classes in a frequency distribution table.

The **height** of each bar is proportional to the number of observations in that interval.

The **number of observations** can be displayed above the bars.

- ✓Frequently used to graphically present interval and ratio data;
- ✓The adjacent bars indicate that a numerical range is being summarized by indicating the frequencies in arbitrarily chosen classes

Histogram

A histogram gives an idea about the distribution of a quantitative variable. The idea is to break the range of values into intervals and count how many observations fall into each interval. Histograms are a bit similar to barplots, but histograms are used for quantitative variables whereas barplots are used for qualitative variables. To draw a histogram in R, use `hist()` :

```
hist(dat$Sepal.Length)
```

Add the arguments breaks = inside the hist() function if you want to change the number of bins.

A rule of thumb (known as Sturges' law) is that the number of bins should be the rounded value of the square root of the number of observations.

Example: The dataset includes 150 observations so in this case the number of bins can be set to 12.

Shape of a Distribution

We can describe graphically the shape of the distribution by a histogram. That is, we can visually determine whether data are evenly spread from its middle or center.

Sometimes the center of the data divides a graph of the distribution into two "mirror images," so that the portion on one side of the middle is nearly identical to the portion on the other side.

- Graphs that have this shape are symmetric;
- those without this shape are asymmetric (or skewed).

Symmetry

The shape of a distribution is said to be **symmetric** if the observations are balanced, or approximately evenly distributed, about its center.

Skewness

A distribution is skewed, or **asymmetric**, if the observations are not symmetrically distributed on either side of the center.

A **skewed-right distribution (**positively skewed) has a tail that extends farther to the right.

A **skewed-left distribution** (negatively skewed) has a tail that extends farther to the left.

