Data Analytics

ECON 1008, Semester 1, 2019

Giulio Zanella

University of Adelaide

School of Economics

Announcements

- Assignment 1 in online, due Sunday night.
- Remember to activate your APLIA account

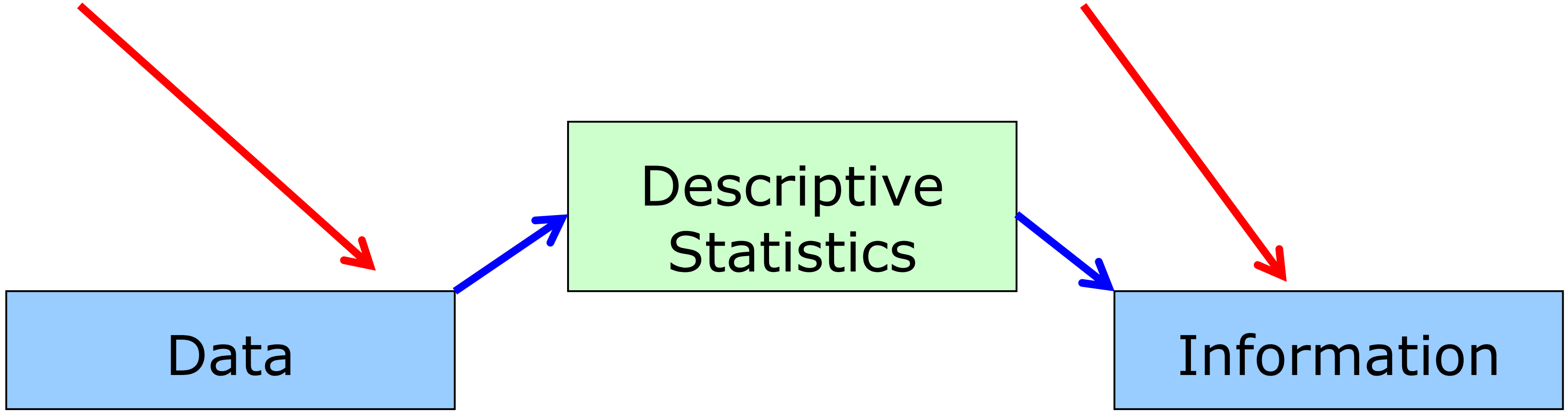
Chapter 2

Types of data, data collection and sampling

Introduction and re-cap…

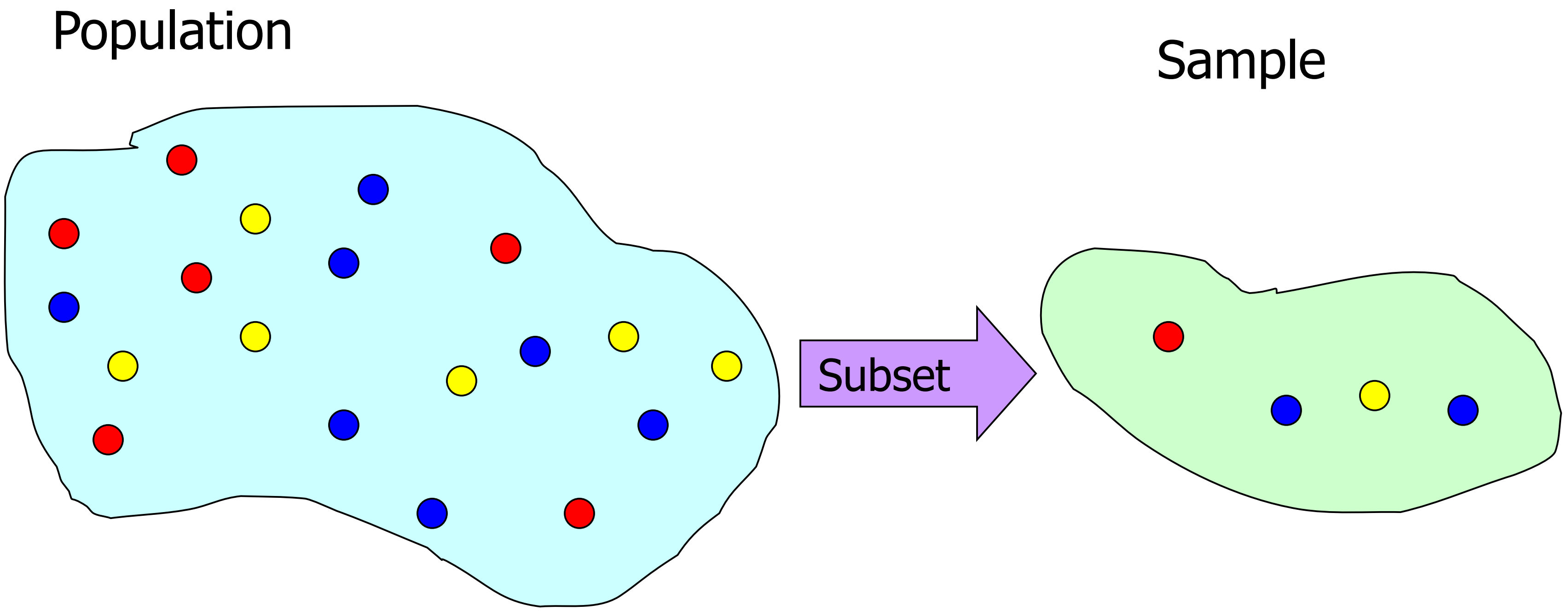
Descriptive statistics

involves arranging, summarising, and presenting a *set of data* in such a way that useful *information* is produced.



Its methods make use of graphical techniques and numerical descriptive measures (such as averages) to summarise and present the data.

Populations and samples



The graphical and tabular methods presented here apply to both entire populations ***and*** samples drawn from populations.

Two key definitions

1. **VARIABLE:** ***some characteristic*** of population or sample
   Example: students' marks in the example of Lecture 2

A variable is typically denoted with a capital letter: X, Y, Z…

The **values** of a variable are the range of possible values for that variable.

Example: possible student marks (0,…,50)

2. **DATA:** the ***observed values*** of a variable.
   E.g. student marks: {14, 16, … 36 … 40 … 46}

Three types of data

1. Numerical data

The values of **numerical** data are *real numbers*.

E.g. grades, eights, weights, prices, waiting time at a medical practice, etc.

Arithmetic operations can be performed on numerical data, thus its meaningful to talk about 2*Height, or Price + $1, and so on.

Numerical data are also called **quantitative**.

Three types of data

2. Nominal Data

The values of **nominal** data are *categories*.

