

Next week: Instructions for and
examples of presentations

After that: See timetable on webpage

If you still need to sign up for a presentation,
do so by the end of Thursday 5th May

You need to select Friday 27th May.

Statistics

Statistic is about the numerical analysis of data

How do we collect data?

E.g., surveys often use questionnaire

Example questionnaire

A survey of residents of EU countries
of opinions on the EU.

What is your opinion
on the EU?

very
negative | -ve | mildly
-ve | neutral | mildly
+ve | +ve | very
+ve

Country of residence

Nationality

Age (yrs)

Annual income (€s)

The questionnaire illustrates several points

- How do we gather data?

We are interested in a huge population (all EU residents)

Impossible to survey the whole population

Instead survey a sample of the population

We want a sample that is :

- representative of the population
(avoiding bias)

Achieve this by choosing the sample randomly

- Large enough that we can trust the result
- Not too large : we want running the survey and processing the results to be feasible and economic.

The survey illustrates the main kinds of data in statistics

Numerical data (either a real number or integer)

e.g., age, income

Numerical data is either discrete (an integer) or continuous (a real number)

Categorical data Only a few fixed possible values.

e.g., country of residence, nationality, opinion of EU

Divide these into:

- Nominal data no ordering on data
e.g. countries
- Ordinal data there is an ordering; e.g. opinion of EU

Summary statistics are values that capture properties of the data

- Averages : summarize the typical value / location / central tendency of the data.
- Measures of spread or dispersion, which measure the variation in the data.

Averages Suppose we have N items of data X_1, \dots, X_N

Mean (arithmetic mean)

$$\left(\bar{x}\right) \mu := \frac{X_1 + \dots + X_N}{N} = \frac{1}{N} \sum_{i=1}^N X_i$$

Only makes sense for numerical data

Median If $X_1, X_2, X_3, \dots, X_N$ are written in ascending order then

$$\text{Median} = \begin{cases} X_{\frac{N+1}{2}} & \text{if } N \text{ is odd} \\ \text{any value between } X_{\frac{N}{2}} \text{ and } X_{\frac{N}{2}+1} & \text{if } N \text{ is even} \end{cases}$$

Makes sense for numerical and ordinal data

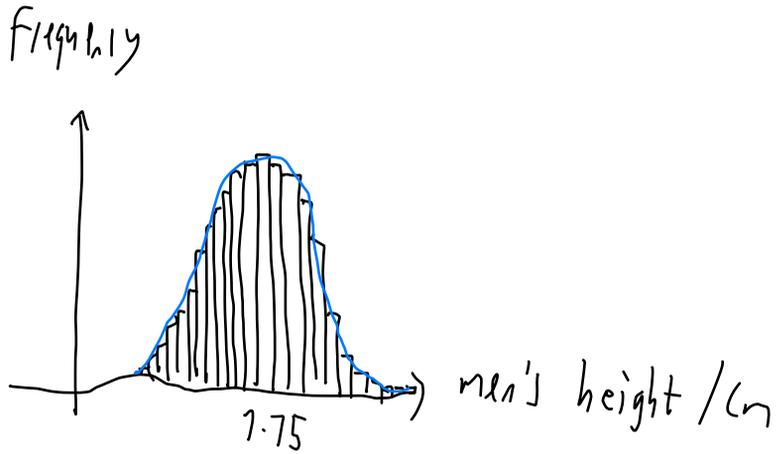
Mode The data value or values that occurs most often in X_1, \dots, X_N .

Makes sense for (discrete) numerical, ordinal and categorical data.

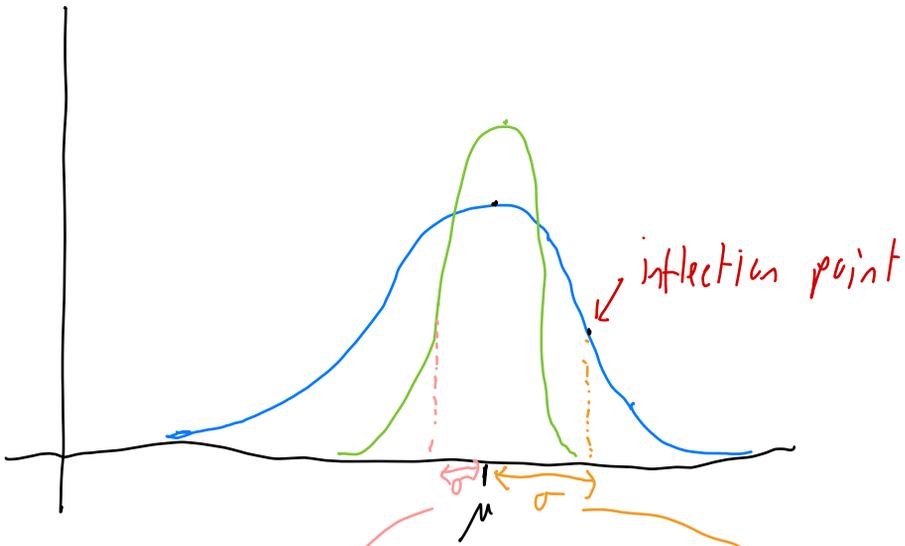
Measures of spread

Often we can visualise the distribution of data as a histogram (or bar chart)

e.g.



When we abstract a graph from such data we very often see a bell-shaped curve (or Gaussian curve)



s.d. of the green curve

s.d. of the blue curve

Data that exhibits this Gaussian bell shape is called normally distributed (it has a Gaussian distribution)

with mean μ and

standard deviation σ

The notion of standard deviation

makes sense for arbitrary numerical data

It is defined by

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

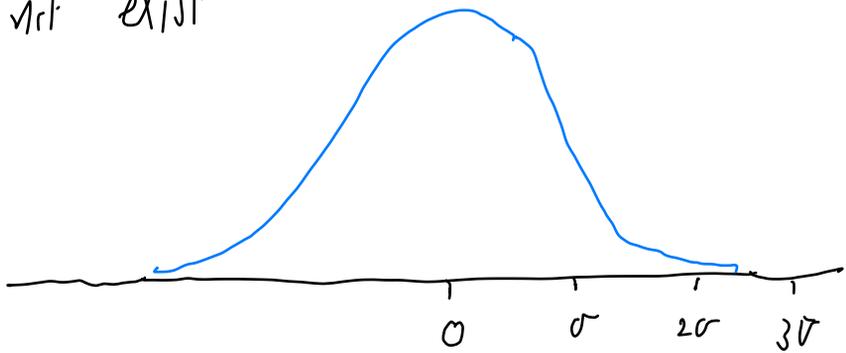
σ^2 is called the variance

Application of S.D. (Standard deviation)

(Watch video linked on course page).

In July 2012 physicists at CERN confirmed the discovery of the Higgs particle with a certainty of 5 σ

If the Higgs particle did not exist



Observed result $> 5\sigma$

If the Higgs particle does not exist then the probability of getting the observed result < 1 in 3.5 million.