Mean=median=mode

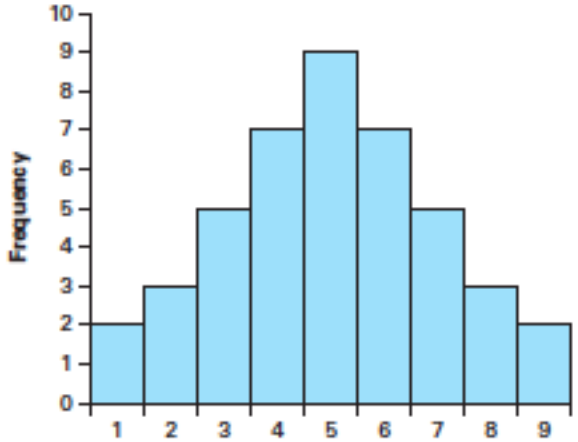


Figure 1.15(a) Symmetric Distribution

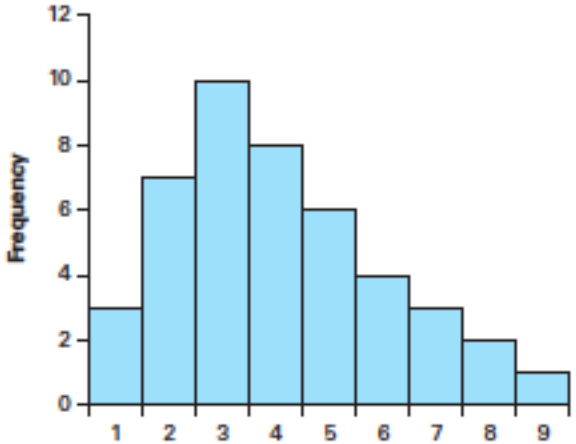


Figure 1.15(b) Skewed-right Distribution

positively skewed

Mean>median>mode

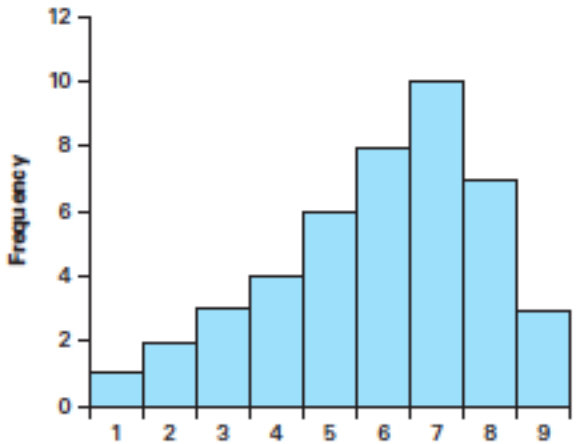


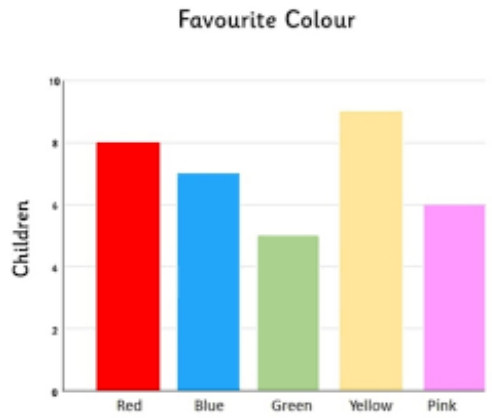
Figure 1.15(c) Skewed-left Distribution

negatively skewed

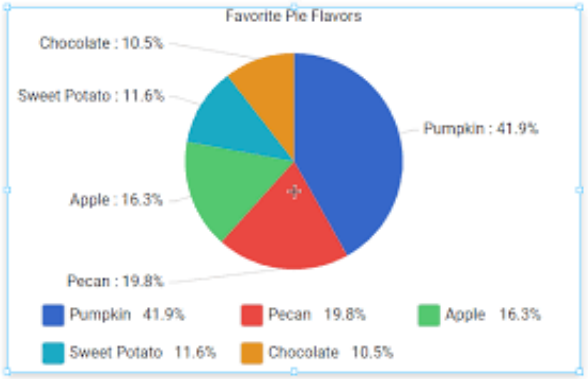
Mean<median<mode

Graphic distributions: discrete and categorical data

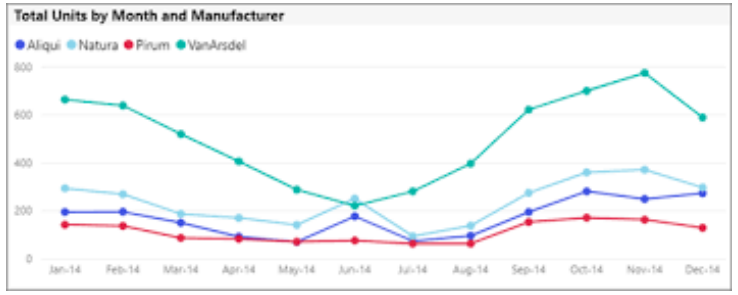
Bar chart



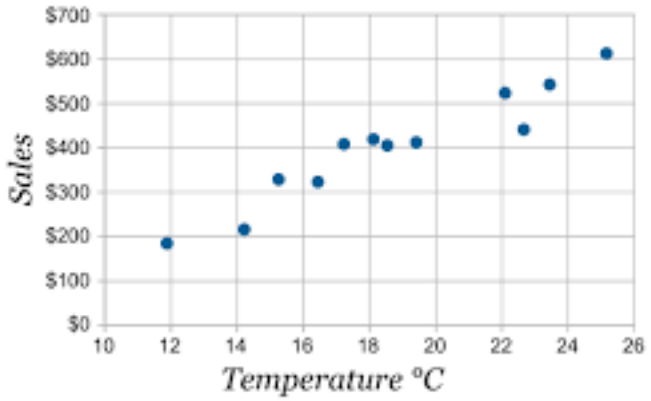
Pie chart



Line chart



Scatter plot



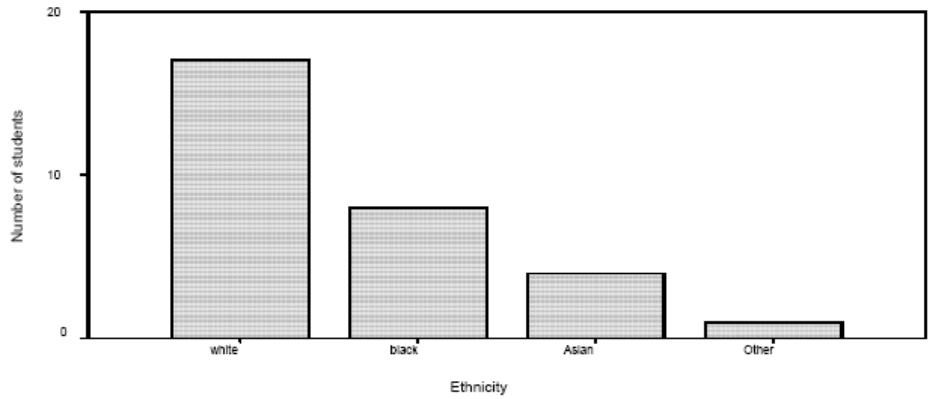
Bar charts – nominal/ordinal variables:

Looks like a histogram, except:

- In a bar chart, the categories which the bars represent do not have to be ordered. A gap is usually left between bars to reflect this.

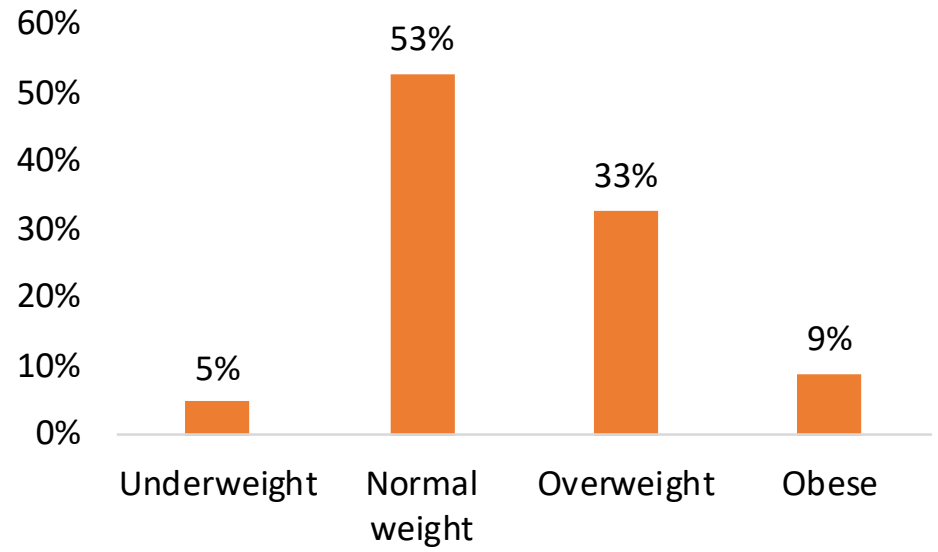
- The bars in a bar chart are all the same width.