E.g. Responses to questions about marital status are categories, coded as:

Single = 1, Married = 2, Divorced = 3, Widowed = 4

These data are **categorical** in nature; arithmetic operations don't make any sense (e.g. does Married ÷ 2 = Divorced?!)

All we can calculate is the proportion of data that falls into each category.

Nominal data are also called **qualitative** or **categorical**.

Three types of data

3. Ordinal Data

Ordinal data appear to be categorical in nature, but their values have an ***order***; a ranking to them:

E.g. University course evaluation system:

Poor = 1, fair = 2, good = 3, very good = 4, excellent = 5

While its still not meaningful to do arithmetic on this data (e.g. does 2*fair = very good?!), we can say things like:

excellent > poor   or   fair < very good

That is, order is maintained no matter what numeric values are assigned to each category.

Ordinal data are also called **ranked**.

Types of data – Examples

Numerical data

age	income
55	75 000
42	68 000
.	.
.	.

weight gain

+10

+5

.

Nominal data

person	married
1	yes
2	no
3	no
.	
.	

computer	brand
1	IBM
2	Dell
3	Compaq
4	IBM
.	.

IBM	Dell	Compaq	other	total
25	11	8	6	50
50%	22%	16%	12%	100%

Ordinal data

exam grade

HD

D

C

P

F

Food quality

Excellent

Good

Satisfactory

Poor

Summary ("hierarchy of data")

Numerical

- Values are real numbers.
- All calculations are valid.
- Data may be treated as ordinal or nominal.

Nominal

- Values are the arbitrary numbers that represent categories.
- Only calculations, such as proportions, based on the frequencies of occurrence are valid.
- Data may not be treated as ordinal or numerical.

