BIG DATA STATISTICS FOR BUSINESS

AY 2023-24

Giovanni De Luca
Parthenope University of Naples

1

Exploratory analysis

- The collection and storage of data are not exhaustive in themselves.
- The processing step is fundamental: it allows the achieving of the goal of supporting business decisions.
- The purpose of collecting and processing large volumes of complex data is to understand the trends of the phenomena of interest, uncover hidden trends, detect anomalies, etc., to make data-driven decisions.

2

2

Data matrix

- Structured data can be arranged in a data matrix.
- The data matrix is a two-dimensional table whose rows are associated with statistical units and columns are associated with variables.
- The statistical units can represent the entire population or constitute a (representative) sample of the population.
- We denote by n the number of statistical units and by p the number of statistical variables ($n > p$).
- The data matrix is the starting point for data analysis.

3

3

Variables

- The number of variables, even if high, has to be lower than the number of statistical units.
- There are two main types of variables:

1. numerical variables
2. categorical variables

4

4

Numerical variables

- Numerical variables are quantitative and are classified into:
  - discrete numerical variables (derive from a counting process, they take integer values, e.g. the number of cigarettes a person smokes a day).
  - continuous numerical variables (can take any value such as heights if measured with enough precision, e.g. 68.1 or 68.09 or 68.092 inches).

5

5

Categorical variables

- Categorical variables are presented in non-numerical form (categories), and do not allow any metric statement on the differences between categories.
- They can be:
  - Ordinal categorical variables (spiciness can be mild, medium, or hot. Even if they are not numbers per se, they can still be ordered)
  - Non-ordinal categorical variables (sex at birth, or regions of a country)

6

6

Data matrix

- Data matrix can contain numerical as well as categorical variables.
- Sometimes, categorical variables are translated into numerical variables.

7

7

- Example: dataset *Diamonds* describes almost 54,000 diamonds using numerical and categorical variables. The data matrix has n = 53,940 rows and p = 10 columns.

	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57.0	336	3.95	3.98	2.47
8	0.26	Very Good	H	SI1	61.9	55.0	337	4.07	4.11	2.53
9	0.22	Fair	E	VS2	65.1	61.0	337	3.87	3.78	2.49
10	0.23	Very Good	H	VS1	59.4	61.0	338	4.00	4.05	2.39
11	0.30	Good	J	SI1	64.0	55.0	339	4.25	4.28	2.73
12	0.23	Ideal	J	VS1	62.8	56.0	340	3.93	3.90	2.46
13	0.22	Premium	F	SI1	60.4	61.0	342	3.88	3.84	2.33

8

- Example: dataset *mtcars* describes the performance of 32 cars based on 11 variables. The data matrix has n = 32 rows and p = 11 columns.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4

Exploratory data analysis and data visualization

Exploratory data analysis

- After getting data, an exploratory data analysis (or preliminary data analysis) is carried out to grasp its main characteristics.
- The analysis includes the calculation of simple statistics (summary measures) and the visualization of the data with the most appropriate graphics.
- This step is also known as pre-processing.

11

11

Summary measures for numerical variables

- The most important summary measures of a numerical variable are:

1. Mean
2. Median
3. Quartiles
4. Mode
5. Skewness

12

12

- The **mean** is the arithmetic average.
- Defined X the variable, the mean is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- The mean is sensitive to outliers (extreme values).

13

13

- After sorting the data in ascending order, the **median** (Me) is the central value of the (preceded by 50% of the data and followed by the remaining 50% of the data).

- The median is not sensitive to outliers (extreme values). It is a robust measure.

14

14

- After sorting the data in ascending order, the **first quartile** (Q_1) is the value preceded by 25% of the data and followed by the remaining 75% of the data.
- The first quartile equals the 25th percentile,

$$Q_1 = P_{25}$$

- After sorting the data in ascending order, the **third quartile** (Q_3) is the value preceded by 75% of the data and followed by the remaining 25% of the data.
- The third quartile equals the 75th percentile,

$$Q_3 = P_{75}$$

- The median is also the second quartile.
- In general, the *p*-percentile is the value preceded by p% of the data.

- The **mode** is the most frequent number. It makes sense when we have discrete numerical variables.
- To find the mode, we first have to organize the data in a table.

17

17

- The **skewness** is the degree of asymmetry observed in data (a measure of the shape of the distribution).
- Skewness can be described as a measure of the extent to which a distribution departs from a symmetric (e.g. normal) distribution.

18

18

- A distribution shows left (negative) skewness if we observe a long tail on the left (values much smaller than the mean).
- A distribution shows right (positive) skewness if we observe a long tail on the right (values much larger than the mean).
- The skewness is usually detected graphically.

19

19

Data visualization

- Sometimes, extracting information just by looking at the numbers is quite difficult.
- Data visualization provides a powerful way to communicate a data-driven finding.
- Data visualization is one of the strongest tool for exploratory data analysis ("A picture is worth a thousand words“).
- “The greatest value of a picture is when it forces us to notice what we never expected to see.” (Tukey).
- Histogram.
- Box-plot.