Bar chart showing number of students by ethnicity

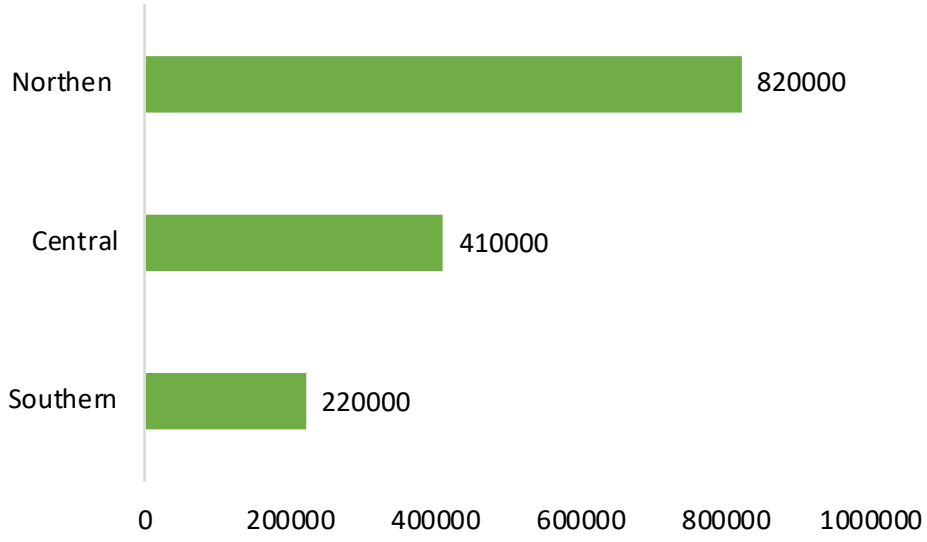


Bar chart – dqualitative data

BMI(xj)	(pj)
Underweight	5%
Normal weight	53%
Overweight	33%
Obese	9%



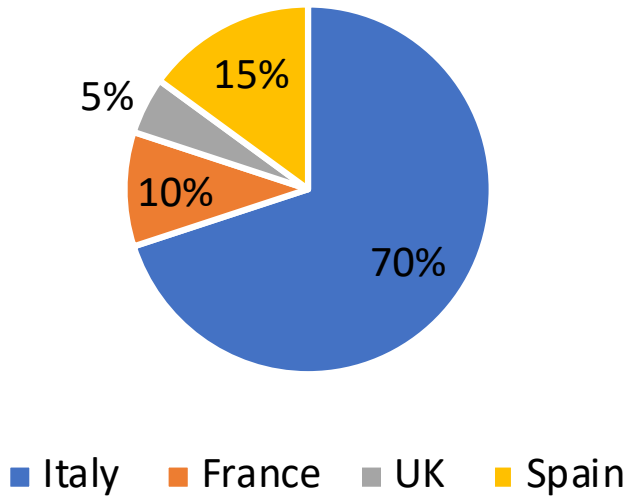
Geographical areas (xj)	Population (nj)
Southern	220000
Central	410000
Northen	820000