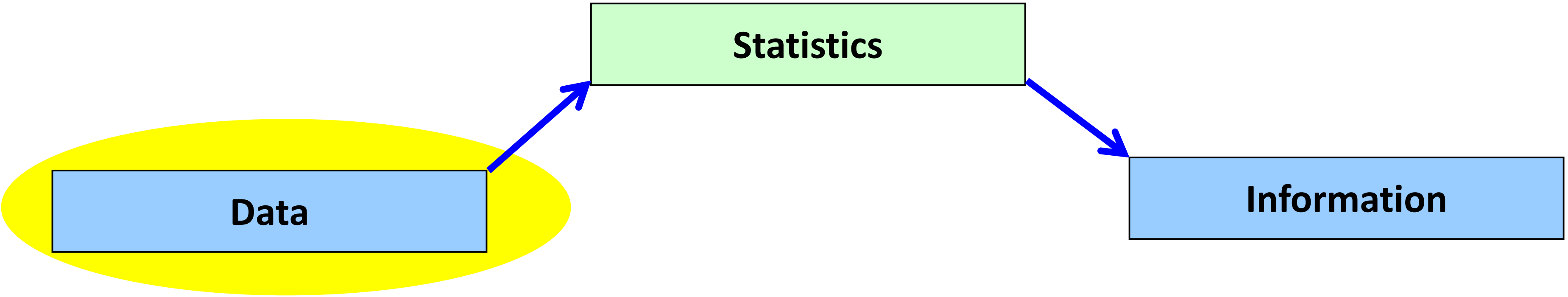
Ordinal

- Values must represent the ranked order of the data.
- Calculations based on an ordering process are valid.
- Data may be treated as nominal but not as numerical.

2.2 Methods of collecting data

Recall,

Statistics is a tool for converting ***data*** into useful ***information***:



1. But where then does data come from? How is it gathered?
2. How do we ensure its **accuracy**? Is the data **reliable**?
3. Is it **representative** of the population from which it was drawn?

Let’s explore some of these **key issues for data analysis.**

Sources of data

Four of the most popular sources of statistical data are:

- Published data
- Data collected from observational studies (Observational data)
- Data collected from experimental studies (Experimental data)
- Data collected from surveys (Survey data)

Published data

This type of data has already been collected by an organization or by a statistical agency and made available for others to use.

This is often a preferred source of data due to **low cost** and convenience.

Published data typically comes in digital form (or, if it's old, as printed material, disks, or tapes).

Types of published data

- Primary data
- Secondary data.

Published data…

Primary data

Data published by the organisation that has collected it is called **primary data**.

E.g. Data published by the *Australian Bureau of Statistics (ABS)*.

Secondary data

Data published by an organisation different from the one that was originally collected and published is called secondary data.

E.g. 1. The *Yearbook of National Accounts Statistics* (United Nations, New York), compiles data from primary sources of various country departments of statistics, like ABS in Australia;

2. The OECD compiles data from national sources for OECD countries

3. Compustat sells a variety of financial data compiled from several primary sources, for a large number of businesses.

Observational and experimental data

When published data is unavailable, one needs to conduct a study to generate the data.

- **Observational study** is one in which measurements representing a variable of interest are observed and recorded, without controlling any factor that might influence their values
  - e.g. measuring the height of a tree in the rainforest over time.
- **Experimental study** is one in which measurements representing a variable of interest are observed and recorded, while controlling factors that might influence their values
  - e.g. measuring the yield of different type of rice using a certain amount of fertilizer (treatment).

Surveys

A ***survey*** solicits information from survey participants;

> e.g. Gallup polls; pre-election polls; marketing surveys, Household, Income and Labour Dynamics in Australia (HILDA, 17k Australians) .

The ***response rate*** (i.e. the proportion of selected participants who completed the survey) is a key survey parameter.

Surveys may be administered in a variety of ways,

e.g.

- Personal interview
- Telephone or computer-based interview
- Self-administered questionnaire.

Sampling

If the data are collected from **the whole population**, we have a **census**. E.g.: the ABS conducts a census every 5 years in Australia.