20

20

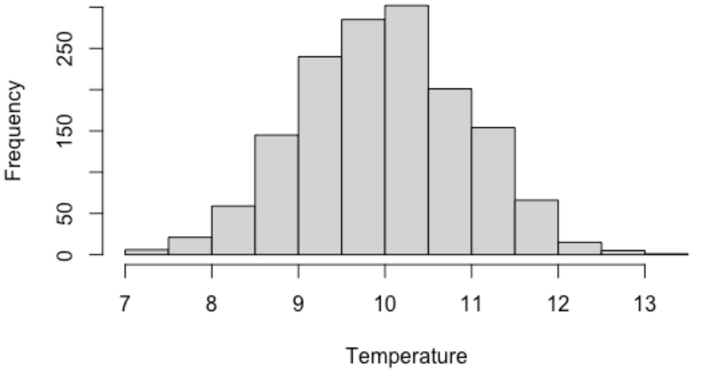
Histogram

- The **histogram** is a representation of the distribution of data.
- It is obtained by dividing the entire range of values into a series of intervals and counting how many values fall into each interval.
- The bins are non-overlapping.

21

21

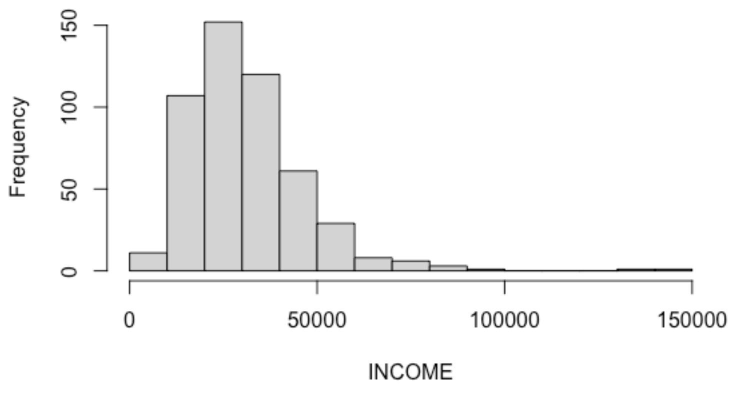
Histogram (example 1)



22

22

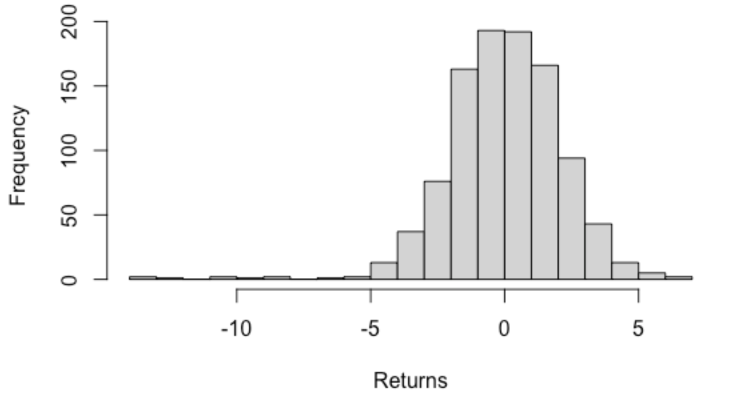
Histogram (example 2)



23

23

Histogram (example 3)



24

24

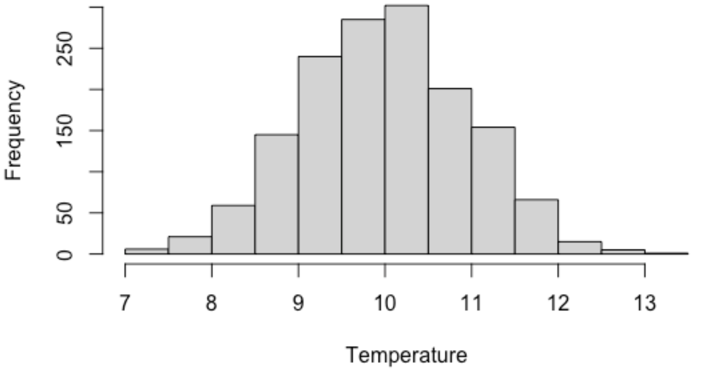
Histogram with density

- We can draw the histogram using frequency densities (on the y-axis) instead of frequencies.
- The frequency densities are computed in such a way that all histogram area is equal to 1.
- We can interpret the bins of the histogram in terms of proportions (or probability).

25

25

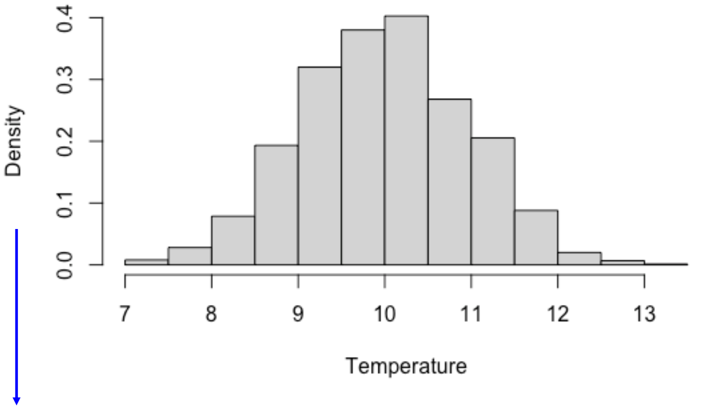
Histogram (example 1)



26

26

Histogram (example 1)

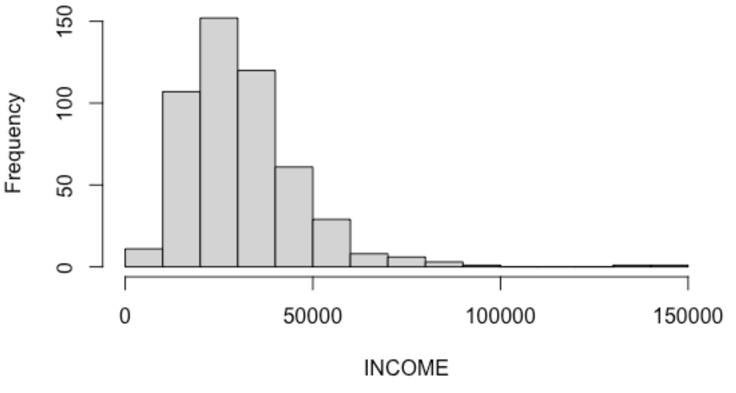


The area of the bins sums to 1.