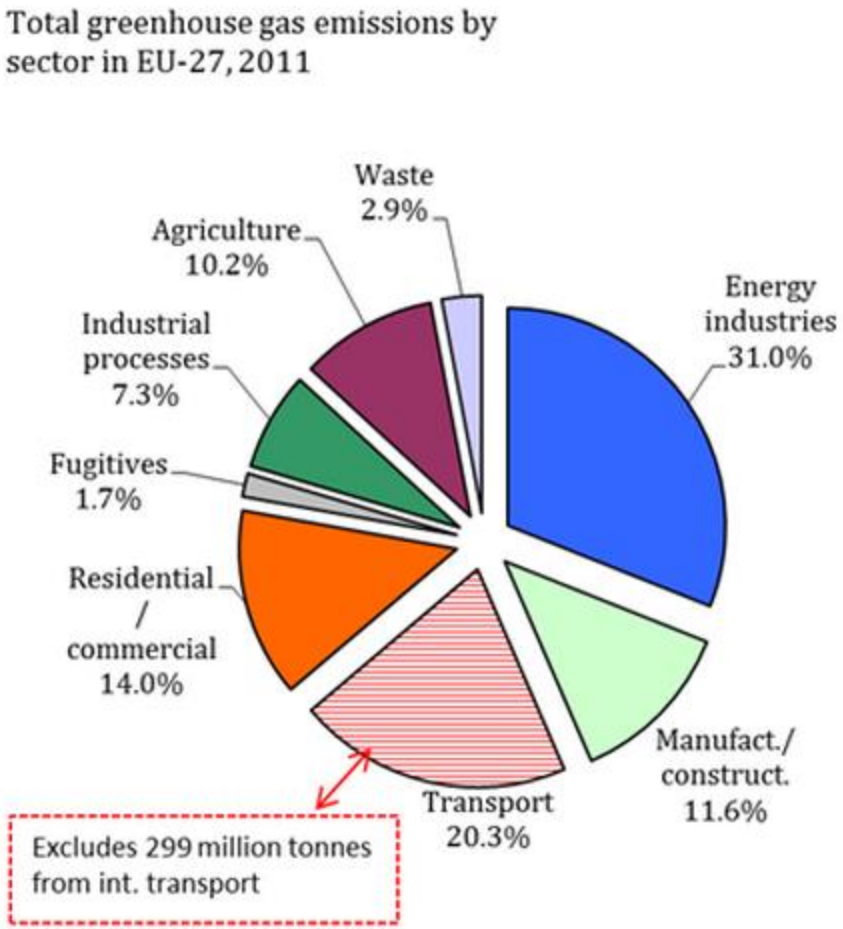
Pie charts

In a pie chart, each category of the variable is represented by a segment of a circle. **The size of the segment is proportional to the number of cases in that category**.

Pie chart showing distribution of students by country of origin



Pie chart

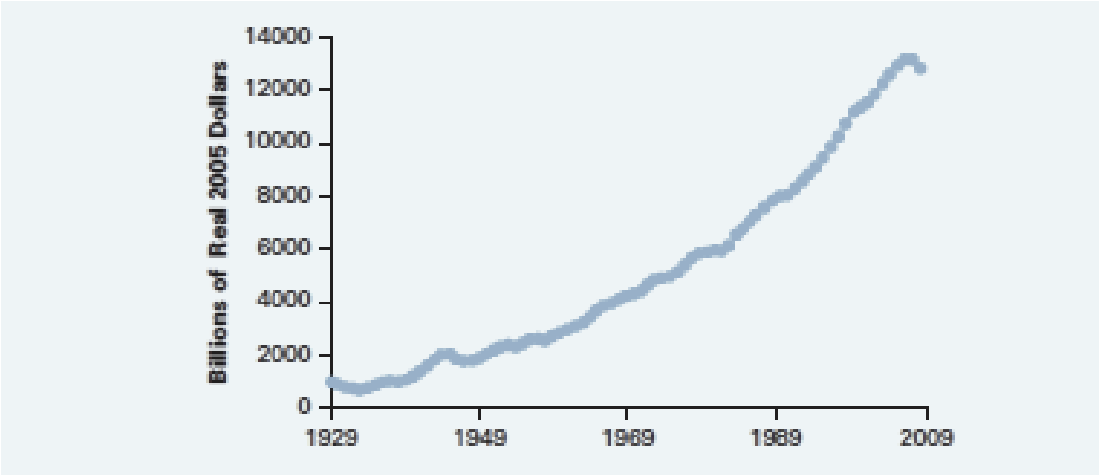

Total greenhouse gas emissions by sector in EU-27, 2011
Waste 2.9%
Agriculture 10.2%
Energy industries 31.0%
Industrial processes 7.3%
Fugitives 1.7%
Residential / commercial 14.0%
Manufact./ construct. 11.6%
Transport 20.3%
Excludes 299 million tonnes from int. transport

Time-series plot

A time series is a set of measurements, ordered over time, on a particular quantity of interest.
In a time series the sequence of the observations is important.

A line chart, also called a **time-series plot**, is a series of data plotted at various time intervals.
Measuring time along the horizontal axis and the numerical quantity of interest along the vertical axis yields a point on the graph for each observation.
Joining points adjacent in time by straight lines produces a time-series plot.

Examples of time-series data include annual university enrollment, annual interest rates, the gross domestic product over a period of years…



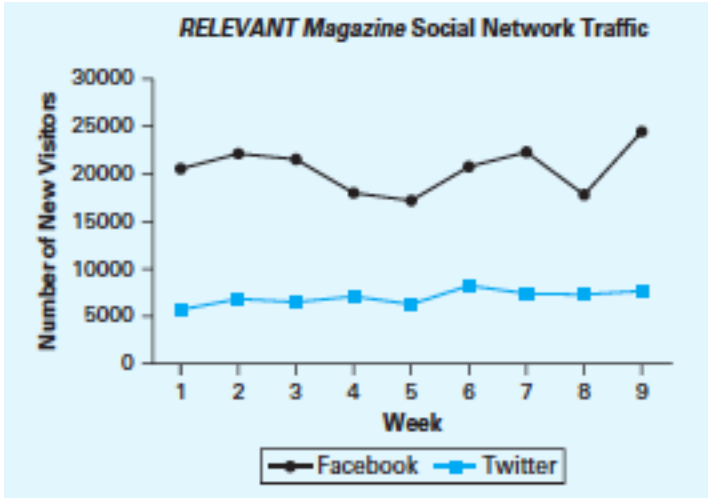
Gross Domestic Product by Time: 1929–2009 (Time-Series Plot)

Example: Social Network Traffic (Time-Series Plot)

RELEVANT Magazine keeps records of traffic (such as the number of weekly new visitors) to its Web site from various social networks such as Facebook and Twitter (Butcher 2011). This information may be helpful to Richard Butcher, Marketing Assistant of RELEVANT Magazine. Plot the number of weekly new visitors for a recent 9-week period from both Facebook and Twitter. Use a time-series plot. The data are stored in the data file RELEVANT Magazine.

