Statistical reasoning for humanities scholars

Lecture I

Nalanda Academy Workshop 17-19 May 2025

Subodh Patil and Kapil Wankhede

What is data?

- Data (pl.) is a collection of/ strings of information.
- Numbers, text characters, words, names, etc…
- In small amounts, data is easy and intuitive to understand and process `by hand' or `by eye'.

What is data?

- Data (pl.) is a collection of/ strings of information.
- Numbers, text characters, words, names, etc…
- In large amounts, one can quickly be overwhelmed by too much information (cf. RTI requests!)

Statistics is the study and analysis of large data sets.

- **It is a *science* in that it involves the use of statistical methods and modelling.**
- **It is an *art* in that involves knowing what questions to ask and what to ignore.**

Data needs to be processed (cleaned) and interpreted before it can be used!

e.g. Imaginary height and weight data of all Std. X children in Wardha taluka:

Sex	Height in cm	Weight in kg
…	…	…
M	@40.997	40.52
F	203.296	50.15
F	150.167	45.29
M	160.002	60.10
F	#45.106	@0.23
…	…	…

Data can be noisy:

e.g. Imaginary height and weight data of all Std. X children in Wardha taluka:

Sex	Height in cm	Weight in kg
…	…	…
M	@40.997	40.52
F	203.296	50.15
F	150.167	45.29
M	160.002	60.10
F	#45.106	@0.23
…	…	…

?

One can `clean' some of the noise with reasonable `priors':

e.g. Imaginary height and weight data of all Std. X children in Wardha taluka:

Sex	Height in cm	Weight in kg
…	…	…
M	140.997	40.52
F	203.296	50.15
F	150.167	45.29
M	160.002	60.10
F	#45.106	@0.23
…	…	…

But some entries are too corrupted to be useful, need to be discarded:

e.g. Imaginary height and weight data of all Std. X children in Wardha taluka:

Sex	Height in cm	Weight in kg
…	…	…
M	140.997	40.52
F	203.296	50.15
F	150.167	45.29
M	160.002	60.10
F	~~#45.106~~	~~@0.23~~
…	…	…

We also need to think about how the data was acquired, and what its information content is:

Sex	Height in cm	Weight in kg
…	…	…
M	140.997	40.52
F	203.296	50.15
F	150.167	45.29
M	160.002	60.10
F	~~#45.106~~	~~@0.23~~
…	…	…

Are all the numbers present in the data fields to be accorded equal significance?

Sex	Height in cm	Weight in kg
…	…	…
M	140.997	40.52
F	203.296	50.15
F	150.167	45.29
M	160.002	60.10
F	~~#45.106~~	~~@0.23~~
…	…	…

Cleaned data needs to reflect the confidence with which we can assign to it:

(accurate only to $\pm\ 0.1\ cm, \pm 0.1\ kg$)

Sex	Height in cm	Weight in kg
…	…	…
M	141.0	40.5
F	203.3	50.2
F	150.2	45.3
M	160.0	60.1
F	~~#45.106~~	~~@0.23~~
…	…	…

… after which it is ready to be manipulated and interpreted✅

Sex	Height in cm	Weight in kg
…	…	…
M	141.0	40.5
F	203.3	50.2
F	150.2	45.3
M	160.0	60.1
F	~~#45.106~~	~~@0.23~~
…	…	…

You are bombarded with numbers every day.
What do these numbers mean?

“30% chance of rain tomorrow”

You are bombarded with numbers every day. What do these numbers mean?

“NCP polling 32% in Vidhan Sabha elections”

You are bombarded with numbers every day.
What do these numbers mean?

“Covid had a 1.5% case fatality rate in India”

You are bombarded with numbers every day.
What do these numbers mean?

"30% ~~chance of rain tomorrow~~"

... ± ? *of simulations show rain*

You are bombarded with numbers every day.
What do these numbers mean?

"NCP polling 32% in Vidhan Sabha elections"
... *with a `margin of error' of 5%*.

You are bombarded with numbers every day.
What do these numbers mean?

"Covid has a 1.5% **case fatality** rate in India"
*... the true **infection fatality** will be lower...*

You are bombarded with numbers every day.

Numbers mean **nothing** unless you can assign a degree of confidence (uncertainty) to them.

The humanities and social sciences encourage critical thinking skills to help you become an empowered citizen. So does statistical and scientific literacy.

Over the coming lectures, we will give you an overview of the basic concepts and applications of statistics.

It is an invitation to a deeper study of the subject, that requires mastering more advanced mathematical and computational tools, which we are happy to point you towards.

- **Statistics**

  A statistical question.

  Has no single answer.

  But many answers. We are interested in the distribution and tendency of the answers.

  How tall are you? -> Not a statistical question.

  How tall are Indians? -> A statistical question.

- **Statistics**

  Average (Central tendency) – mean, mode, median

  Probability distribution.

  How many Indians are 175 cms (~69 inches) tall?

- **Statistics**

  Average (Central tendency) – mean, mode, median

  Probability distribution.

  How many Indians are 175 cms (~69 inches) tall?

  Not a well defined question!

  A better question: How many Indians are between 174.5 and 175.5 cms tall?

Average (Central tendency) – mean, mode, median

Probability distribution.

How many Indians are between 174.5 and 175.5 cms tall?

interval histogram bin