27

27

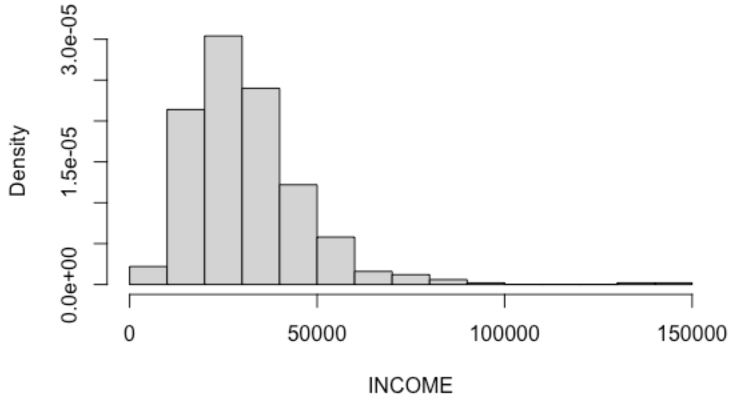
Histogram (example 2)



28

28

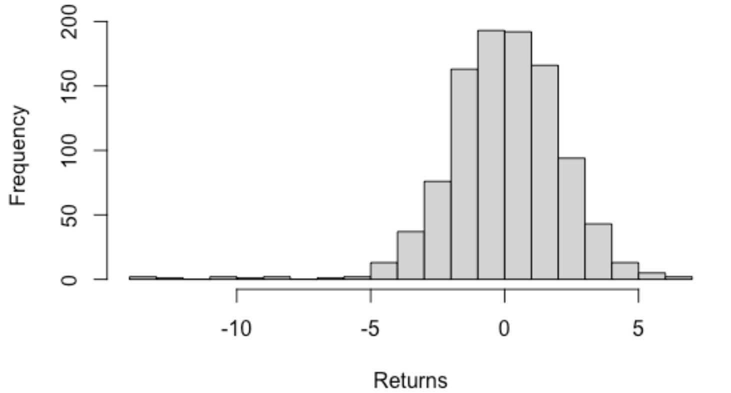
Histogram (example 2)



29

29

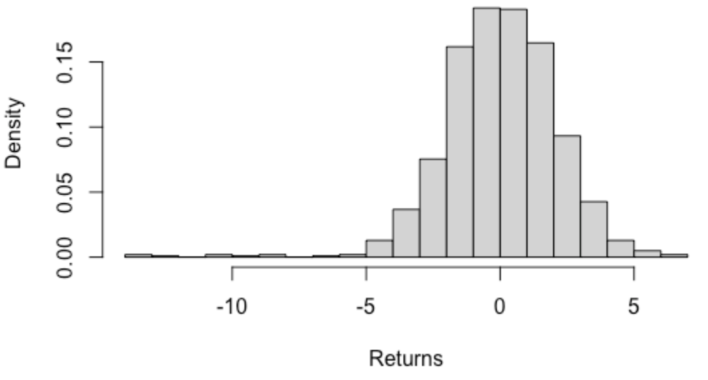
Histogram (example 3)



30

30

Histogram (example 3)



31

31

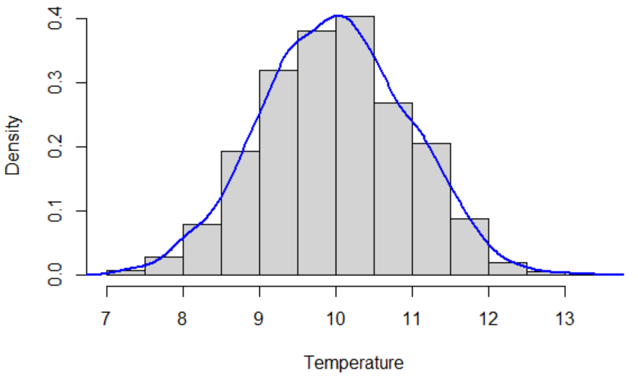
Smooth density plot

- When the histogram is represented with densities, we can draw a smooth density plot.
- Smooth density plots are similar to histograms but are aesthetically more appealing.

32

32

Density plot (example 1)



33

33

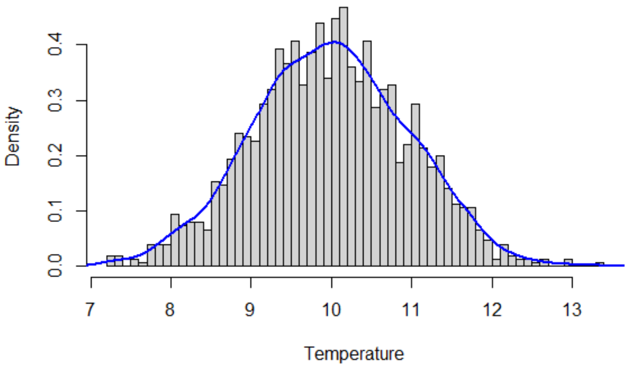
Smooth density plot

- Instead of making a histogram with tiny bins, we can draw this smooth curve.
- Note that “smooth” is a relative term. We can control the degrees of smoothness of the curve.

