

STATISTICAL LABORATORY

Lecture 3: Exploratory Data Analysis (Practice)

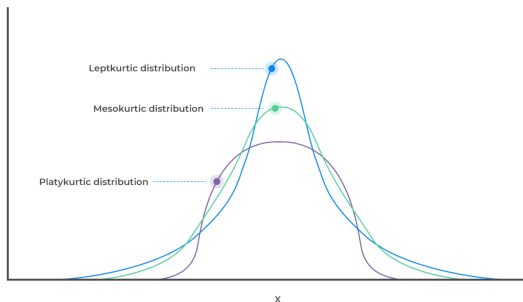
Yiu Lim LUI, Allen

Lisbon School of Economics & Management
Universidade de Lisboa

`allenlui@iseg.ulisboa.pt`

Applied Mathematics for Economics and Management
1st Year / 1st Semester – 2026/2027

Kurtosis Graph



- **Leptokurtic Distribution** (high peak, heavy tails): Most data concentrated near the mean. More extreme values (outliers) likely.
- **Platykurtic Distribution** (low peak, light tails): Data more evenly spread. Fewer extreme values.
- **Mesokurtic Distribution** (moderate peak, normal tails): Moderate concentration of data around the mean.

Exercise 2.1

For quality-control purposes, the product is shipped in boxes of six units. A simple random sample without replacement of 40 boxes is drawn from the lot. For each box, the number of defective units is recorded. The observed sample is:

2, 0, 1, 3, 0, 3, 6, 4, 5, 3, 0, 1, 0, 3, 2, 0, 1, 0, 2, 3, 1, 0, 3, 0, 1, 0, 0, 1, 5, 6, 0, 2, 5, 1, 0, 3, 0, 2, 3, 0

- a) Define the statistical unit and the variable under analysis, and classify the variable.
- b) Present the frequency table.
- c) Provide a graphical representation of the data.
- d) Compute the range and the interquartile range (IQR).
- e) Assess the symmetry (skewness) of the distribution.

Exercise 2.1 (a) (Answer) (last week)

- a) Statistical Unit: Box of 6 units; Variable: number of defective products per box; Classification of the variable: Quantitative discrete variable (since it takes numerical integer values representing counts).

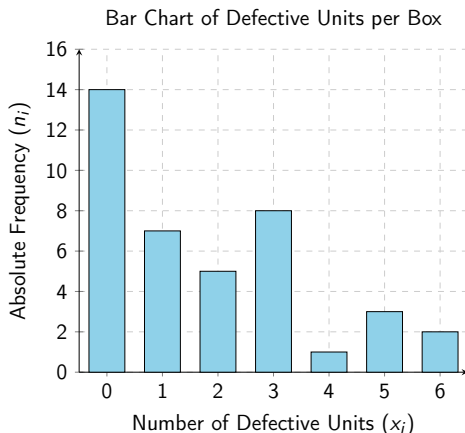
Exercise 2.1 (b) (Answer) (last week)

Let x_i represent the number of defective units, n_i the absolute frequency, f_i the relative frequency, N_i the cumulative absolute frequency, and F_i the cumulative relative frequency.

Table: Frequency Distribution of Defective Units per Box

x_i	n_i	f_i	N_i	F_i
0	14	0.350	14	0.350
1	7	0.175	21	0.525
2	5	0.125	26	0.650
3	8	0.200	34	0.850
4	1	0.025	35	0.875
5	3	0.075	38	0.950
6	2	0.050	40	1.000
Total	40	1.000	–	–

Exercise 2.1 (c) (Answer) (last week)



For discrete variables such as $\{0, 1, 2, \dots, 6\}$, both bar chart and histogram are acceptable. **But to be more precise, it is a bar chart, since there are gaps between bars.**

Exercise 2.1 (d) (Answer)

- Range = Max - Min = 6 - 0 = 6. Domain (D) = [0, 6].
- Interquartile Range (IQR) = $Q_3 - Q_1$.

$$(n + 1) \times 0.75 = 41 \times 0.75 = 30 + 0.75 \quad (\text{For } Q_3)$$

$$Q_3 = 0.25x_{(30)} + 0.75x_{(31)} = 3.$$

Note: We can check the frequency table to know the value of $x_{(30)}$ and $x_{(31)}$.

$$(n + 1) \times 0.25 = 41 \times 0.25 = 10 + 0.25 \quad (\text{For } Q_1)$$

$$Q_1 = 0.75x_{(10)} + 0.25x_{(11)} = 0.$$

Exercise 2.1 (e) (Answer)

- Mode: 0
- Median: 1
- Mean: 1.80

Mean $>$ Median $>$ Mode. This implies positive skewness or right skewness.

