Innovation and Data Analytics

Yesoda Bhargava

January 20, 2021

What is Data?

- Potential for information/insight/knowledge.

Figure 1: Some examples of types of data

What determines this potential?

- Quality of data.
- Quality of statistical analysis - robustness and the thinking behind it.
- **Theoretical basis for analysis of the data.**
- Assumptions made during the analysis.
- Applicability of findings beyond the sample data.

Digging Deeper...

- What is Quality of data? How do you determine it?
- Why care about the theoretical basis for analysis?
- What is the importance of awareness of the assumptions used for analysis?
- How to determine the quality of my statistical analysis?
- Why is applicability outside of sample important?

Some suggestions..

- Source of data has loads to say about the quality of data. WHO or World Bank data would be more reliable than any news article data which is not referenced or based on peer-reviewed research.
- Always check more about the source of data and see if they are instead propagating some agenda.
- **Theoretical basis of analysis**: Data analysis is most times testing of hypothesis, discovering pathways/mechanisms in data which substantiate theoretical under-pinnings. If the conceptual thoguht is defensible, data analysis would make more sense.
- The analysis has to mean something and not just number crunching , therefore theoretical basis is important.
- For eg. Disease prediction models. You must know about factors that influence the risk of a disease. Know the likely directionality of the relationships.
- If theoretical understanding is unclear the modelling process could be ambiguous and hence, erroneous.
- Real world differs from experimental world, hence, results must be extrapolated in the context of the population from which sample is drawn.
- Before doing any statistical analysis: reason why you are doing that.

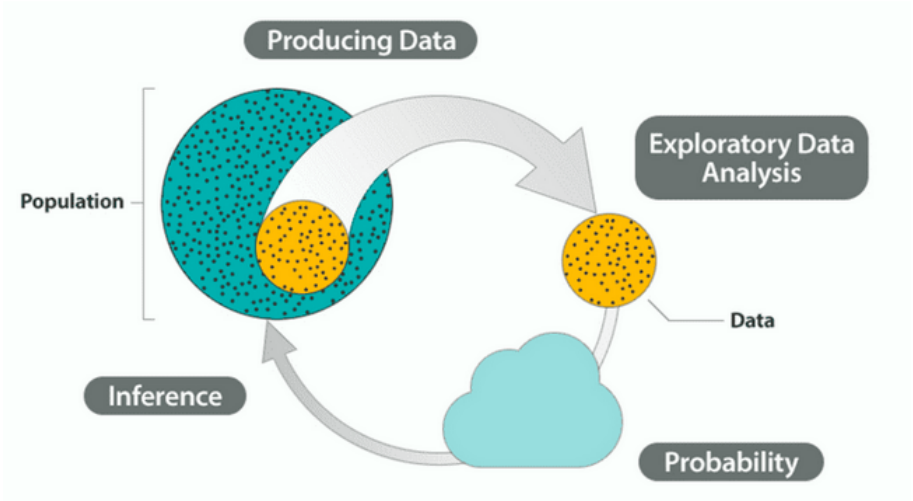

Figure 2: Basics of statistics.

Type of data format

- Structured : Survey data, numerical data, Electronic health record data, clinical data
- Unstructured: Textual data, newspaper data
- Image Data
- Data in the form of audio, video.
- Know your data before jumping to analysis - get familiar with it. How?

Data Analysis

- Statistics is the tool that helps analyse data.
- Two types of statistics: **Descriptive Statistics** and **Inferential Statistics**.
- **Descriptive Statistics** : Useful to understand/explore/get familiar with data. Exploratory Data Analysis.
- **Inferential Statistics**: Hypothesis testing, Modelling, Predictive Modelling.
- A scientific researcher must be a bit detached from his/her work and not too over enthusiastic for it could risk objectivity required for analyses.
- Look at your results/data as they are, **not as you want to look at it due to pre-conceived notions**.
- If unexpected result appears: dig deeper. Check assumptions. Check analyses. Try to find out the reason for such finding than being upset.

Data Deluge

- Current world is rapidly amassing data.
- Everyone wants to be a Data Scientist but they confuse it with Number Crunching.
- The vision to analyse data has to be as clear as possible.
- What do you want to extract depends on how you perceive your data.
- **Relating to the world outside of our statistical codes is always important**.
- Think practically what makes sense and discard rest, after thinking over it.

Best way to learn?

- Go back to slide 2 and slide 7 and begin with analysing each type of data.
- Identify a research domain you wish to explore using data analysis.
- Learn the skills (R, Python, theoretical conceptualization for analysis).
- Read about statistical problem solving in a practical manner.
- Doing projects, mini-projects is the best way to learn.

Recommended Readings

- Python Vs. R for Data Science
- The most desired skill in data science.
- Data Dredging
- UCI Machine Learning Repository
- Kaggle: Machine Learning and Data Science Community

Next Lecture

Ethics in Innovation.