However, a census is expensive! Recall that statistical inference permits us to draw conclusions about a population from a sample.

SAMPLING means **selecting a sub-set of a whole population.** This is often done instead of a census for a number of reasons including

- ***cost***

  For example, it's less expensive to sample 1000 television viewers than 20 million TV viewers
- ***practicality***

  For example, performing a crash test on every automobile produced is impractical.

Sampling…

Target population

The population about which we want to draw inferences (example: the Australian population today)

Sampled population

The actual population from which the sample has been drawn (e.g. Australian households occupying private dwellings in 2001, at the beginning of the project)

In any case, the ***sampled population*** and the ***target population*** should be similar to one another. Otherwise the sample selected may become self-selected.

Sampling…

Example:

A survey of opinion on a radio talk-back show topic

Target population: All radio listeners who listen to the talk-back show.

Sample selected: Those listeners who are interested in the topic and managed to contact the radio station.

Sampled population: Those listeners who are interested in the topic.

Sampling plans

A *sampling plan* is just a method or procedure for specifying how a sample will be taken from a population.

Most commonly used sampling plans,

- Simple random sampling
- Stratified random sampling
- Cluster sampling.

Simple random sampling

A simple random sample is a sample selected in such a way that every possible sample of the same size is equally likely to be chosen.

For example, drawing three names from a hat containing all the names of the students in a class of 200 is an example of a *simple random sample*: any group of three names is as equally likely as picking any other group of three names.

Simple random sampling…

To conduct simple random sampling…

- assign a number to each element of the chosen population (or use already given numbers),

  e.g. Medicare card number of each Australian resident

- randomly select the sample numbers (members) using a software (e.g., Excel).

Example 1

A government income-tax auditor is responsible for 1000 tax returns. The auditor will randomly select 30 returns to audit. Use Excel’s random number generator to select the returns.

Solution:

We generate 50 numbers between 1 and 1000 (we need only 30 numbers, but the extra numbers might be used if duplicate numbers are generated.)

Stratified random sampling

A stratified random sample is obtained by dividing the population into *mutually exclusive sets* (or strata), and then drawing simple random samples from each stratum.

Population 1

Occupation

- Professional
- Clerical
- Blue-collar
- Other

Population 2

Age

- Under 20
- 20–30
- 31–40
- 41–50
- 51–60
- > 60

Population 3

Sex

- Male
- Female

Stratified random sampling…

With this procedure we can acquire information or make inferences about

- the whole population
- each stratum
- the relationships among strata.

Cluster sampling

Cluster sample is a simple random sample of groups or clusters of elements (vs. a simple random sample consists of individual objects).

This procedure is useful when

- it is difficult and costly to develop a complete list of the population members (making it difficult to develop a simple random sampling procedure).
- the population members are widely dispersed geographically.

Cluster sampling…

Cluster sampling may increase sampling error, because of probable similarities among cluster members.

For example, to draw a cluster sample of residents in Adelaide, first select a number of streets in the Adelaide city area using a simple random sampling method and then include all residents in those selected streets to form the cluster sample.

Sample size

Numerical techniques for determining sample sizes will be described later. As a rule of thumb:

the larger the sample size, the more accurate we can expect the sample estimates to be

That is, the closer the sample estimate to the unknown population parameter we wish to estimate.

2.5 Sampling and non-sampling errors

Two major types of errors can arise when a sampling procedure is performed.