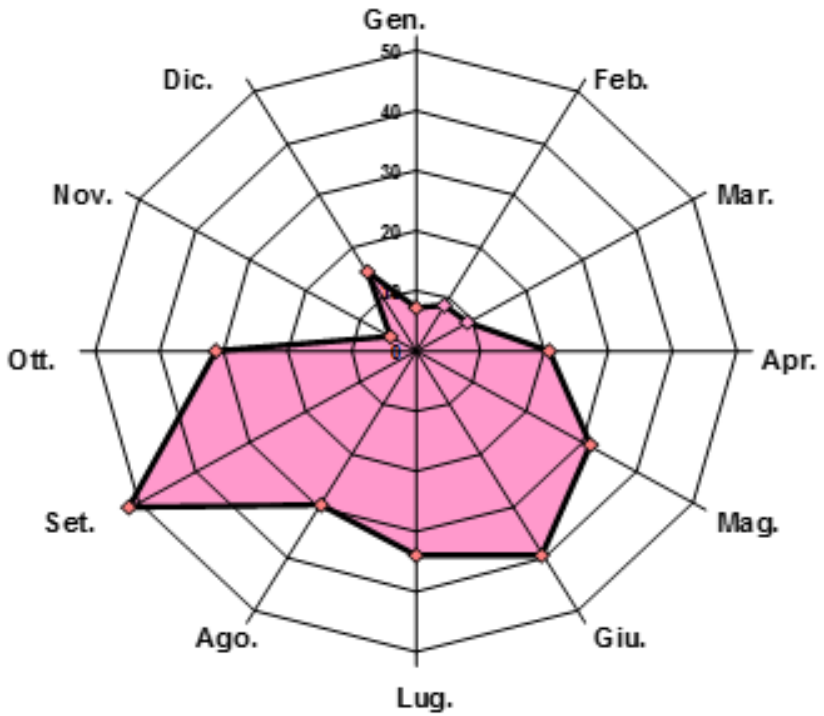
In the table below we obtain the number of weekly new visitors for a recent 9-week period from both Facebook and Twitter.

Week	Twitter	Facebook
1	5,611	20,499
2	6,799	22,060
3	6,391	21,365
4	6,966	17,905
5	6,111	17,022
6	8,101	20,572
7	7,370	22,201
8	7,097	17,628
9	7,531	24,256



Radar or spike graph

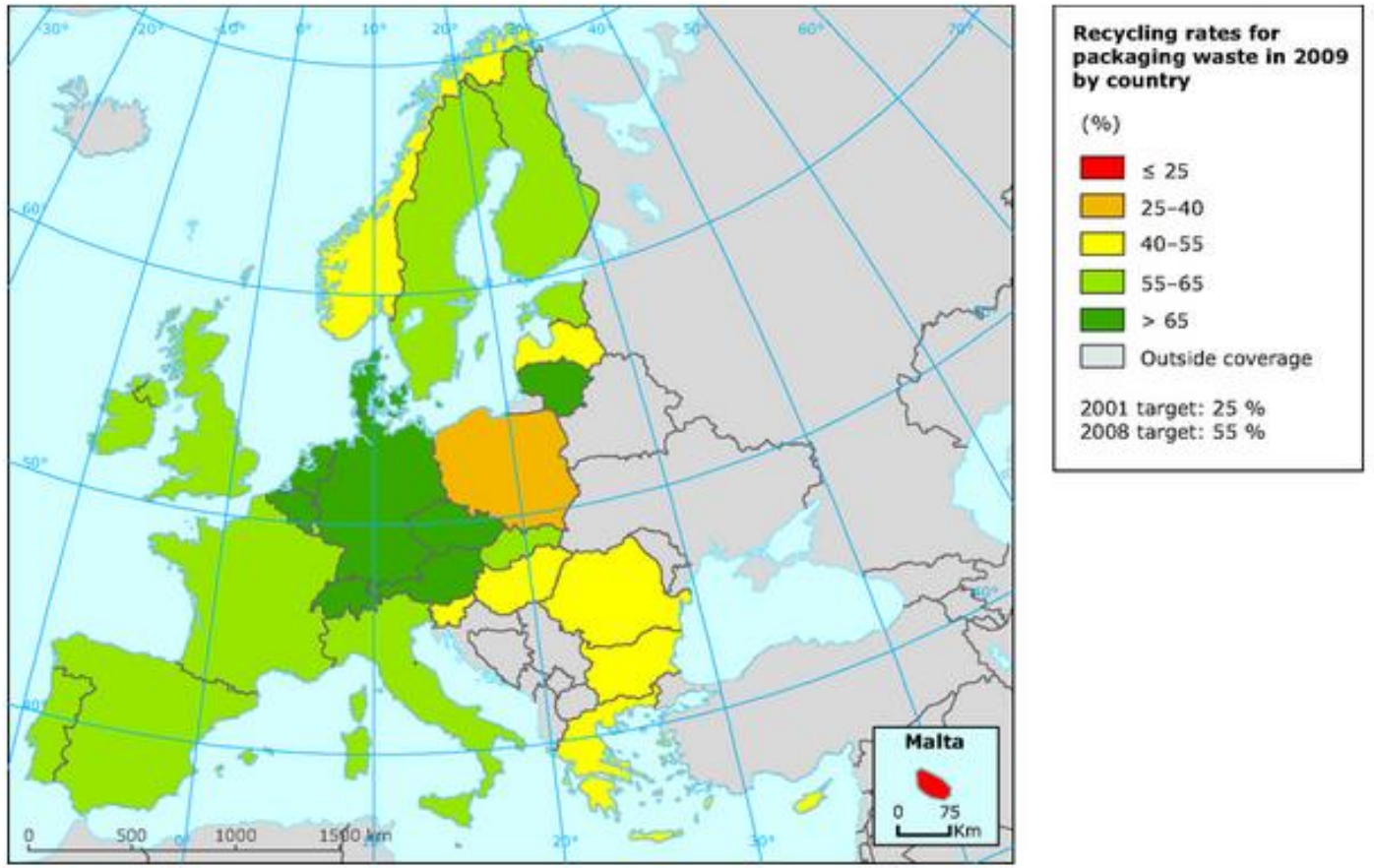
N. Of wedding in italy (by month) year 1998.