34

34

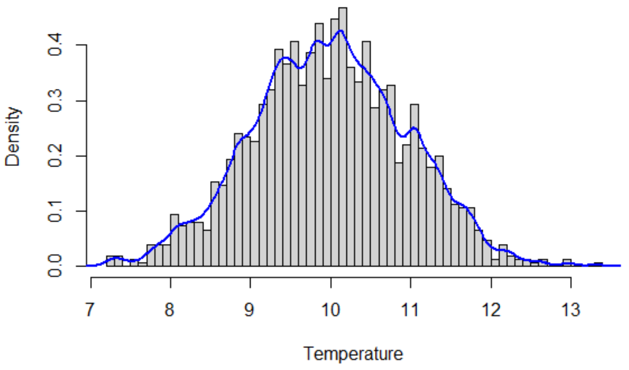
Density plot (example 1)



35

35

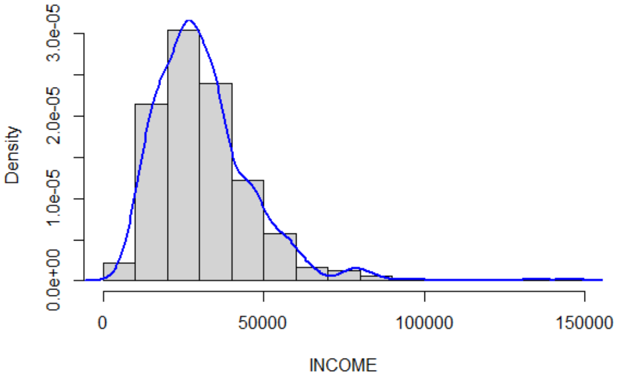
Density plot (example 1)



36

36

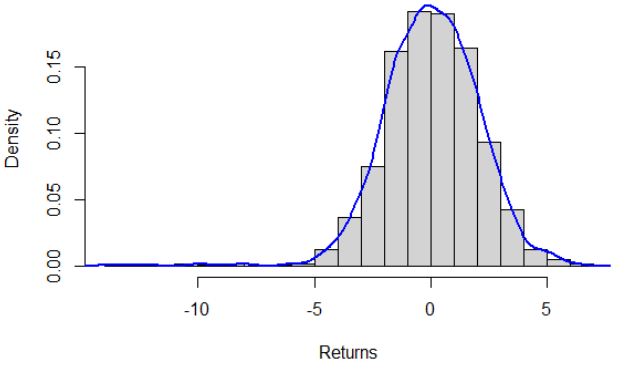
Density plot (example 2)



37

37

Density plot (example 3)



38

38

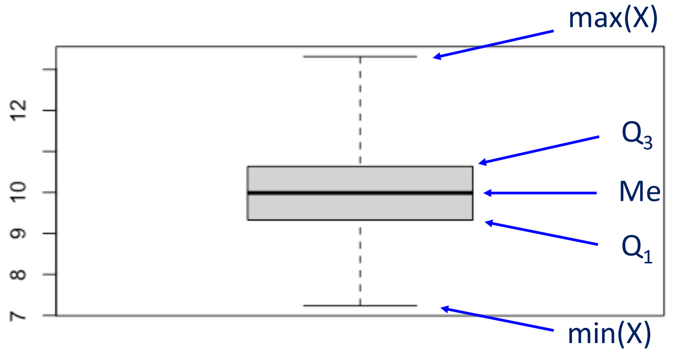
Box plot

- The **box and whiskers plot** (or **box plot**) is a graphical representation to describe the distribution of a set of data through simple indexes.
- In its simplified version, the plot shows
  - min(X)
  - 1° quartile of X (Q_1)
  - Median of X, Me(X)
  - 3° quartile of X (Q_3)
  - max(X)

39

39

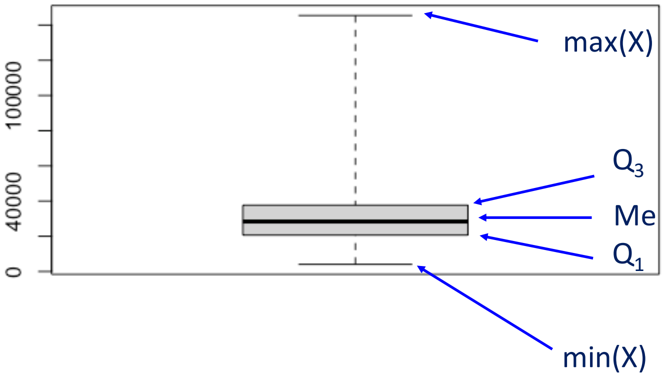
Box plot (example 1, Temperature)



40

40

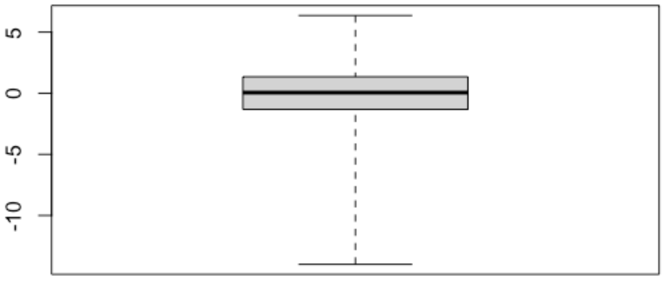
Box plot (example 2, Income)