- Sampling error
- Non-sampling error

Sampling errors

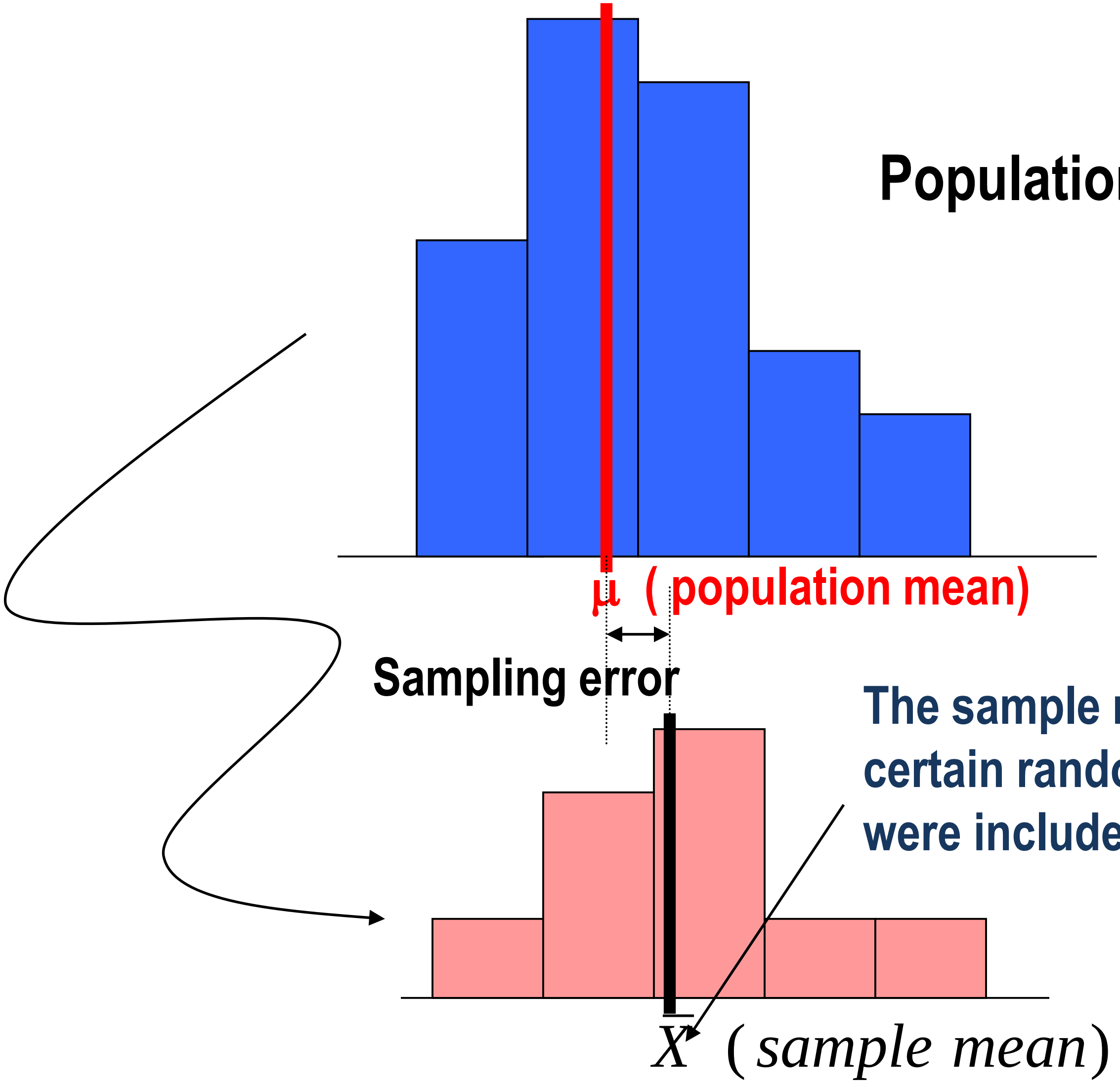
Sampling error refers to differences between the sample and the population, because of the specific observations that happen to be selected (“**good or bad luck**”)

Example: estimating a **population mean** using a **sample mean,**

sampling error = sample mean – population mean

Increasing the sample size will of course reduce the sampling error.

Sampling errors…



Population income distribution

μ (population mean)

Sampling error

The sample mean falls here only because certain randomly selected observations were included in the sample.

$\bar{X}$ (*sample mean*)

Non-sampling errors

Mistakes made along the process of data acquisition

Sample observations being selected improperly.

There are three types of non-sampling errors:

- Errors in data acquisition,
- Non-response errors,
- Selection bias.

Increasing the sample size will NOT reduce this type of error.

