COMP229: Introduction to Data Science
Lecture 2: descriptive statistics

Olga Anosova, O.Anosova@liverpool.ac.uk
Autumn 2023, Computer Science department
University of Liverpool, United Kingdom

Plan for the next sessions

Gentle start with descriptive statistics.

Formalize the sample-based intuition into probability theory.

Progress to developed statistical apparatus of inferential statistics (distributions, hypotheses).

Add geometric tools to the analysis.

Move to high-dimensional vector data via linear algebra.

Apply obtained knowledge to do clustering, PCA and SVD.

Notations

$\mathbb{Z}$ denotes all **integers**,

$\mathbb{R}$ denotes all **real** numbers (e.g. 0, 1, -0.42, $\frac{1}{3}$, $-\sqrt{3}$, π),

$A = \{a, b, c\}$ is a **set** (of 3 elements),

$a \in A$ means that the element **belongs to** the set A,

$B = \{b, c\} \subset A$ means that B is a **subset** of A,

$$\sum_{i=1}^{n} a_i = a_1 + \cdots + a_n.$$

Focus on logic, not on methods, because the same problem can be solved by many methods.

Descriptive vs inferential statistics

Descriptive statistics quantitatively summarises features of *sample* data by numbers, diagrams.

This approach differs from **inferential statistics** that aims to learn about the whole *population* (all objects or subjects) from a smaller sample.

This class can be viewed as a sample of all students studying in the UK. If we only find the average age in the class, it's a simple description of the sample.

If we *infer* (make a conclusion about) the average age of all UK students, it's an *inference* problem.

The sample mean of scalar data

Definition 2.1. The sample **mean** (or the *arithmetic average*) of n values $a_1, \ldots, a_n \in \mathbb{R}$ is defined as $\bar{a} = \dfrac{1}{n}\sum_{i=1}^{n} a_i = \dfrac{a_1 + a_2 + \cdots + a_n}{n}$.

Problem 2.2. What's the average temperature in Liverpool? Average temperature by months is given by the set $A = \{5, 6, 7, 9, 12, 15, 17, 16, 14, 11, 8, 6\}$. The data A above has the sample mean $\bar{a} =$

$$= \frac{1}{12}(5+6+7+9+12+15+17+16+14+11+8+6) = 10.5$$

Weighted average

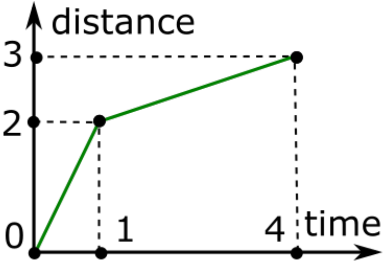
The formula $\bar{a} = \frac{1}{n}\sum_{i=1}^{n} a_i$ is simplified when data values are repeated (weighted). Let a sample have distinct values $b_1, \ldots, b_m$ that appear $k_1, \ldots, k_m$ times, respectively. The full sample has $\sum_{j=1}^{m} k_j$ values in total: k_j repeated values of b_j, $j = 1, \ldots, m$. Then $\bar{a} = \frac{1}{n}\sum_{i=1}^{n} a_i$ becomes $\bar{a} = \frac{1}{n}\sum_{j=1}^{m} k_j b_j$.

Why do I need to know the formula?

Problem 2.3. A corporation reports that its policies are democratically controlled by its 50 stockholders as they have 600 votes in total with an average of 12 votes per person. Evaluate this corporation's claim.

Solution 2.3. What if we knew that 5 stockholders have 84 votes each? Those 5 important stockholders have accumulated $84 \times 5 = 420$ votes and have compete control of the corporation. The average is indeed $\dfrac{84 \times 5 + 4 \times 45}{50} = 12$ votes.

Average speed



Problem 2.3. Find the average speed from the graph describing how the speed depends on time.

Solution 2.3. From the picture, $v_1 = 2/1 = 2$ (unitless here for simplicity) and $v_2 = \frac{3-2}{4-1} = \frac{1}{3}$.

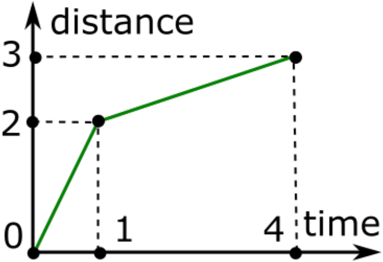
So the the average speed is $\frac{v_1+v_2}{2} = \frac{1}{2}\left(2 + \frac{1}{3}\right) = \frac{7}{6}$.

By another definition the average speed v is

$\frac{\text{total distance}}{\text{total time}} = \frac{3}{4} \neq \frac{v_1+v_2}{2} = \frac{7}{6}$.