41

41

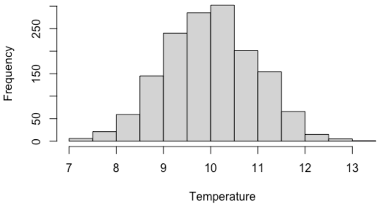
Box plot (example 3, Returns)

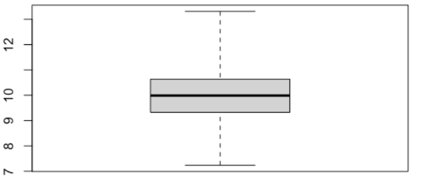


42

42

Histogram and Box plot (example 1)

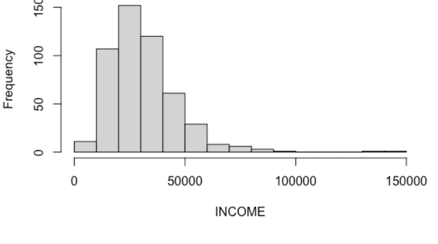


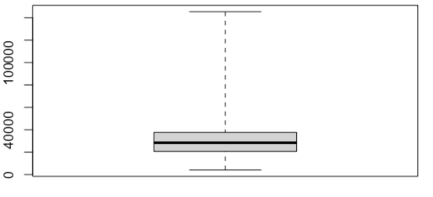


43

43

Histogram and Box plot (example 2)

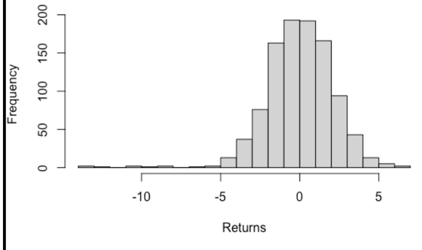


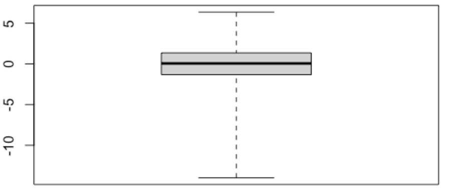


44

44

Histogram and Box plot (example 3)





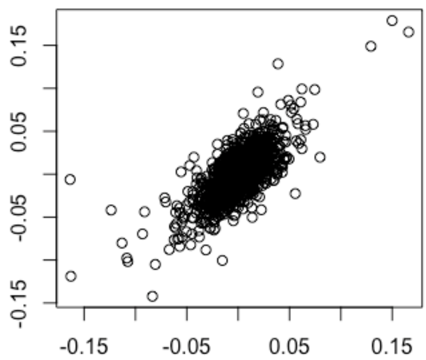
45

Scatterplot

- Scatterplot is one of the best known bivariate graphs.
- It highlights the (positive or negative) association between two variables.
- Positive association (or positive relationship, or concordance): the two variables tend to move in the same direction and a straight line (regression line) with a positive slope can be drawn.
- Negative association (or negative relationship or discordance): the two variables tend to move in opposite directions and a straight line (regression line) with a negative slope can be drawn.

46

Scatterplot (example 1)

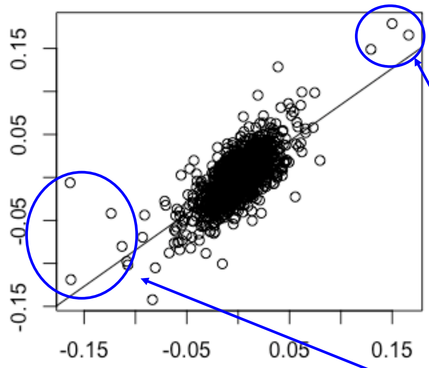


Positive association (or concordance).

Scatterplot

- A more accurate analysis also tends to interpret the behavior also in the tails (thus evaluating the association between extreme values).
- Association between extremely high values: upper tail dependency
- Association between extremely low values: lower tail dependency

Scatterplot (example 1)



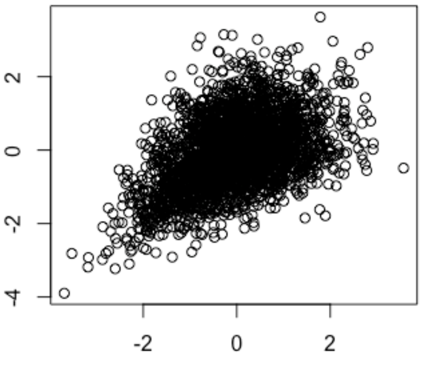
Positive association (or concordance).

Extremely high values show strong association

Extremely low values show weak association

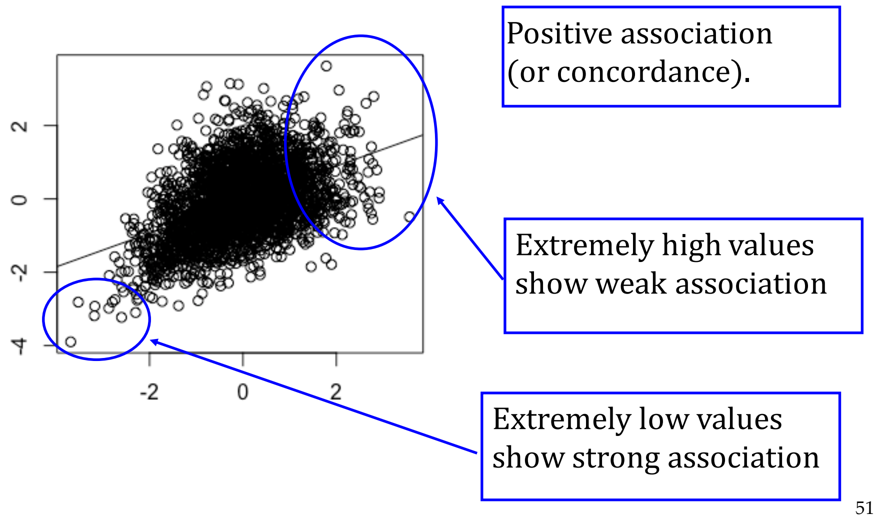
49

Scatterplot (example 2)