Cyclic information

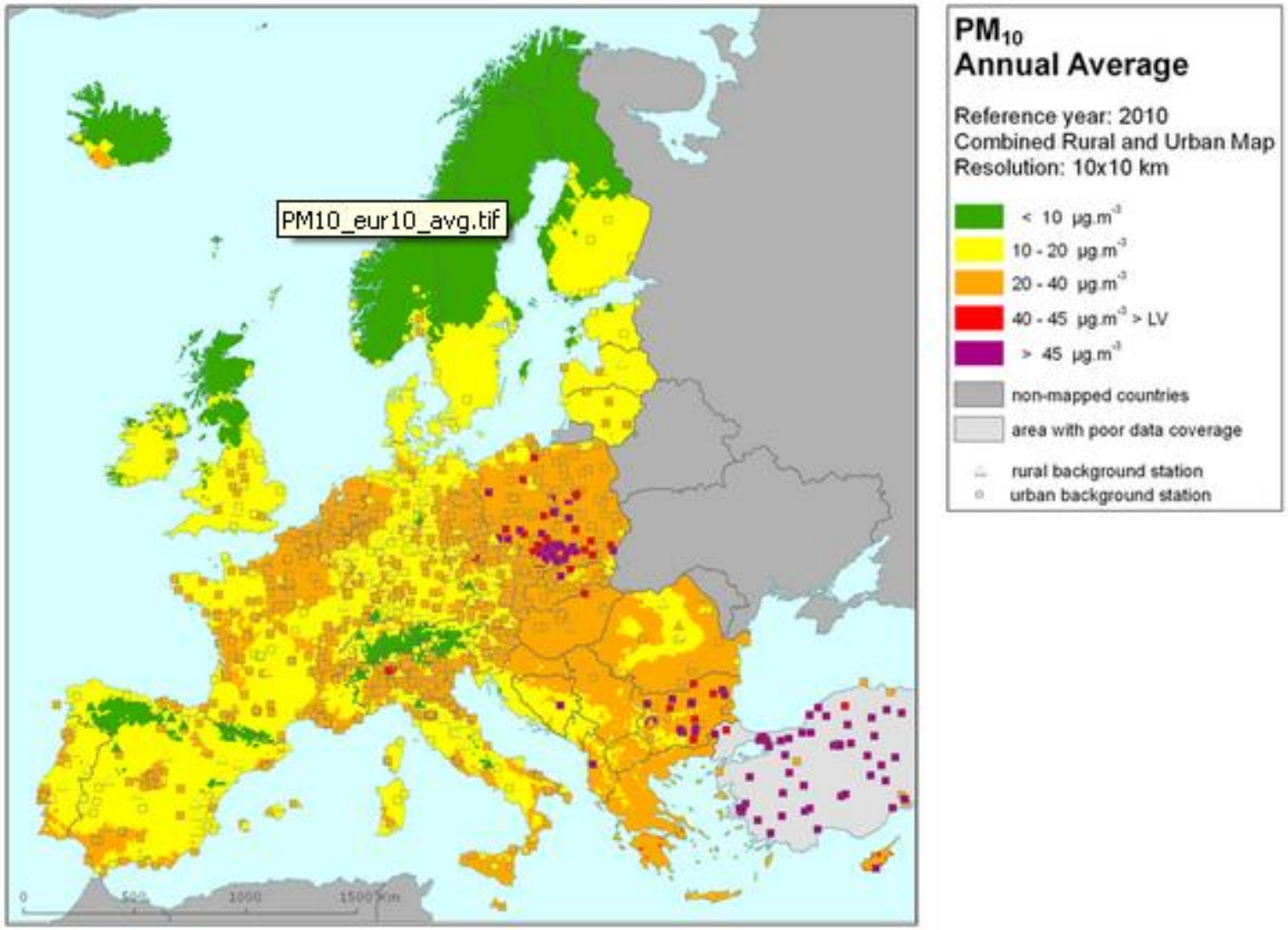
Cartogram

Cartograms focuses on the magnitude of a (variable) indicator analysed thus enabling everybody to sight the performance of a specific area and making comparison with the performance of other neighbours at the same level and with similar characteristics