Which is correct?

Average speed



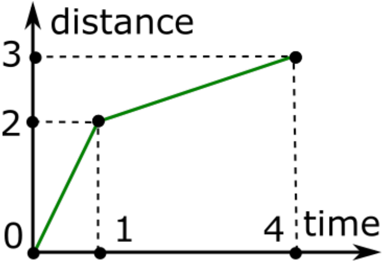
Take into account that the speed $v_1 = 2$ was maintained for $t \in [0,1]$ (time ratio $\frac{1}{4}$), while $v_2 = \dfrac{1}{3}$ for $t \in [1,4]$ (time ratio $\frac{3}{4}$).

Then $v = \dfrac{1}{4}v_1 + \dfrac{3}{4}v_2 = \dfrac{1}{4} \times 2 + \dfrac{3}{4} \times \dfrac{1}{3} = \dfrac{3}{4}$.

Different methods can give the same correct result!

Library functions can mislead

You should be able to program yourself how to compute the average of a data sample, because library functions often output different results.



If measuring as a time series, e.g. at regular times $t = 0, 2, 4$ with $v = 0, \frac{7}{3}, 3$, then the brute force application gives:

$\frac{1}{2}\left(\left(\frac{7}{3} - 0\right) + \left(3 - \frac{7}{3}\right)\right) = \frac{3}{2}$, another wrong answer!

Mean as a good descriptor

Claim 2.4. Let the mean $\bar{a} = \dfrac{1}{n}\sum_{i=1}^{n} a_i > 0$ be positive. Then all values $a_i > 0$.

For example, $a_1 = 1$, $a_2 = 3$ give $\bar{a} = 2 > 0$.

For example, the Liverpool temperatures example $\bar{a} = 10.5 > 0$ and all $a_i > 0$.

Examples prove nothing

Examples do not prove the general case as counter-examples can also exist, e.g. $a_1 = -1$, $a_2 = 3$ have $\bar{a} = 1 > 0$.

What if we have tried hard and couldn't find any counter-examples (data values $a_1, \ldots, a_n$ with some $a_i < 0$ and a mean $\bar{a} > 0$), what does it imply?

The average temperature in Liverpool is 10.5°. Will there be no ice?

(Counter-)examples disprove claims

If we couldn't find counter-examples, we can't call a claim correct or wrong, it remains a conjecture.

If we found a counter-example, then the claim is certainly disproved. Often it's easy to find some counter-examples, but it's hard to prove claims.

Claim 2.4 stating if $\bar{a} > 0$, then all $a_i > 0$ has been *disproved* by a counter-example $a_1 = -1$, $a_2 = 3$ with $\bar{a} = 1 > 0$.

Let all $a_i > 0$. Does it imply that $\bar{a} > 0$?

Why proove?

Because we can!

And it's one of the few things left that gives us a real advantage compared to AI:

The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A".

Is it enought to add several rules of logic?

Counting ravens

Claim 1: *All ravens are black.*

By *contraposition*

$(A \Rightarrow B) \iff (\neg B \Rightarrow \neg A)$

this claim is equivalent to

Claim 2: *If we see a non-black thing, then it is not a raven.*

Scenario 1: If we see a black raven, this increases the likelihood of Claim 1.

Scenario 2: If we see a green apple, this increases the likelihood of Claim 2, and hence of Claim 1.

Raven paradox explanation

Number of all ravens is **finite** and can possibly be observed.
The number of all other things is **infinite**, so they can not be all checked.

It's not enough to master several logic rules, it's important to master combining them into something meaningful!

Other more sophisticated ways to explain this paradox are also possible, see "Raven paradox".

Proof example

Claim 2.5. If all $a_1, \ldots, a_n > 0$, then $\bar{a} > 0$.

Proof. The sum of the given inequalities $a_1 > 0, a_2 > 0, \ldots, a_n > 0$ gives $\sum_{i=1}^{n} a_i > 0$, then we can divide by $n > 0$ and get $\bar{a} = \frac{1}{n} \sum_{i=1}^{n} a_i > 0$. □

Similarly: if the average exam mark is 60, it doesn't imply that all students have passed.

What should be an average mark within [0,100] to guarantee all passes? Come to tutorial 1 to discuss.

The range of a data sample

The average doesn't guarantee any bounds: 1000 & -1000 have the same average as 1 & -1, so we need another descriptor.

Definition 2.6. The **range** of scalar data is the interval from the minimum to the maximum value.

Example 2.7. What's the range of the Liverpool monthly temperatures
$A = \{5, 6, 7, 9, 12, 15, 17, 16, 14, 11, 8, 6\}$?