Exercise 2.2

The following table shows the number of weekly study hours outside classes observed in a sample of MAEG students:

6.7	8.2	8.5	5.7	9	8	18.2	9.5	5.6	11.2	9.8	3.1	7.6	8.8	6.3
-----	-----	-----	-----	---	---	------	-----	-----	------	-----	-----	-----	-----	-----

- a) Define the statistical unit and the variable under study.
- b) Compute the five-number summary for this dataset.
- c) Analyse the symmetry of the data using a boxplot.

Exercise 2.2 (Suggested Answer)

- 1 Statistical Unit: student; Variable: weekly study hours outside classes.
- 2 Min: 3.1, First Quartile Q_1 : 6.3, Median Q_2 : 8.2, Third Quartile Q_3 : 9.5, Max: 18.2
- 3 Positive skewness.

Exercise 2.3

The statistical information on the monthly wages of workers in a textile company is presented below:[cite: 3]

Salary (euros)	Workers (%)
[600 ; 900]	9
]900 ; 1000]	18
]1000 ; 1100]	35
]1100 ; 1200]	22
]1200 ; 1300]	16

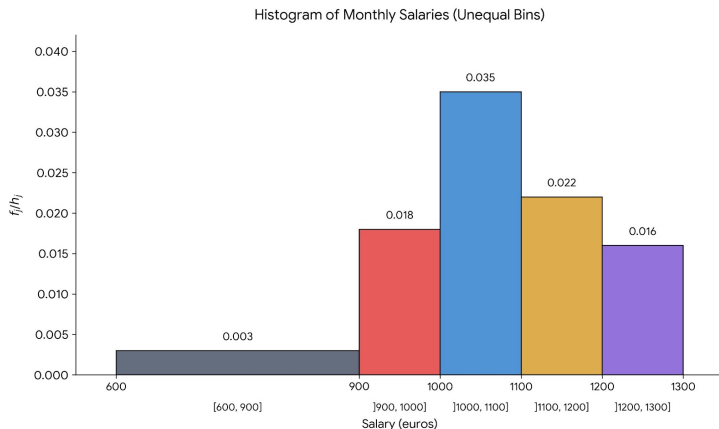
- a) Define and classify the variable under study.
- b) Provide a graphical representation of the data.
- c) Assess the symmetry of the data based on the graphical representation.

Exercise 2.3 (Answer)

- a) Continuous variable: salary, in euros;
- b) When class intervals have unequal widths (like in this exercise, where the first class has a width of 300 and the rest have a width of 100), you cannot just use the raw frequencies as the heights of the bars in a histogram. You divide the frequency by the width (f_j/h_j). This ensures that the **area** of each bar (Height $\times$ Width) is what actually represents the **frequency**, making comparisons fair and accurate across different bin sizes.

Exercise 2.3 (Answer)

So we have the following histogram:



c) Negative skewness.

Exercise 2.4

Consider the grades of 50 students in this course from the previous academic year:

93, 77, 67, 72, 52, 83, 66, 84, 59, 63,
75, 97, 84, 73, 81, 42, 61, 51, 91, 87,
34, 54, 71, 47, 79, 70, 65, 57, 90, 83,
58, 69, 82, 76, 71, 60, 38, 81, 74, 69,
68, 76, 85, 58, 45, 73, 75, 42, 93, 65

- a) Represent the data in an ordered stem-and-leaf plot.
- b) Compute the five-number summary from the plot.
- c) Construct a boxplot and assess the symmetry (or asymmetry) of the data.
- d) Group the data into 5 classes of equal width, with the lower bound of the first class set to 25 and construct a histogram.