Source: European Environmental Agency

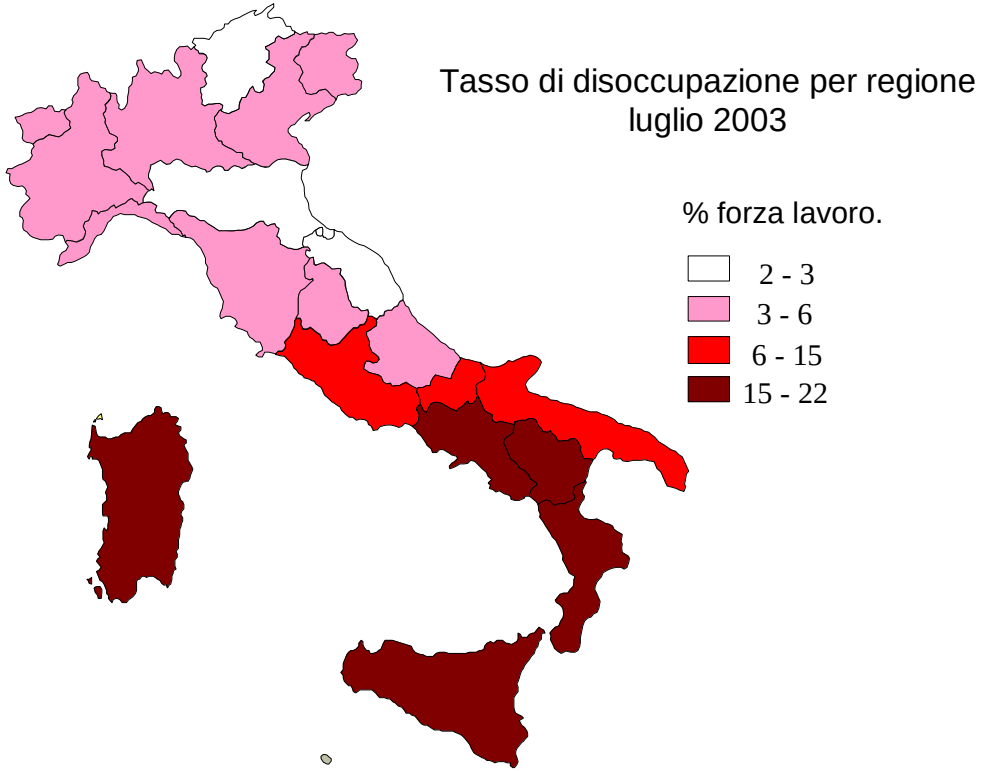
Cartogram



Source: European Environmental Agency

Cartogram

Cartogram: Unemployment rate by italian regions (July 2003)



Graphical Presentation of Bivariate Data

Questions in deciding on an appropriate graph:

i) *Which variable is the dependent variable?* Values of the dependent variable should be plotted on the vertical ('y') axis. Values of the explanatory variable are plotted on the horizontal ('x') axis.

ii) How many categories does the dependent variable have?

iii) What is the level of measurement of the explanatory variable?