50

Scatterplot (example 2)



51

51

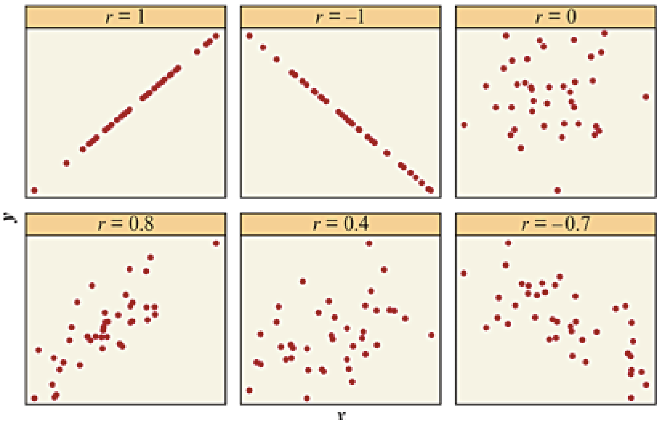
Correlation

- The correlation *r* measures the strength and the direction of the association between two numerical variables.
- Correlation always falls between -1 and +1.
- Sign of correlation denotes direction:
  - (-) indicates a negative association.
  - (+) indicates a positive association.
- Correlation does not depend on the variables' units.
- Two variables have the same correlation no matter which is treated as the response variable.
- Correlation is not resistant to outliers.
- Correlation only measures the strength of a linear relationship.

52

52

Correlation



53

53

Correlation

- Hypothesis testing is usually carried out.
- The null hypothesis (no correlation) is

$$H_0: r = 0$$

- We reject H_0 if the p-value of the test is less than 0.05,

$$p\text{-value} < 0.05$$

54

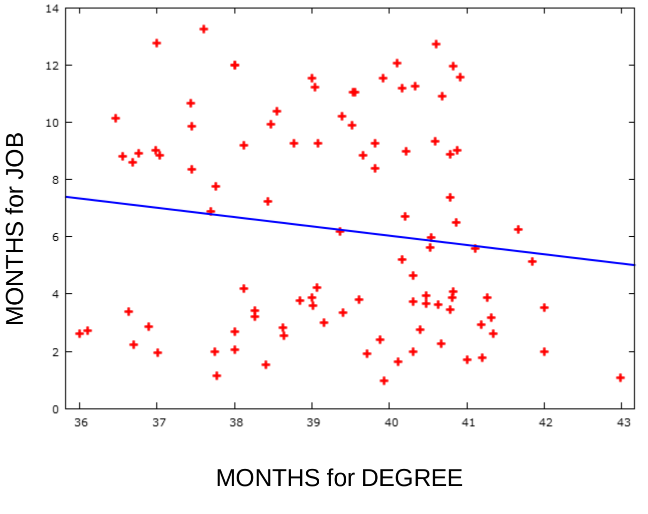
54

Caution in analyzing association (1)

- The direction of an association between two variables can change after including a third variable and analyzing the data at separate levels of that variable (Simpson's paradox).
- Example: we analyze the data of 114 employed three-year graduates; the time (in months) to graduate and the time (in months) to find employment are recorded.
- I assume that companies prefer those who graduate quickly.
- Therefore, I expect a direct relationship between the time to graduate and the time to find a job.

