FOUNDATION YEAR 2025-2026

COURSE - DATA ANALYSIS

Lesson 1

Niccolò Maffezzoli, Ca' Foscari University of Venice

January 12, 2026

Course information

Where and when

- Monday 8.45-10.15 - room B1, Palazzina Briati, Fondamenta Briati, Venice
- Thursday 8.45-10.15 - room 0L, Tesa 4 Polo didattico San Basilio, Venice
- My office hours for questions/doubts/anything: Monday 15-16. niccolo.maffezzoli@unive.it. Skype: nicomaffe

Coding

- Programming language: python
- We will use colab. Free, but gmail account necessary. Please make one.
- Internet is the best source of material for coding.

Statistics optional textbook

Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences by R. J. Barlow. Link book.

Shared drive with material

Link to course materials

Course Grading

There will be:

1. Mid-Term Exam
2. Final Exam
3. Coding project + report

Note: **Attendance** is mandatory for 50% of each individual course. Students who do not reach 50% attendance will automatically receive a failing grade.

Grades are provided in “trentesimi” (scale 0 to 30). The student’s final grade will consider all evaluations along with active participation in class. Passing grade: 18/30.

Descriptive statistics

One of the best ways to describe a set of measurements of the same quantity

Arithmetic mean

Given a set of measurements $\{x_1, x_2, x_3, ..., x_n\}$,

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

Example 1

You measure the mass of a ball 5 times: m = {2.3, 2.1, 1.8, 1.9, 2.0}. You may extract a meaningful quantitative value to describe the data.

$$\mu_m = \frac{1}{5}\sum_{i=1}^{5} m_i = \frac{1}{5}(m_1+m_2+m_3+m_4+m_5) = \frac{1}{5}(2.3+2.1+1.8+1.9+2.0) = 2.02kg$$

The mean

Example 2

In a survey of 100 cars passing a checkpoint, 72 contained only 1 occupant, 23 had 2, 2 had 3, 3 had 4. The mean number of occupants per car:

$$\mu = \frac{1}{100}(72 \cdot 1 + 23 \cdot 2 + 2 \cdot 3 + 3 \cdot 4) = 1.36$$

Other descriptive statistics

- Geometric mean: $\sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$
- Median: halfway point. Half of the values lie below, half lie above. For example, in the data set $\{1, 3, 5, 6, 9\}$, the median is 5 because it is the middle value. In the data set $\{2, 4, 6, 8\}$, the median is $(4 + 6) / 2 = 5$
- Mode: the most frequent value.

Measuring the spread: the variance

The variance

The variance V (also found as σ^2) expresses the dispersion/spread of the data around the mean. Given a set $\{x_1, x_2, x_3, ..., x_n\}$,

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

$$V = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{3}$$

Exercise

Consider the following two sets of measurements:

$$set_1 = \{2.01, 2.04, 2.02, 2.03, 2.07\}$$

$$set_2 = \{0.75, 1.24, 2.59, 3.41, 2.18\}$$

Calculate the mean and the variance of these two sets.

Exercise

Consider the following two sets of measurements:

$$set_1 = \{2.01, 2.04, 2.02, 2.03, 2.07\}$$

$$set_2 = \{0.75, 1.24, 2.59, 3.41, 2.18\}$$

Calculate the mean and the variance of these two sets.

$$\mu_1 = 2.034; \sigma_1^2 = 0.02059$$

$$\mu_2 = 2.034; \sigma_2^2 = 0.94899$$

Although the means are the same, the measurements of set_2 are more dispersed. This is reflected by $\sigma_2^2 > \sigma_1^2$.

Measuring the spread: the standard deviation

The standard deviation

The standard deviation σ (a.k.a. width or Root Mean Squared, RMS) is given by:

$$\sigma = \sqrt{V} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \tag{4}$$

The standard deviation

This formula gives the square root of the average of the squared differences between each observation and the mean. It is a measure of the amount of variation or spread in the data set, and is commonly used in statistical analysis to describe the distribution of data. It has the same unit as the x_i.

Weighted mean

Weighted mean

$$\mu_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i},$$

where $\overline{\mu}_w$ is the weighted mean, x_i is the i-th value in the set, w_i is the weight assigned to the i-th value, and n is the total number of values in the set.

When dealing with errors

If x_i are affected by errors σ_i, i.e. we have a set $\{x_i \pm \sigma_i\}$, we can compute a weighted mean:

$$\mu_w = \frac{\sum_{i=1}^{n} x_i/\sigma_i^2}{\sum_{i=1}^{n} 1/\sigma_i^2},$$

The weights are $w_i = 1/\sigma_i^2$. More precise measurement are 'weighted' more.