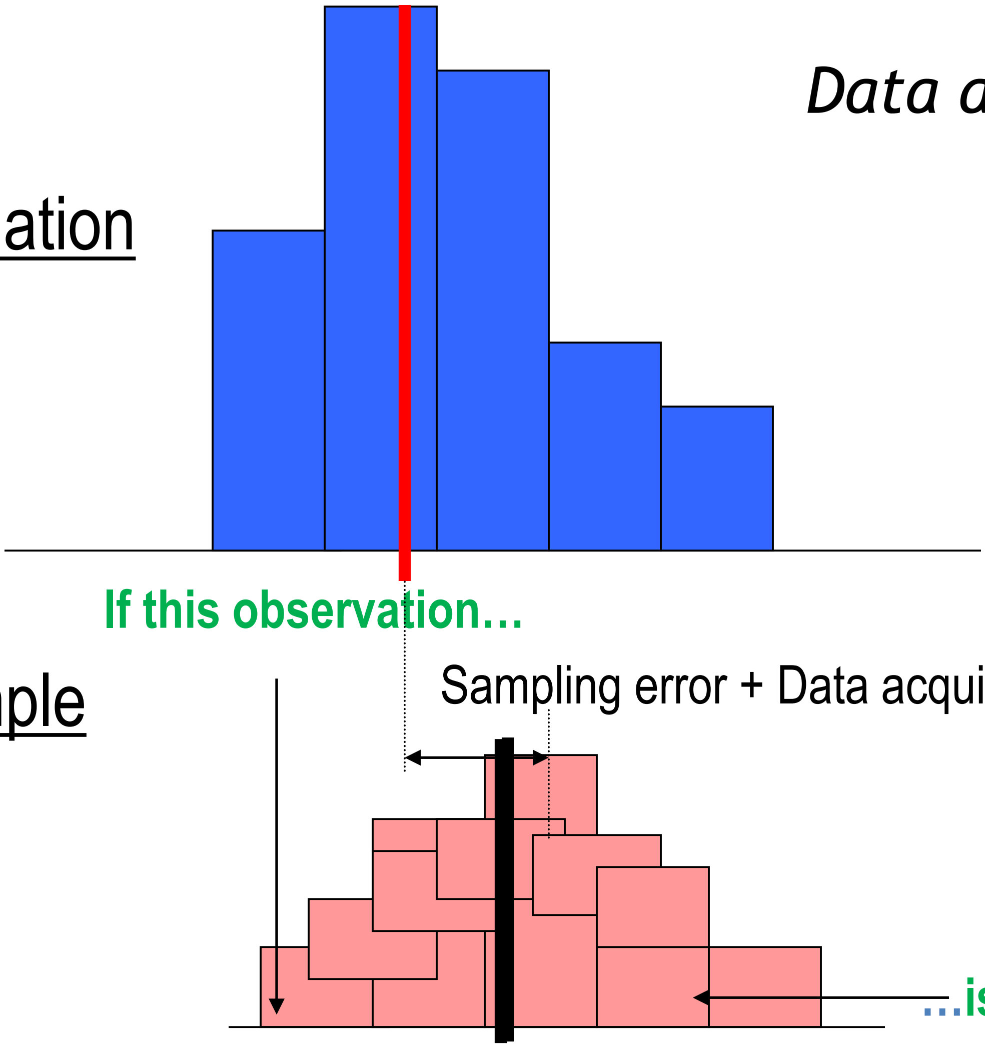
Errors in data acquisition

Errors in data acquisition arises from the recording of incorrect responses, due to:

- incorrect measurements being taken because of faulty equipment,
- mistakes made during transcription from primary sources,
- inaccurate recording of data due to misinterpretation of terms,
- inaccurate responses to questions concerning sensitive issues, or
- clerical mistakes when transferring/recording data.

Errors in data acquisition

Data acquisition error



Population

If this observation…

Sample

Sampling error + Data acquisition error

…is wrongly recorded here…

…then the sample mean is affected.