Exercise 2.4 (Answer)

a)

		n_i
3	4 8	2
4	2 2 5 7	4
5	1 2 4 7 8 8 9	7
6	0 1 3 5 5 6 7 8 9 9	10
7	0 1 1 2 3 3 4 5 5 6 6 7 9	13
8	1 1 2 3 3 4 4 5 7	9
9	0 1 3 3 7	5

Exercise 2.4 (Answer)

- b) Min: 34, First Quartile Q_1 : 58.75, Median Q_2 : 71, Third Quartile Q_3 : 81.25, Max: 97
- c) Slight negative skewness (Slightly longer Lower Whisker);

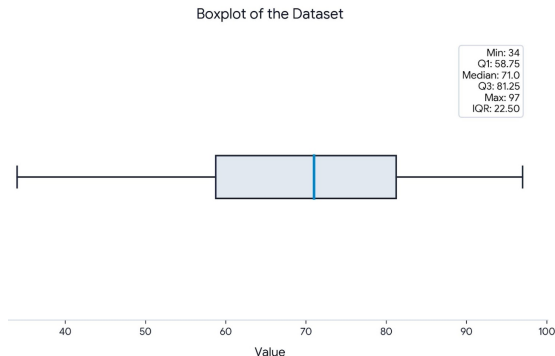
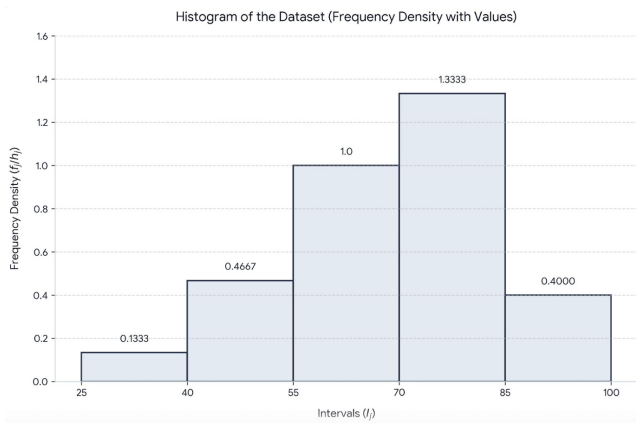


Figure: Box-plot with Min/Max Method

Exercise 2.4 (Answer)

d)



Exercise 2.5

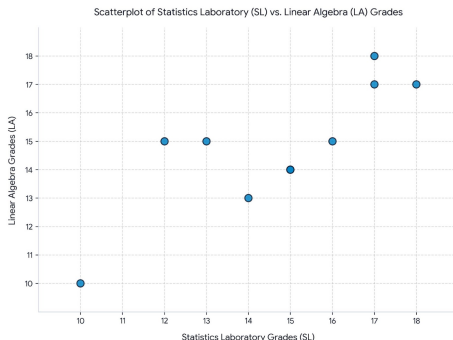
The grades of 10 students in Statistics Laboratory (SL) and Linear Algebra (LA) are presented below:

SL	15	12	18	14	13	10	15	16	17	17
LA	14	15	17	13	15	10	14	15	18	17

- a) Provide a graphical representation of these data.
- b) Assess the possibility of a linear relationship between the grades in these two courses.

Exercise 2.5 (Answer)

a) Scatterplot



b) Relatively strong positive linear relationship.

Exercise 2.6

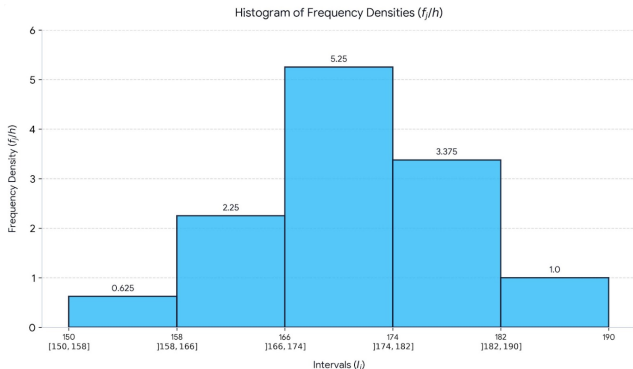
The statistical information on the heights of students in a school is presented below:

Height (cm)	Number of Students
[150 ; 158]	5
]158 ; 166]	18
]166 ; 174]	42
]174 ; 182]	27
]182 ; 190]	8

- a) Identify the statistical unit and the variable, and classify the variable.
- b) Briefly assess the data based on its graphical representation.

Exercise 2.6 (Answer)

- a) Statistical unit: student ; Continuous variable: height, in cm;
b)



Negatively skewed (more data concentrated on the right.)