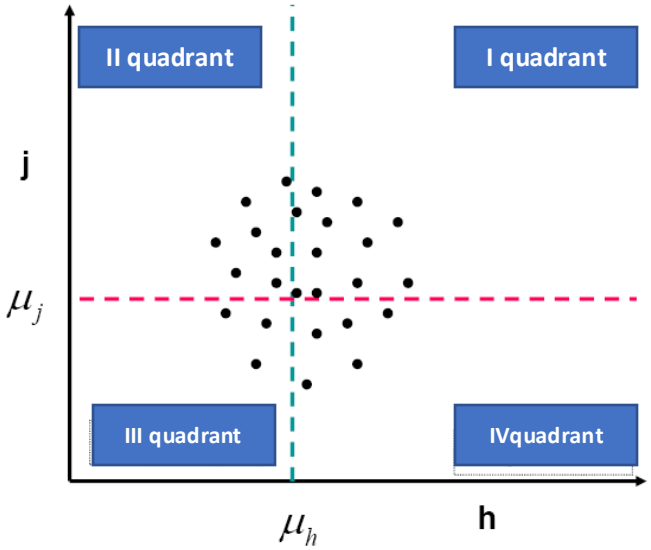
Dependent and Explanatory Variables

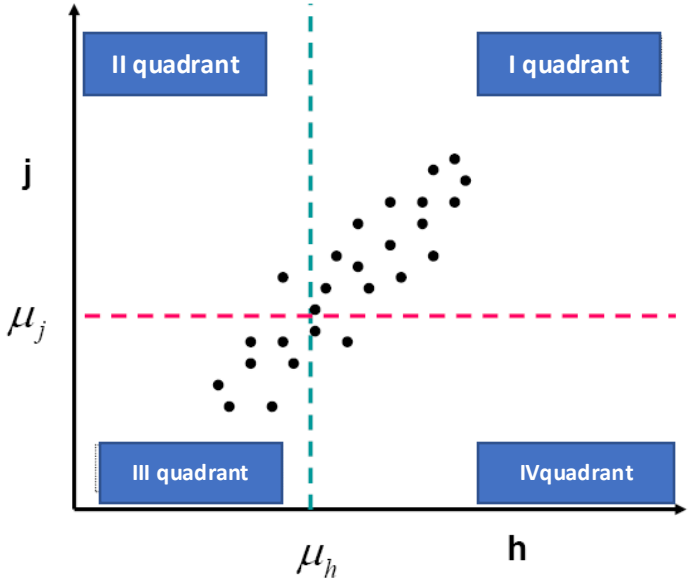
The **dependent variable** (also called the *response or outcome variable*) is the variable being predicted or explained by the **explanatory variable** (also called a *predictor or independent variable*).

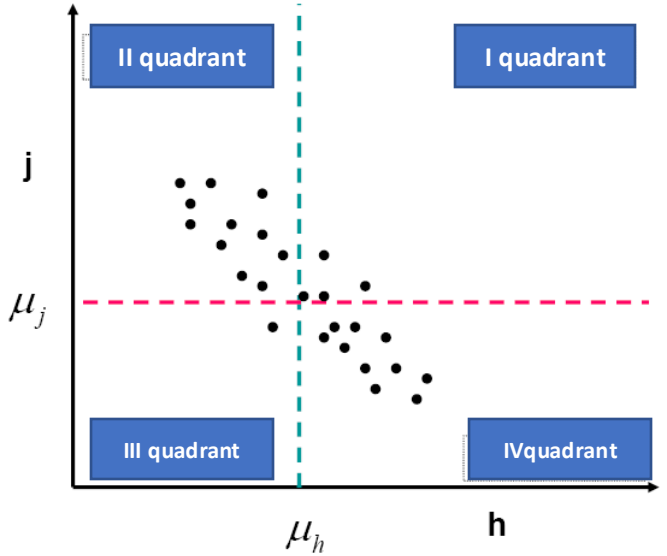
Displaying the Relationship Between 2 Interval/Ratio Variables: Scatter Plots

Each individual (statistical unit) is represented by a point (*x*,*y*) on a plot, where *x* is their value on the explanatory variable and *y* is their value on the dependent variable.

A (graphical) introduction to the study of relationship between two variables



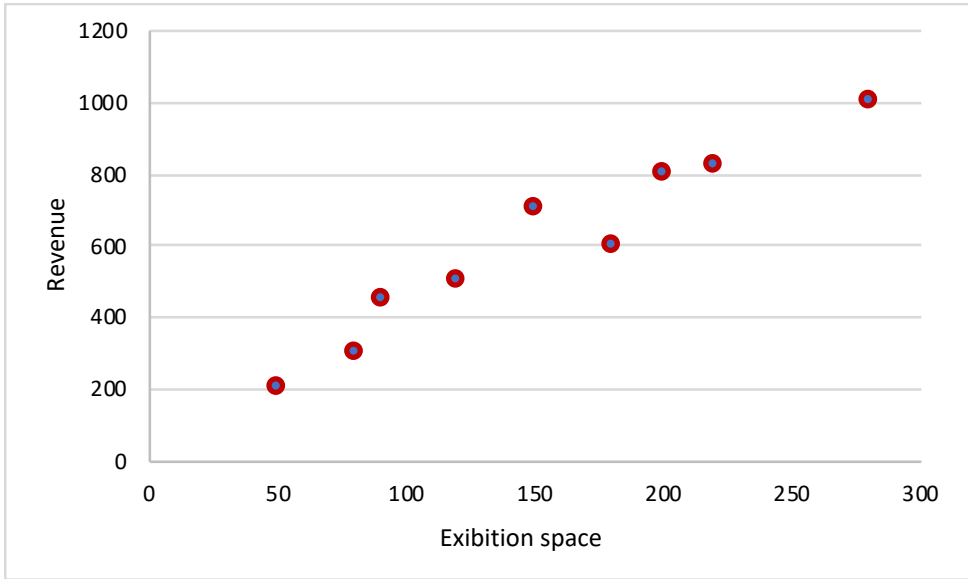




Scatter plot

For quantitative data

Exhibition space (square meters)	Revenue (€)
50	200
80	300
90	450
120	500
150	700
180	600
200	800
220	820
280	1000



Type of chart	Type of data
Bar graph	*Qualitative data – Nominal variables*
Histogram	*Quantitative data – Interval and ratio variablese*
Pie chart	*Qualitative data – Nominal variables*
Radar or spike chart	*Qualitative data*
Cartogram	*Territorial data*
Scatter plot	*Times series data*