Non-response error

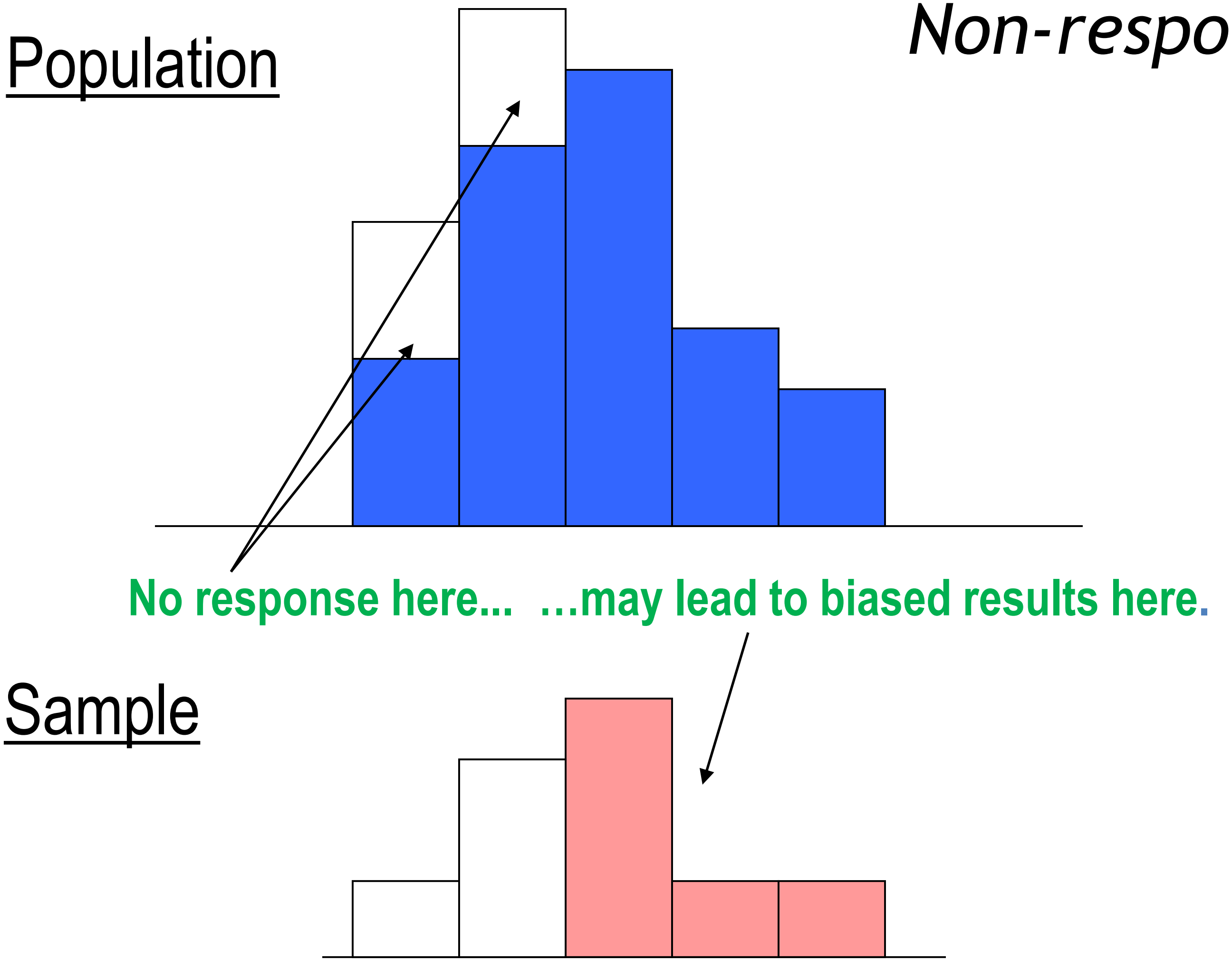
- **Non-response error** refers to error (or ***bias***) introduced when responses are not obtained from some members of the sample surveyed due to refusal by them to respond for some reason

- The sample observations that are collected may not be **representative** of the target population.

- The ***Response Rate*** (i.e. the proportion of all people selected who complete the survey) is a key survey parameter and helps in the understanding in the validity of the survey.

Non-response error

Non-response error



Population

No response here… …may lead to biased results here.

