Toolkit 4: DATA ANALYSIS AND COMMUNICATION

Research methods

Science refers to a systematised production of knowledge. In this sense, it implies methods to verify causality between variables.

Methods might be divided into quantitative and qualitative. Quantitative studies comprise many cases trying to verify general trends, while qualitative studies usually approach a smaller number of cases or observations in more detail. Statistics is usually related to a quantitative design, but it can be used also in qualitative studies.

Statistical techniques might be useful for:

- Analysing general characteristics related to a group or a phenomenon.
- Building hypothesis.
- Testing hypothesis.
- Inferring causality between two phenomena.

Basic concepts in Statistics

- Descriptive statistics: it provides summary statistics. The goal is to describe the data: summarising information into some elements, such as the most frequent value, the mean or the median. It gives a clue about the distribution of the data.
- Inferential statistics: it aims to draw inferences about the population from a smaller sample. When we are trying to understand a relationship between two variables, we usually employ correlation tests or regressions.
- Variables: refer to an estimate that varies. For example, in an economic study, one of the variables might be economic growth, measured by macroeconomic indicators such as GDP.
- Independent variable: it is the consequence that the research wants to analyse.
- Dependent variable: it is the cause of the phenomenon analysed. Usually, a model contains more than one dependent variable as phenomena are multi-causal.
- Mean: it is the average calculated by the sum of all numbers divided by the quantity of numbers.
- Median: the median is the value located in the middle of the list.
- Standard error: as the standard deviation, it is a measure of spread. The standard error of the man indicates how far the sample mean deviates from the population mean.
- Variance: it is also a measure of spread within a set of data. The standard deviation, which indicates how the data sets are distributed around the mean, is the positive square root of the variance.

- Confidence level: expressed as a percentage, it represents the chance of getting the same results when repeating the experiment.

Correlation does not imply causality!

If two events happen at the same time, it does not mean the first causes the second. It might be a coincidence, or they might be related through another cause (˜intervening variable˜).

Data visualisation

Building graphs and expressing data through images help the audience to visualise the information easily and quickly. There are several software and tools that might be used for that: R, Python, Power BI, Tableau, and Flourish (Google). There are different types of graphs which are more adequate to certain data sets depending on what the researcher wants to communicate. The Financial Times has a chart explaining which type of data might be helpful for visualising certain types of data. The thumb rule is to maintain the graph as simple as possible.

Data sharing

- Before sharing data, verify if you have the rights to do it.
- Regulations that are relevant for data sharing: copyright/intellectual property and data protection.

References and further readings:

Statistics How To. https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/what-is-the-standard-error-of-a-sample/ and https://www.statisticshowto.com/probability-and-statistics/confidence-interval/#WhatisCI

Newcastle University maths resources: https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/descriptive-statistics/variance-and-standard-deviation.html

Beware Spurious Correlations (hbr.org)

Descriptive vs. Inferential Statistics: What's the Difference? - Statology