It might help to put the values in the increasing order from the minimum to the maximum value.

The mid-point of the range

$A = \{5, 6, 7, 9, 12, 15, 17, 16, 14, 11, 8, 6\}$ can be ordered as $\{5, 6, 6, 7, 8, 9, 11, 12, 14, 15, 16, 17\}$ and has the range $[5, 17]$.

The minimum value of A is $\min\limits_{i=1,\ldots,n} a_i = 5$.
The maximum value of A is $\max\limits_{i=1,\ldots,n} a_i = 17$.

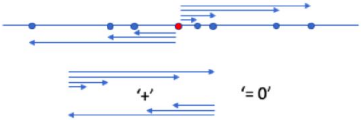
Problem 2.8. Is the mid-point of the range $\dfrac{1}{2}\left(\min\limits_{i=1,\ldots,n} a_i + \max\limits_{i=1,\ldots,n} a_i\right)$ always equal to the sample mean $\bar{a}$? If not, are they always different?

Hint: the mid-point of the range $[5, 17]$ is 11.

Measuring variation

How widely are the value distributed around the mean sample $\bar{a}$? What is the amount of variation (or dispersion) of a set of data values?

What if we just find the average of all deviations from the mean?

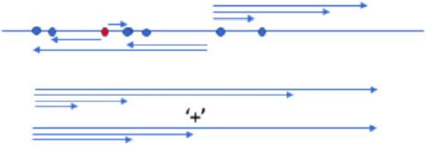


$$\sum_{i=1}^{n}(a_i - \bar{a}) = \sum_{i=1}^{n} a_i - n\bar{a} =$$

$$\sum_{i=1}^{n} a_i - n\sum_{j=1}^{n}\frac{a_j}{n} =$$

$$\sum_{i=1}^{n} a_i - \sum_{j=1}^{n} a_j = 0, \text{ no good.}$$

The sample standard deviation



We need to square the differences $a_i - \bar{a}$.

Definition 2.9. A sample of n values $a_1, \ldots, a_n$ with a mean $\bar{a}$ has the **sample variance** defined as $Var = s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (a_i - \bar{a})^2$ and the (unbiased) **sample standard deviation**

$$s = \sqrt{Var} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (a_i - \bar{a})^2}.$$

The factor $\frac{1}{n-1}$ instead of $\frac{1}{n}$ is Bessel’s correction.

The partial case $n = 2$

For two values $a_1 \leqslant a_2$, the mean is $\bar{a} = \dfrac{a_1 + a_2}{2}$, the sample deviation $s = \sqrt{(a_1 - \bar{a})^2 + (a_2 - \bar{a})^2} =$ $\sqrt{\left(\dfrac{a_1 - a_2}{2}\right)^2 + \left(\dfrac{a_2 - a_1}{2}\right)^2} =$ $\sqrt{\dfrac{2(a_1^2 - 2a_1a_2 + a_2^2)}{4}} == \sqrt{2}\dfrac{a_2 - a_1}{2}$ is the half-distance between a_1 and a_2 multiplied by $\sqrt{2}$.

The partial case $n = 2$ continued

Claim 2.11. The sample of two values a_1, a_2 is uniquely determined by the mean $\bar{a}$, deviation s.

Proof. Without loss of generality we can assume that $a_1 \leqslant a_2$ Then the mean and the deviation provide 2 linear equations with 2 unknowns a_1, a_2:

$$\begin{cases} \bar{a} = \dfrac{a_1 + a_2}{2} \\ s = \dfrac{a_2 - a_1}{\sqrt{2}} \end{cases}$$

with the unique solution

$a_1 = \bar{a} - \dfrac{s}{\sqrt{2}}$ and $a_1 = \bar{a} + \dfrac{s}{\sqrt{2}}$. □

Liverpool temperature deviations

Problem 2.12. Find the sample deviation s of monthly temperatures in Liverpool $A = \{5, 6, 7, 9, 12, 15, 17, 16, 14, 11, 8, 6\}$.

Solution 2.12. A set $\{5, 6, 6, 7, 8, 9, 11, 12, 14, 15, 16, 17\}$ has $\sum_{i=1}^{n} (a_i - \bar{a})^2 =$

$$(5-10.5)^2+2(6-10.5)^2+(7-10.5)^2+(8-10.5)^2+(9-10.5)^2+(11-10.5)^2$$

$$+(12-10.5)^2+(14-10.5)^2+(15-10.5)^2+(16-10.5)^2+(17-10.5)^2=199,$$

and $s = \sqrt{\dfrac{199}{11}} \approx 4.25.$