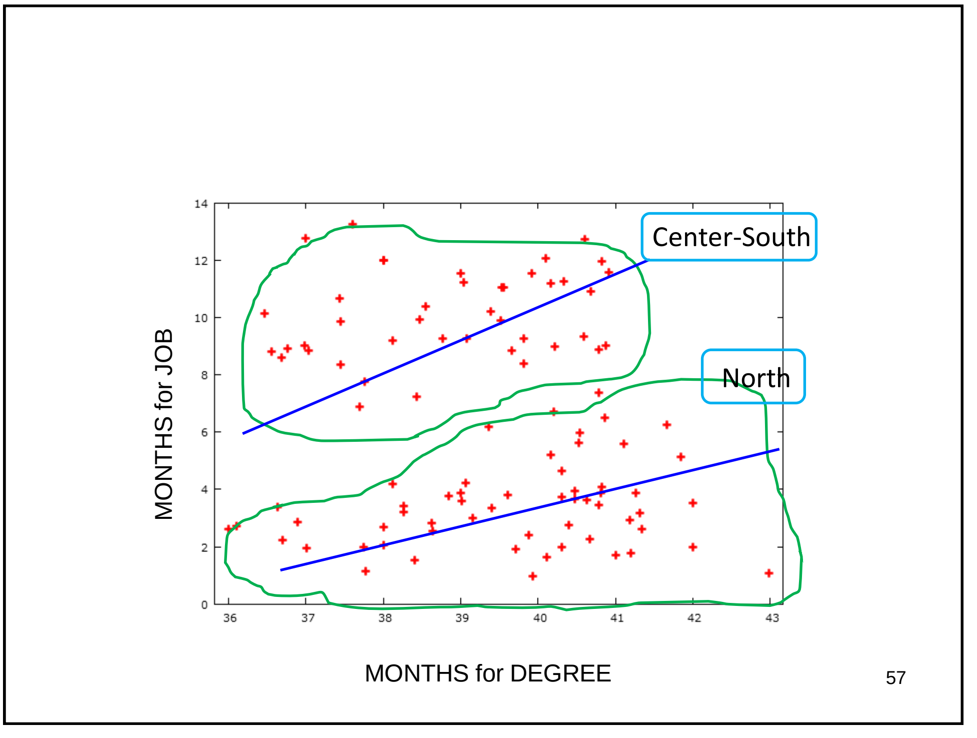
55

55

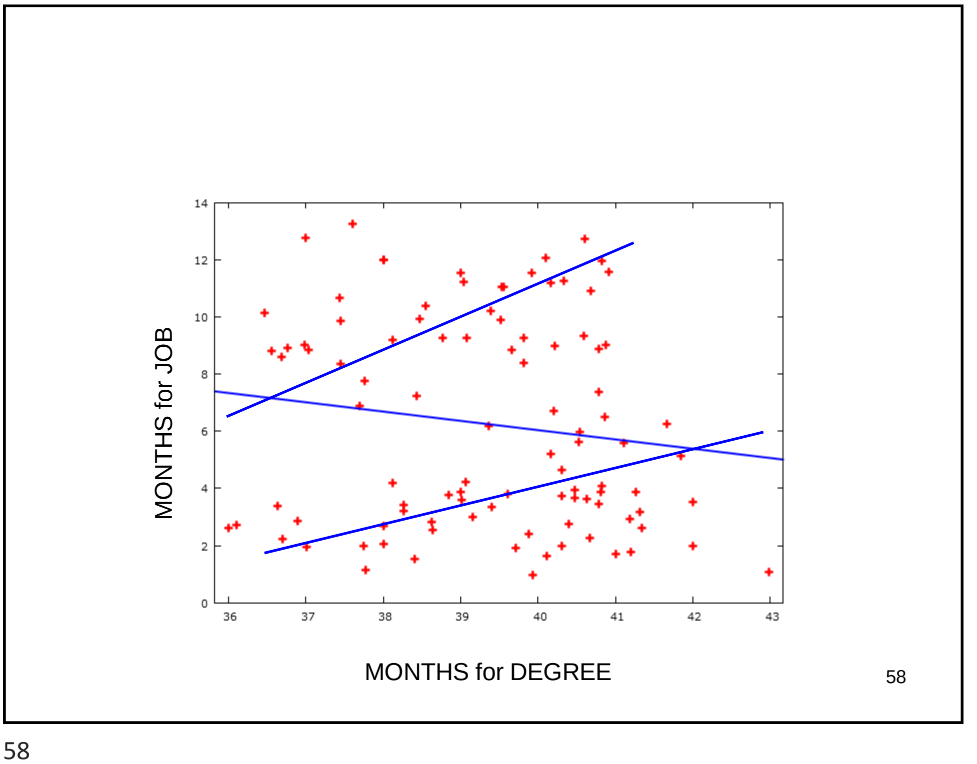


56

56


Center-South
North
MONTHS for JOB
MONTHS for DEGREE
57

57


MONTHS for JOB
MONTHS for DEGREE
58

58

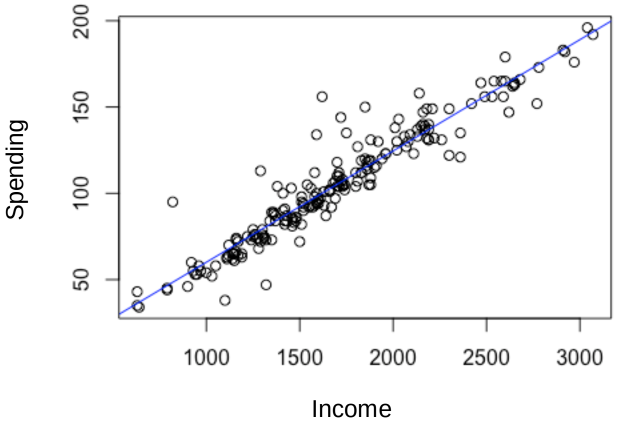
Coplot

- Coplot shows the relationship between two numerical variables, conditionally on the value (category) assumed by a third variable.
- Consider spending on insurance services as a function of income.

59

59

Example



60

60

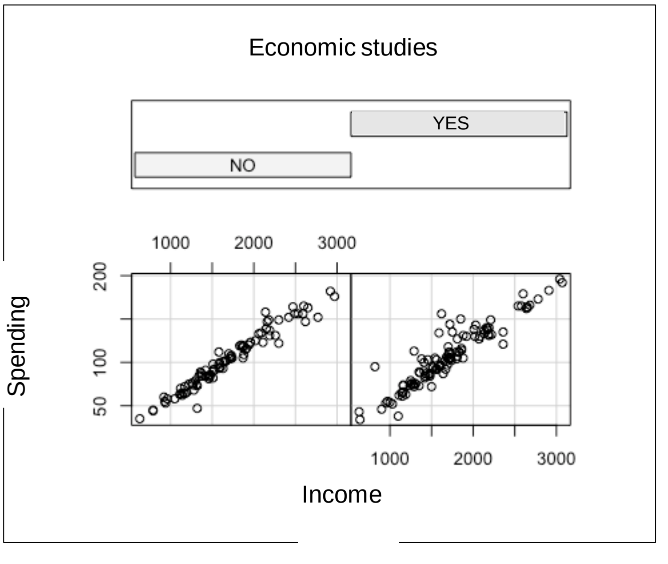
Coplot

- Let us introduce a third (categorical) variable: Economic Studies («YES», «NO»)