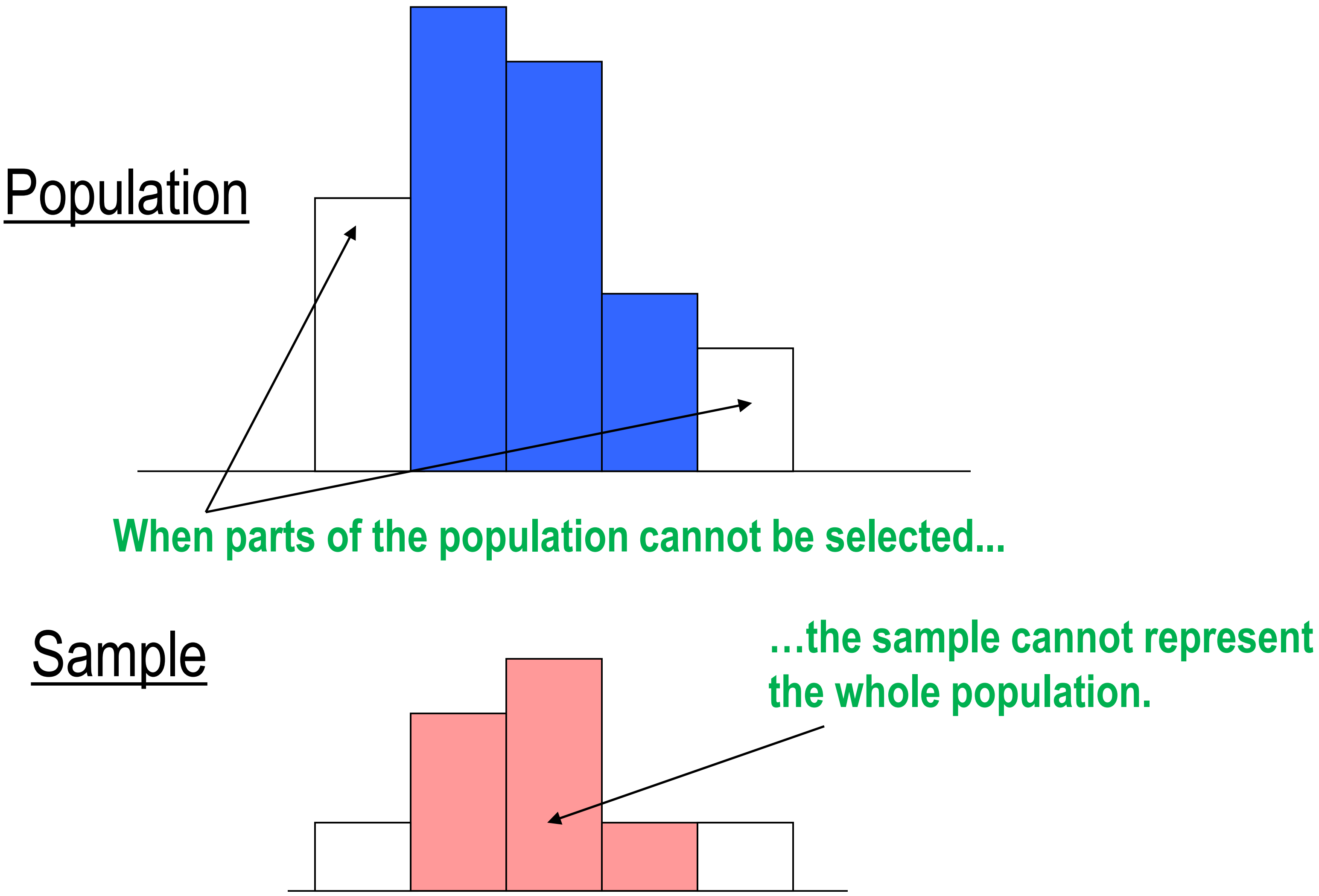
Sample

Selection bias

Selection bias occurs when the sampling plan is such that some members of the target population cannot possibly be selected for inclusion in the sample.

For example, selecting a sample of households in NSW using telephone numbers listed in NSW White Pages, as not every NSW household telephone number is listed in the White Pages.

Selection bias



Population

When parts of the population cannot be selected...

Sample

…the sample cannot represent the whole population.