Sampling and hypothesis tests

Part 1: Introduction to sampling

By: Noureddin Sadawi, PhD

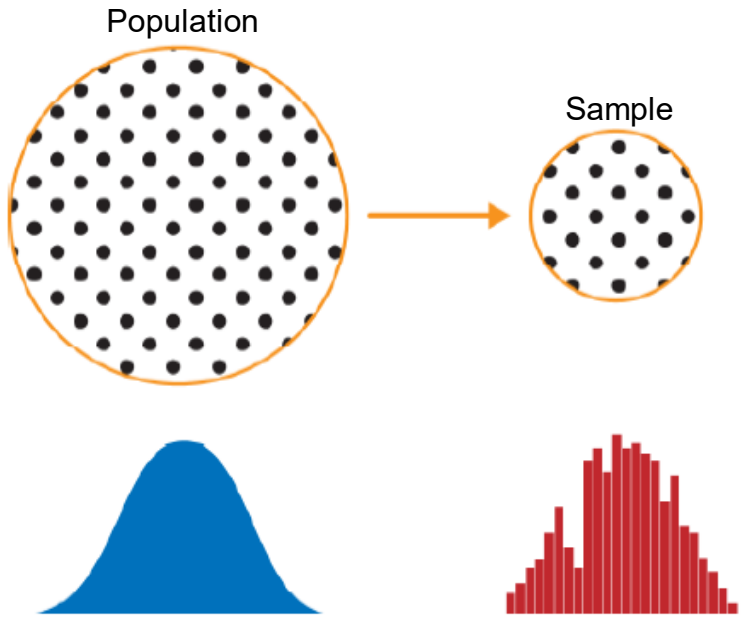
University of London

Introduction

- ‘A popular misconception holds that the era of big data means the end of a need for sampling.
- In fact, the proliferation of data of varying quality and relevance reinforces the need for sampling as a tool to work efficiently with a variety of data and to minimize **bias**.
- Even in a big data project, predictive models are typically developed and piloted with samples.
- Samples are also used in tests of various sorts (e.g., comparing the effect of web page designs on clicks).’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Population vs sample



Source: (Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

- ‘The lefthand side represents a population that, in statistics, is assumed to follow an underlying but unknown distribution.
- All that is available is the sample data and its empirical distribution, shown on the righthand side.
- To get from the lefthand side to the righthand side, a sampling procedure is used (represented by an arrow).’

Random sampling

- ‘A **sample** is a subset of data from a larger data set.
- Statisticians call this larger data set the population.
- **Random sampling** is a process in which each available member of the population being sampled has an equal chance of being chosen for the sample at each draw.
- The sample that results is called a simple random sample.
- Sampling can be done **with replacement**, in which observations are put back in the population after each draw for possible future reselection.
- Or it can be done **without replacement**, in which case observations, once selected, are unavailable for future draws.’

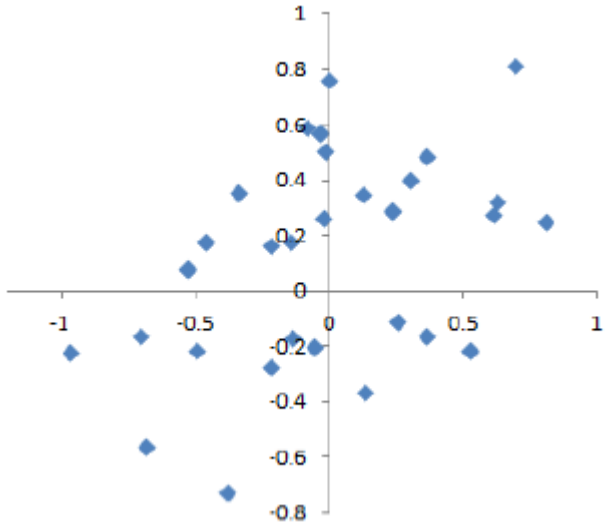
(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Bias

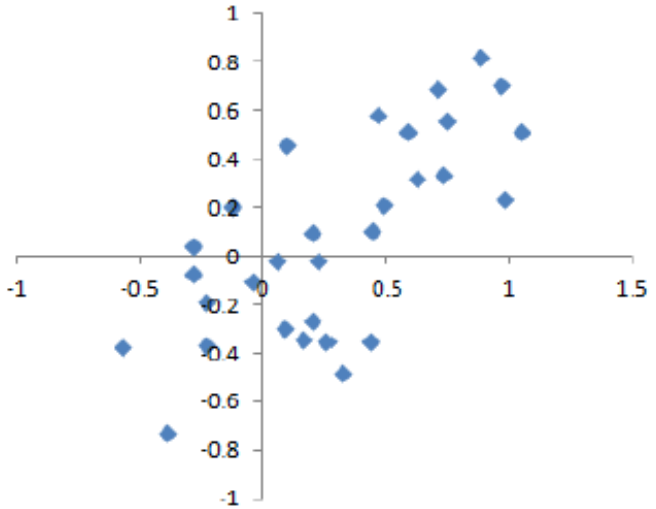
- ‘Statistical bias refers to measurement or sampling errors that are systematic and produced by the measurement or sampling process.
- An important distinction should be made between errors due to random chance and errors due to bias.
- Consider the physical process of a gun shooting at a target.
- It will not hit the absolute center of the target every time, or even much at all.
- An unbiased process will produce error, but it is random and does not tend strongly in any direction [see next slide].
- The results show a biased process-there is still random error in both the x and y direction, but there is also a bias. Shots tend to fall in the upper-right quadrant.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Example



Scatterplot of shots from a gun with true aim.



Scatterplot of shots from a gun with biased aim.

Random selection

- As its name suggests, random sampling needs to generate random samples.
- This is not always easy.
- One key aspect of it is: what is the population we have access to?
- Another aspect is the sampling procedure: with or without replacement?
- **Stratified sampling?**
- ‘In stratified sampling, the population is divided up into strata, and random samples are taken from each stratum.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Key ideas

- ‘Even in the era of big data, random sampling remains an important arrow in the data scientist’s quiver.
- Bias occurs when measurements or observations are systematically in error because they are not representative of the full population.
- Data quality is often more important than data quantity, and random sampling can reduce bias and facilitate quality improvement that would otherwise be prohibitively expensive.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Sampling distribution of a statistic

- ‘The term sampling distribution of a statistic refers to the distribution of some sample statistic over many samples drawn from the same population.
- Much of classical statistics is concerned with making inferences from (small) samples to (very large) populations.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Key terms

- ‘**Sample statistic:** A metric calculated for a sample of data drawn from a larger population.
- **Data distribution:** The frequency distribution of individual values in a data set.
- **Sampling distribution:** The frequency distribution of a sample statistic over many samples or resamples.
- **Central limit theorem:** The tendency of the sampling distribution to take on a normal shape as sample size rises.
- **Standard error:** The variability (standard deviation) of a sample statistic over many samples (not to be confused with standard deviation, which by itself, refers to variability of individual data values).’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).