61

61

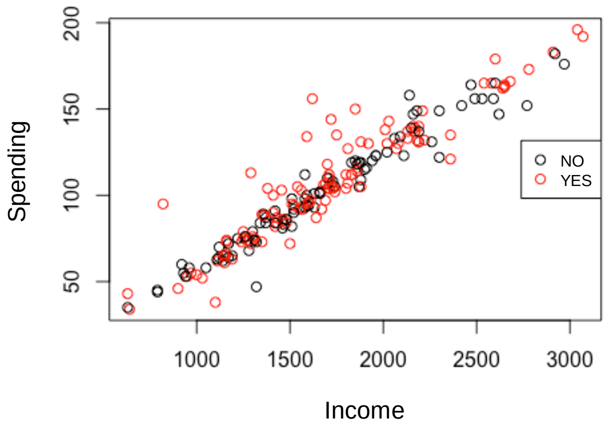
Coplot (1° version)



62

62

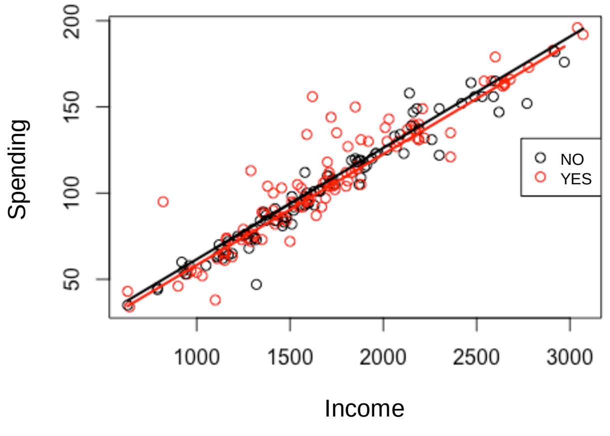
Coplot (2° version)



63

63

…with two regression lines



64

64

Caution in analyzing association (2)

- A lurking variable is a variable not measured in a study that influences the association between two variables.
- Example: the positive relationship between ice cream sales and the number of drowned is apparent because the temperature is a lurking variable.

65

65

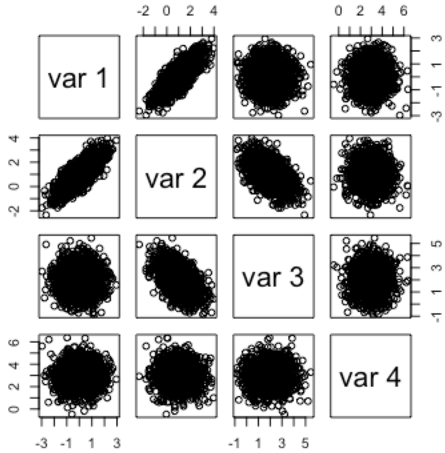
Scatterplot matrix

- A scatterplot matrix highlights the two-by-two relationships between *p* variables.
- It is constituted in the form of a matrix (*p* x *p*).
- The scatterplots are plotted on the cells above and below the main diagonal.

66

66

Scatterplot matrix (*p*=4)

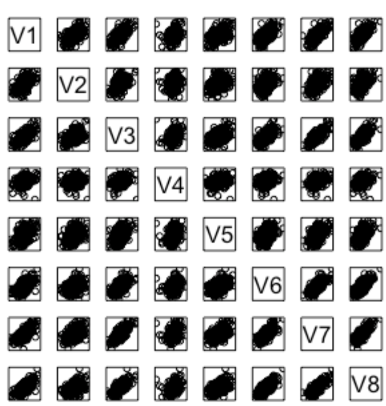


We detect:
-concordance between var1 and var2;
-no association between var1 and var3;
- no association between var1 and var4;
-discordance between var2 and var3;
-no association between var2 and var4;
-no association between var3 and var4.

67

67

Scatterplot matrix (*p*=8)



68

68

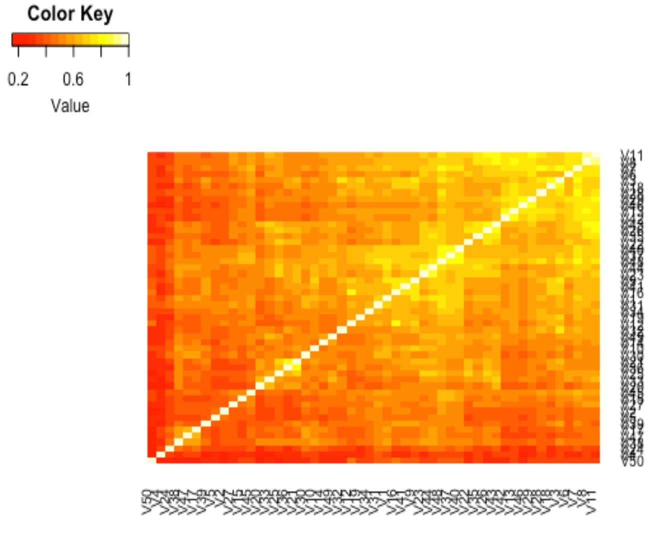
Heatmap

- As the data size increases dramatically, a new graph, called heatmap, can effectively replace the scatterplot matrix.
- The heatmap is a mosaic of different colors associated with different degrees of correlation.

69

69

Heatmap (p=50)



70

70